Estadística para la resolución de problemas en Ingeniería

Ferreri, Carnevali, Barrea, Catalano, Chiapella, Leoni, Cardone, Fernández de Luco, Fusari

FCEIA

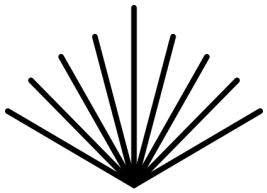
UNR Universidad Nacional de Rosario

UNR EDITORA

Estadística para la resolución de problemas en Ingeniería

Estadística para la resolución de problemas en Ingeniería

Noemí M. Ferreri
Graciela H. Carnevali
Leonardo D. Barrea
Mara L. Catalano
Luciana C. Chiapella
Valeria A. Leoni
Paula Cardone
María E. Fernández de Luco
Julieta L. Fusari



UNR
EDITORA

Estadística para la resolución de problemas en Ingeniería / Noemí M. Ferreri ... [et al.]. -
1a ed. - Rosario : UNR Editora, 2026.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-987-702-786-0

1. Estadísticas. 2. Ingeniería. I. Ferreri, Noemí M.
CDD 600

Luciana C. Chiapella: diseño y maquetación del libro
Paula Cardone: diseño de tapa

Queda hecho el depósito que marca la Ley N° 11.723.}
Copyright © 2026
Universidad Nacional de Rosario

UNR editora
Editorial de la Universidad Nacional de Rosario
Secretaría de Extensión Universitaria
Urquiza 2050 - S2000AOB / Rosario, República Argentina
www.unreditora.unr.edu.ar / editora@sede.unr.edu.ar

Sobre los autores

Los autores de esta obra poseen formación de grado y posgrado en estadística, matemática, docencia universitaria, ingeniería de la calidad y áreas afines. Cuentan con una amplia trayectoria docente en estadística, participando en la formación de estudiantes de ingeniería industrial, mecánica, eléctrica, electrónica, civil y agrimensura. Asimismo, se desempeñan como docentes en otras carreras de grado y posgrado, y como capacitadores en cursos orientados al ámbito profesional y empresarial.

En el campo de la investigación, han integrado numerosos proyectos vinculados con la educación estadística, el desarrollo del pensamiento estadístico, la aplicación de la estadística en organizaciones y la resolución de problemas ingenieriles de diversa complejidad. Además, varios de los autores ejercen profesionalmente como estadísticos en el ámbito empresarial, aportando una perspectiva aplicada que enriquece el enfoque de esta obra.

Agradecimientos

Los autores agradecen a los colegas que, a lo largo de los años, han formado parte de la cátedra de Estadística de la Facultad de Ciencias Exactas, Ingeniería y Agrimensura de la Universidad Nacional de Rosario.

El reconocimiento se extiende tanto a los docentes que integraron la cátedra antes del inicio de la escritura de este libro, cuyos aportes e ideas enriquecieron los materiales elaborados para los estudiantes y constituyen el punto de partida de esta obra, como a aquellos que se incorporaron cuando el proceso de escritura estaba finalizando, y contribuyeron a su mejora mediante observaciones y sugerencias.

Agradecen también a los estudiantes que utilizaron este material a lo largo de los años, cuyas preguntas y comentarios contribuyeron de manera significativa a su desarrollo y mejora.

Prólogo

La formación estadística constituye una competencia fundamental para quienes se desempeñan en el campo de la ingeniería. Los profesionales no solo deben interpretar información estadística incluida en informes, sino también generar datos, analizarlos críticamente, comunicar adecuadamente sus conclusiones fundamentándolas en los resultados obtenidos y utilizarlas en la toma de decisiones. Esto requiere tanto el dominio de conceptos y procedimientos como el desarrollo de un pensamiento estadístico que permita comprender la variabilidad presente en los procesos y utilizarla como base para la toma de decisiones.

El pensamiento estadístico implica formular preguntas relevantes, reconocer fuentes de variabilidad, planificar estudios estadísticos, recolectar datos, analizarlos e interpretar los resultados en contexto. Se trata de una forma de razonamiento compleja y articulada, cuyo desarrollo se inicia en un curso introductorio y se perfecciona con la experiencia. Este pensamiento se asume como principio orientador de los cursos de Estadística para estudiantes de las diferentes carreras de Ingeniería de la Facultad de Ciencias Exactas, Ingeniería y Agrimensura y de este texto, que se utiliza como soporte, promoviendo la resolución sistemática de problemas y el desarrollo de una actitud crítica frente a los resultados obtenidos.

En este marco, el libro presenta distintas etapas de resolución de problemas de naturaleza estadística vinculados con situaciones propias de la ingeniería. Los problemas abordados se desarrollan de manera simplificada, acorde con un curso introductorio, y se circunscriben al estudio de una única variable de interés. No obstante, se procura favorecer la integración progresiva de conceptos, procedimientos y estrategias de resolución que permitan construir formas de razonamiento transferibles a situaciones de mayor complejidad.

La organización general del texto sigue el ciclo de resolución de problemas de naturaleza estadística, conocido como ciclo PPDAC —Planteo del problema, Planificación del estudio estadístico, Recolección de datos, Análisis de los datos y Obtención de conclusiones—, propuesto por Wild y Pfannkuch en el año 1999. Este ciclo constituye un eje articulador de los contenidos y permite comprender la estadística no como un conjunto aislado de técnicas, sino como una disciplina orientada a la resolución de problemas.

La obra se estructura en nueve capítulos, cuyo contenido se resume a continuación.

El Capítulo 1 introduce los problemas de naturaleza estadística y presenta, de manera general, las etapas del ciclo PPDAC. Se describen las tareas y decisiones involucradas en cada etapa, con especial énfasis en el planteo del problema y la planificación del estudio estadístico. Asimismo, se presentan

nueve problemas que pretenden abarcar varias áreas del ámbito de la Ingeniería. Estos problemas se retoman a lo largo del libro y se utilizan como ejemplos y actividades para el desarrollo de los distintos contenidos, funcionando como hilo conductor del texto.

El Capítulo 2 aborda el análisis descriptivo de un conjunto finito de datos correspondientes a una única variable de interés. Se desarrollan herramientas para organizar, representar y resumir la información mediante tablas, gráficos y medidas descriptivas. Cuando los datos corresponden a la totalidad de una población finita, el análisis descriptivo permite obtener conclusiones definitivas. En cambio, cuando los datos provienen de una muestra, las conclusiones obtenidas se consideran preliminares y deben complementarse posteriormente mediante herramientas inferenciales.

Los Capítulos 3 al 6 desarrollan los conceptos teóricos vinculados con la probabilidad y las distribuciones de probabilidades como modelos para describir el comportamiento de variables aleatorias en poblaciones infinitas. Se introducen los conceptos de variable aleatoria, distribución de probabilidades y parámetros asociados (Capítulo 3); se presentan aspectos formales de la probabilidad (Capítulo 4) y algunos modelos frecuentemente utilizados en problemas de ingeniería (Capítulo 5); y se estudian funciones de variables aleatorias y sus distribuciones de probabilidades resultantes (Capítulo 6). Los conceptos desarrollados en estos capítulos constituyen el soporte teórico de las herramientas inferenciales presentadas posteriormente.

Los Capítulos 7 al 9 están dedicados al análisis inferencial. Allí se presentan herramientas que permiten obtener conclusiones sobre una población a partir de información muestral. Se introduce el concepto de muestra aleatoria simple y las distribuciones muestrales asociadas a distintos estimadores (Capítulo 7). Posteriormente se desarrollan procedimientos de estimación —incluyendo intervalos de confianza e intervalos de tolerancia— (Capítulo 8) y pruebas de hipótesis sobre parámetros y distribuciones de probabilidades (Capítulo 9). Asimismo, se enfatiza la consideración de los riesgos asociados a inferencias erróneas y la necesidad de interpretar los resultados en el contexto del problema estudiado.

En todos los capítulos se incluyen ejemplos resueltos y actividades propuestas para favorecer el trabajo autónomo y la integración de los contenidos desarrollados. Asimismo, se incorporan apartados específicos dedicados al uso de R y RStudio como herramientas de apoyo para el análisis de datos, la representación gráfica y la implementación de procedimientos estadísticos. En estos apartados se escriben y explican los códigos de R que se pueden utilizar para obtener los resultados correspondientes a los ejemplos presentados en cada capítulo, permitiendo al lector reproducirlos y profundizar en su comprensión. La inclusión de estos recursos busca acompañar el aprendizaje de la disciplina mediante la incorporación de prácticas actuales de análisis y trabajo con datos.

Es el deseo de los autores que este libro constituya una herramienta útil para quienes se inician en el estudio de la estadística y que contribuya no solo a la adquisición de técnicas y procedimientos, sino también al desarrollo de una manera de pensar necesaria para el ejercicio profesional de la ingeniería.



Índice general

1 El ciclo de resolución de problemas 1

1.1 Introducción 1

1.2 Los procesos y la variabilidad 2

1.3 Problemas en un marco de variabilidad e incertidumbre 5

1.4 Problemas de naturaleza estadística para trabajar 7

1.5 El ciclo para la resolución de problemas de naturaleza estadística 9

1.5.1 Etapa de Planteo del problema (P) 10
1. Conceptos estadísticos asociados 10
2. Tareas que se realizan en la etapa de Planteo del problema 16

1.5.2 Etapa de Planificación del estudio estadístico (P) 18
1. Conceptos estadísticos asociados 18
2. Tareas que se realizan en la etapa de Planificación del estudio estadístico 20

1.5.3 Etapa de Recolección de datos (D) 22
1. Conceptos estadísticos asociados 22
2. Tareas que se realizan en la etapa de Recolección de datos 24

1.5.4 Etapa de Análisis de los datos (A) 25
1. Conceptos estadísticos asociados 25
2. Tareas que se realizan en la etapa de Análisis de los datos 25

1.5.5 Etapa de Obtención de conclusiones (C) 26
1. Conceptos estadísticos asociados 26
2. Tareas que se realizan en la etapa de Obtención de conclusiones 27

1.6 Síntesis 30

1.7 Actividades propuestas 32

1.8 Estadística con R 35

1.8.1 Introducción 35
1.8.2 Instalación 35
1.8.3 Entorno 36
1.8.4 Trabajo con paquetes 37
1. Instalación de paquetes 37
2. Uso de paquetes 38

1.8.5 Importación de datos . . . 39
1. Importación de datos desde Excel . . . 39
2. Importación de datos en texto plano . . . 40

2 Análisis descriptivo de un conjunto de datos . . . 43

2.1 Introducción 43

2.2 Organización de los datos 44

2.2.1 Distribuciones de frecuencias para variables cualitativas o categóricas . . . 46
1. Gráfico de barras . . . 48
2. Gráfico de sectores . . . 48
3. Diagrama de Pareto . . . 48

2.2.2 Distribuciones de frecuencias para variables cuantitativas discretas . . . 52
1. Gráfico de bastones . . . 54
2. Gráfico escalonado . . . 54

2.2.3 Distribuciones de frecuencias para variables cuantitativas continuas . . . 56
1. Histograma . . . 60
2. Polígono de frecuencias . . . 61
3. Gráfico de frecuencias acumuladas o polígono acumulativo . . . 61

2.2.4 Otras formas de representar las distribuciones de frecuencias . . . 62
1. Diagrama de puntos . . . 63
2. Diagrama de tallo y hoja . . . 64
3. Gráfico de series cronológicas o series de tiempo . . . 65

2.3 Medidas de resumen de una variable 68

2.3.1 Proporción de unidades que satisfacen cierta condición . . . 69

2.3.2 Medidas de localización o posición . . . 69
1. Valores mínimo y máximo . . . 70
2. Percentiles . . . 70
3. Promedio o media aritmética . . . 73
4. Media truncada . . . 77
5. Media geométrica . . . 78
6. Moda . . . 79

2.3.3 Medidas de dispersión o variabilidad . . . 82
1. Rango . . . 83
2. Rango intercuartílico . . . 83
3. Variancia y desviación estándar . . . 84
4. Coeficiente de variación . . . 87

2.3.4 Diagrama de caja y bigotes . . . 89

2.3.5 Aspectos que involucran a las medidas de localización y de variabilidad . . . 94

2.3.6 Medidas de resumen cuando se aplica una transformación lineal . . . 99

2.4 Síntesis 102

2.5 Actividades propuestas 105

2.6 Estadística con R 123

2.6.1 Análisis descriptivo de variables cualitativas . . . 123

2.6.2 Análisis descriptivo de variables cuantitativas discretas . . . 127

2.6.3 Análisis descriptivo de variables cuantitativas continuas . . . 131

2.6.4 Otros gráficos . . . 134
1. Diagrama de puntos . . . 134
2. Diagrama de tallo y hojas . . . 135
3. Serie de tiempo . . . 135
4. Diagrama de caja y bigotes . . . 136

2.6.5 Medidas de resumen . . . 138
2.7 Apéndice 140
2.7.1 Medidas de resumen obtenidas en la resolución de los problemas 1, 4 y 5 . . . 140

3 Distribuciones de probabilidades para variables aleatorias . 143
3.1 Introducción 143
3.2 Variable aleatoria y su distribución de probabilidades 144
3.2.1 Variable aleatoria continua. Función de densidad de probabilidad . . . 145
3.2.2 Variable aleatoria discreta. Función de probabilidad puntual . . . 149
3.2.3 Función de distribución acumulada . . . 151
3.3 Parámetros de una variable aleatoria 155
3.3.1 Parámetros de localización o posición . . . 155
3.3.2 Parámetros de dispersión o variabilidad . . . 157
3.3.3 Parámetros que resultan cuando se aplica una transformación lineal . . . 159
3.3.4 Algunas propiedades de los procedimientos de medición . . . 161
3.4 Síntesis 164
3.5 Actividades propuestas 166
3.6 Estadística con R 173
3.6.1 Variables aleatorias continuas . . . 173
1. Funciones de densidad de probabilidad y de distribución acumulada . 173
2. Medidas de resumen . . . 176
3.6.2 Variables aleatorias discretas . . . 177
1. Funciones de probabilidad puntual y de distribución acumulada . . . 177
2. Medidas de resumen . . . 179
3.7 Apéndice 180
3.7.1 Sobre la variancia de una variable aleatoria . . . 180
3.7.2 Sobre la esperanza y la variancia de una variable X, función lineal de Y . . . 181

4 Experimentos aleatorios y probabilidad . . . 183
4.1 Introducción 183
4.2 Algunas definiciones preliminares 183
4.3 La probabilidad de un suceso 189
4.3.1 ¿Cómo se pueden asignar probabilidades? . . . 189
1. Encontrando relaciones que permitan determinar probabilidades . . . 189
2. Conociendo la ley de probabilidad . . . 190
3. Estudiando las frecuencias relativas al repetir el experimento . . . 191
4.3.2 La probabilidad en la toma de decisiones . . . 192
4.4 Probabilidad. Definición axiomática. Algunas consecuencias 193
4.5 Probabilidad condicional. Independencia entre sucesos 194
4.6 Teorema de la Probabilidad Total 200
4.7 Síntesis 202
4.8 Actividades propuestas 203
4.9 Apéndice 209
4.9.1 Probabilidad. Definición axiomática. Consecuencias . . . 209
4.9.2 Construcción de un diagrama de árbol . . . 210

4.9.3 Teorema de la Probabilidad Total . . . 211

5 Distribuciones de probabilidades de uso frecuente . . . 213

5.1 Introducción 213

5.2 Variables aleatorias continuas 213

5.2.1 Distribución Normal . . . 213

5.2.2 Distribución Uniforme . . . 223

5.2.3 Distribución Triangular . . . 227

5.2.4 Distribución Exponencial . . . 232

5.3 Variables aleatorias discretas 236

5.3.1 Distribución Bernoulli . . . 236

5.3.2 Distribución Binomial . . . 239

5.3.3 Distribución Hipergeométrica . . . 244

5.3.4 Distribución Geométrica . . . 249

5.3.5 Distribución Poisson . . . 252

1. Proceso Poisson . . . 254
2. Relación entre la distribución Poisson y la distribución Exponencial . . . 256

5.4 Síntesis 258

5.5 Actividades propuestas 260

5.6 Estadística en R 282

5.6.1 Distribución Normal . . . 283

5.6.2 Distribución Uniforme . . . 287

5.6.3 Distribución Binomial . . . 288

5.6.4 Otras distribuciones . . . 291

5.7 Apéndice 294

6 Funciones de un vector aleatorio . . . 297

6.1 Introducción 297

6.2 Vector aleatorio y su comportamiento en probabilidad 298

6.2.1 Sobre la relación entre las variables del vector . . . 300

6.3 Variables definidas como funciones de un vector 300

6.3.1 Esperanza y desviación estándar de la nueva variable . . . 304

1. Esperanza de la nueva variable G . . . 304
2. Desviación estándar de la nueva variable G . . . 304

6.3.2 Distribución de probabilidades de la nueva variable . . . 305

1. Propiedad reproductiva de la suma de variables . . . 305
2. Teorema Central del Límite . . . 313
3. Simulación . . . 317

6.4 Introducción al estudio de las mediciones y sus errores 327

6.5 Síntesis 328

6.6 Actividades propuestas 330

6.7 Estadística con R 335

6.8 Apéndice 343

6.8.1 Distribuciones de probabilidades conjunta, marginales y condicionales . . . 343

1. Vectores bivariados discretos . . . 343
2. Vectores bivariados continuos . . . 348

6.8.2 Variables relacionadas y variables independientes. Covariancia y correlación . . 350
1. Variables independientes entre sí 351
2. Variables relacionadas entre sí 352
6.8.3 Esperanza y variancia de una combinación lineal de variables aleatorias 353
1. Esperanza de una combinación lineal de variables aleatorias 354
2. Variancia de una combinación lineal de variables aleatorias 354
6.8.4 Propiedad reproductiva para los modelos Binomial y Poisson 354

7 Muestras aleatorias simples y distribuciones muestrales 361

7.1 Introducción 361
7.2 Muestra aleatoria simple 362
7.3 Parámetros y estadísticos 363
7.4 Distribuciones muestrales 364
1. Antes de seleccionar la muestra 364
2. Después de seleccionar la muestra 364
7.4.1 Distribución de probabilidades del promedio o media muestral 365
7.4.2 Distribución de probabilidades de la frecuencia relativa de la muestra 370
7.4.3 Distribución de probabilidades de la variancia muestral 373
7.4.4 Uso de simulación para la obtención de la distribución de un estimador 375
7.4.5 Otros métodos para obtener la distribución del estimador 376
7.4.6 Propiedades de los estimadores 377
7.5 Síntesis 378
7.6 Actividades propuestas 379
7.7 Estadística con R 384
7.8 Apéndice 386
7.8.1 Estadístico: Promedio o media muestral 386
7.8.2 Estadístico: Proporción o frecuencia relativa muestral 386
7.8.3 Estadístico: Variancia muestral 387
7.8.4 Distribuciones t de Student y chi-cuadrado 388
7.8.5 Propiedades de los estimadores 389

8 Inferencia estadística. Estimación 391

8.1 Introducción 391
8.2 Inferencias sobre los parámetros 392
8.3 Intervalos de confianza 394
8.3.1 Intervalo de confianza para la media poblacional 394
1. Cuando el desvío estándar poblacional es conocido 394
2. Cuando el desvío estándar poblacional es desconocido 405
8.3.2 Intervalo de confianza para la variancia y el desvío estándar poblacionales 409
8.3.3 Intervalo de confianza para la proporción 411
8.4 Intervalo de tolerancia 417
8.5 Evaluación de los requerimientos de las herramientas presentadas 419
8.6 Síntesis 422
8.7 Actividades propuestas 426
8.8 Estadística con R 438
8.8.1 Diagrama de probabilidad Normal 438

8.8.2 Intervalos de confianza 439
1. Intervalo de confianza para la media poblacional 439
2. Intervalo de confianza para la variancia poblacional 440
3. Intervalo de confianza para la proporción poblacional 441
8.8.3 Intervalos de tolerancia 442

9 Inferencia estadística. Pruebas de hipótesis 445

9.1 Introducción 445

9.2 Pruebas de hipótesis 446

9.3 Pruebas de hipótesis sobre parámetros 448

9.3.1 Pruebas de hipótesis para la media poblacional 449
1. Cuando el desvío estándar poblacional es conocido 449
2. Cuando el desvío estándar poblacional es desconocido 467
9.3.2 Pruebas de hipótesis para la variancia y el desvío estándar poblacionales 471
9.3.3 Prueba de hipótesis para la proporción 474
9.3.4 Prueba de hipótesis para la mediana poblacional 482

9.4 Pruebas no paramétricas. Pruebas de bondad de ajuste 486

9.5 Síntesis 493

9.6 Actividades propuestas 495

9.7 Estadística con R 515

9.7.1 Pruebas de hipótesis respecto de parámetros 515
1. Pruebas de hipótesis para la media poblacional 515
2. Pruebas de hipótesis para la variancia y el desvío estándar poblacionales 517
3. Prueba de hipótesis para la proporción poblacional 518
4. Prueba de hipótesis para la mediana poblacional 520
9.7.2 Pruebas de bondad de ajuste 524

Bibliografía 527



1. El ciclo de resolución de problemas

1.1 Introducción

Quienes ejercen profesionalmente la ingeniería son responsables del diseño y de la producción de la mayoría de los productos que se utilizan en la actualidad, desde simples tornillos hasta vehículos y maquinarias de gran porte. También se ocupan del diseño, la gestión y la puesta en marcha de sus procesos productivos; están involucrados en la generación y en la distribución de la energía, así como en el uso eficiente de la misma y tienen injerencia en el desarrollo de nuevos materiales y equipos. Hay trabajo de ingeniería en el diseño, la construcción y el mantenimiento de obras de infraestructura, viviendas familiares y edificios de usos variados, así como en la elaboración de equipos de medición y diagnóstico que se utilizan, entre otras cosas, para mejorar la salud y la calidad de vida de las personas. Quienes trabajan en ingeniería son, además, responsables del avance en la comunicación y la informática.

En el ejercicio de estas actividades surgen problemas ingenieriles de distinta naturaleza y complejidad, que deben ser resueltos considerando múltiples aspectos simultáneamente, tales como costos, restricciones técnicas, factibilidad y desempeño de los sistemas involucrados.

En el abordaje de estos problemas, se presentan fuentes de variabilidad que afectan el comportamiento de características de interés, ya sea del proceso en sí o de alguna de sus salidas. Esta variabilidad genera incertidumbre, por lo que no es posible predecir con exactitud los valores de dichas características en una unidad en particular.

Como consecuencia de esta situación, en muchos de estos problemas ingenieriles aparecen preguntas específicas que no pueden responderse de manera determinística, precisamente por la presencia de la variabilidad y la incertidumbre. Estas preguntas constituyen lo que en adelante se denominarán *problemas de naturaleza estadística*, cuya resolución exige la realización de estudios estadísticos adecuados.

La estadística no solo proporciona conceptos y procedimientos para abordar estos problemas, sino que, más importante aún, promueve una forma particular de razonar: el *pensamiento estadístico*. Este

enfoque implica comprender por qué y cómo se llevan a cabo las investigaciones estadísticas, así como las ideas fundamentales que las sustentan. Entre ellas se destacan la omnipresencia de la variabilidad, la necesidad de disponer de datos adecuados, el uso de métodos apropiados para su recolección y análisis, la importancia del muestreo, el diseño de experimentos y la utilización de modelos matemáticos para la toma de decisiones informadas.

En la resolución de problemas de naturaleza estadística se sigue un ciclo compuesto por una serie de etapas, más o menos complejas según el caso, que permite arribar a conclusiones satisfactorias. Cuando estos problemas se enmarcan en contextos ingenieriles, dichas conclusiones contribuyen a la resolución del problema general.

Diversos autores han desarrollado este ciclo de resolución, asignándole distintas denominaciones. En este texto se considera el Ciclo PPDAC (Planteo del problema, Planificación del estudio estadístico, Recolección de datos, Análisis de los datos, Obtención de conclusiones) propuesto por Wild y Pfannkuch en 1999 [1].

En este primer capítulo se presenta el Ciclo PPDAC y cada una de sus etapas, con especial énfasis en la primera y la segunda. Además se proponen problemas del campo ingenieril, que se irán resolviendo a lo largo de todo el libro.

Los objetivos de este capítulo son:

- Presentar las etapas del proceso de resolución de problemas de naturaleza estadística, y aplicarlas a los problemas propuestos.
- Presentar los conceptos y procedimientos estadísticos básicos de cada etapa, los cuales se irán profundizando en capítulos posteriores y, con más detalle, los conceptos y procedimientos estadísticos asociados a las etapas de Planteo del problema y de Planificación del estudio estadístico.

1.2 Los procesos y la variabilidad

El término “proceso” presenta muchos significados, según el contexto o área de aplicación que corresponda. En general, se lo puede definir como un conjunto de fases sucesivas de un fenómeno natural o de una operación artificial o, también, como una serie de etapas, ordenadas con cierta lógica, para obtener un resultado. Un proceso de producción es un sistema de acciones que se encuentran interrelacionadas de forma dinámica y que se orientan a la transformación de ciertos elementos.

De estas definiciones se desprende que en todo proceso hay un conjunto de etapas que conducen a un resultado final y que, a lo largo de dichas etapas, se produce alguna transformación o cambio (Figura 1.1).



Figura 1.1 *Esquema de un proceso*

El esquema de la Figura 1.1 es una simplificación de un proceso: no siempre las etapas se dan de manera secuencial o lineal.

Considere, a modo de ejemplo, un proceso de producción industrial que puede definirse como un conjunto de actividades orientadas a la transformación de recursos o factores productivos en bienes. Esto incluye a una gran variedad de procesos, como por ejemplo la producción de harinas o quesos, automóviles, sustancias químicas o acero (Figura 1.2) hasta la generación de energía. También pueden considerarse procesos de producción de servicios como la entrega de productos comprados en línea (Figura 1.3), entre otros.

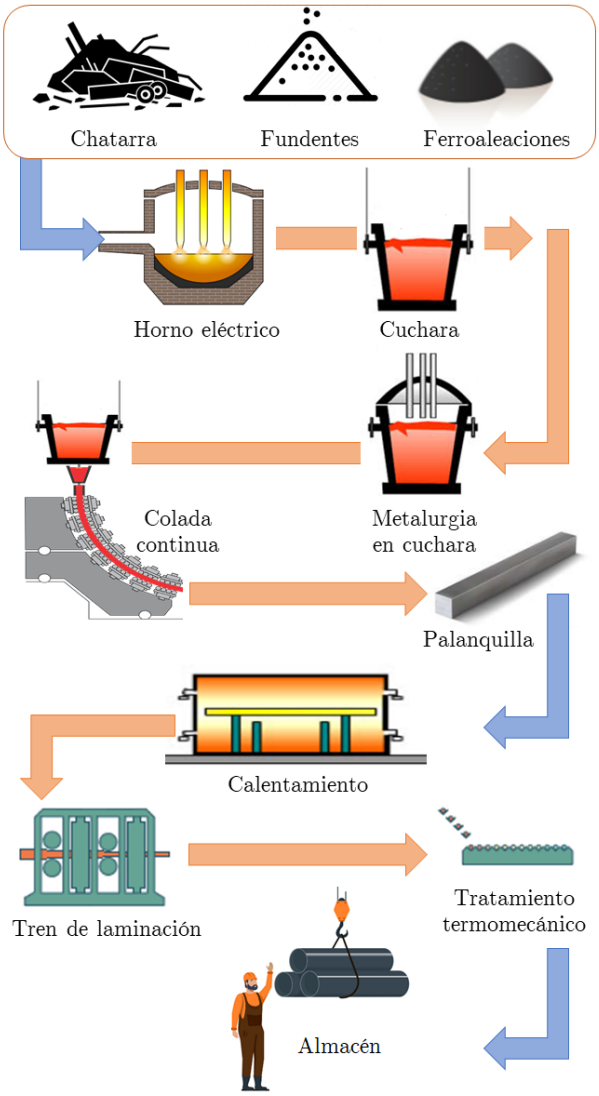


Figura 1.2 *El proceso de fabricación de tubos de acero*

En todos los procesos se desarrolla un conjunto definido de etapas que culminan con la obtención del producto o servicio de interés. Si existe control sobre el proceso, se intenta que cada vez que este se ejecute, siempre se trabaje de la misma manera y que, además, los productos o servicios satisfagan requerimientos o especificaciones impuestos por algún cliente. Sin embargo, la evidencia indica que estos productos o servicios varían entre sí en relación a diferentes características (peso, diámetro, viscosidad, dureza, sabor, resistencia a la corrosión, tiempo de espera del cliente, etc.) e, incluso, suelen presentarse algunos que no cumplen con lo pretendido (producción defectuosa o no conforme).

¿Por qué se presenta variabilidad en los procesos? En todos ellos actúan múltiples factores que no se mantienen constantes: las condiciones de operación cambian, los materiales o insumos pueden presentar calidad variable, los instrumentos de medición introducen incertidumbre, al igual que el

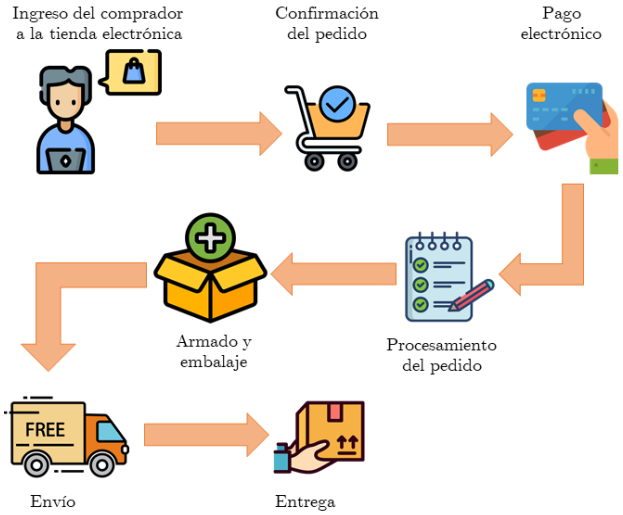


Figura 1.3 *El proceso de entrega de productos comprados en línea*

entorno externo. Además, estos factores pueden variar en un momento dado y también a través del tiempo. Si se piensa en un proceso de producción de bienes o servicios, los diferentes factores que lo afectan se pueden agrupar en las seis categorías que se presentan en la Figura 1.4.

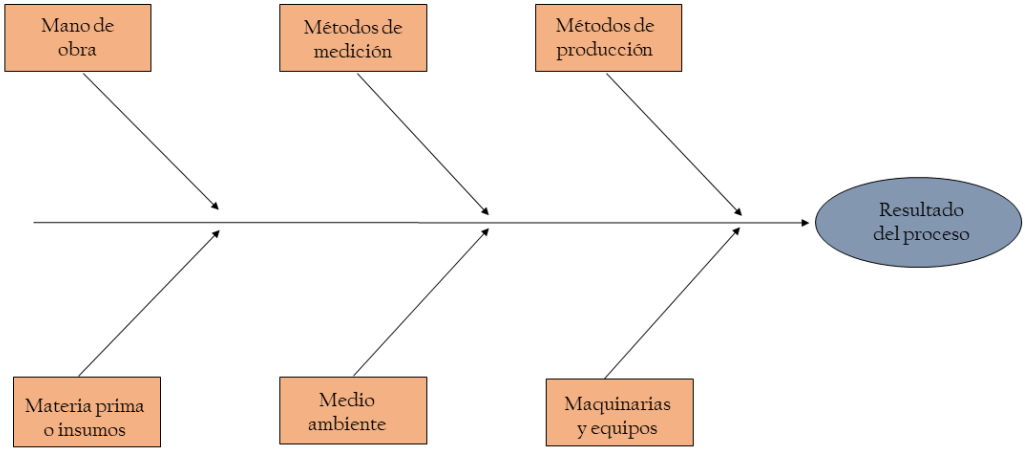


Figura 1.4 *Categorías en las que se pueden agrupar los factores de variabilidad para cualquier proceso de producción de bienes o servicios*

El esquema presentado en la Figura 1.4 es una versión básica de un "Diagrama de Espina de Pescado" o "Diagrama de Causa/Efecto" que fue perfeccionado por el japonés Kaoru Ishikawa [1]. Este diagrama pone de manifiesto **el sistema de causas o factores que producen variabilidad en las características de interés** de un proceso. Además, es válido no solo para los procesos de producción de bienes o servicios sino para cualquier proceso en el que puedan identificarse factores causales y características

[1] Kaoru Ishikawa (1915-1989) fue uno de los principales exponentes de la Calidad en Japón y uno de los principales referentes mundiales de la Calidad, junto con Deming y Juran. Tan grande fue su aporte a la mejora de la calidad que la Sociedad Americana para la Calidad (ASQ) creó en su honor un premio anual denominado Medalla Ishikawa, que se otorga a una persona o equipo que haya desempeñado un papel excepcional en la mejora de la calidad.

observables en el proceso mismo o en alguna de sus salidas. Puede afirmarse entonces que las mediciones, el diseño de equipos o sistemas, la administración de recursos humanos, la gestión de compras, el mantenimiento de maquinarias y el diseño de carreteras, entre otros, constituyen también procesos. Según de qué proceso se trate, puede ser que algunas de las categorías consideradas en la Figura 1.4 estén presentes, que el número de factores sea pequeño o muy grande, que el "producto" sea un servicio o cualquier otra salida, y por último, que interese estudiar la influencia de los factores a través del tiempo.

En síntesis, el concepto de **proceso** incluye a la **variabilidad**, la cual se manifiesta en una o varias características del producto o servicio (o de cualquier otra salida) o del proceso mismo. Esta variabilidad genera *incertidumbre* ya que no se pueden predecir con exactitud los valores que asumirán esas características para una salida en particular.

1.3 Problemas en un marco de variabilidad e incertidumbre: el rol de la estadística

Las personas especializadas en ingeniería trabajan en distintos procesos de áreas diversas (industria, construcción, logística, comunicaciones, producción de energía, gestión de recursos, etc.). En estos se presentan habitualmente situaciones, como por ejemplo:

- Devoluciones de pedidos por parte de los clientes.
- Necesidad de diseñar un puente.
- Posibilidad de añadir un nuevo cliente interesado en comprar barras metálicas.
- Evaluación de la calidad de una ruta.
- Evaluación del comportamiento de un sistema de medición.
- Evaluación de las fallas de un sistema de distribución de energía eléctrica.
- Necesidad de medir magnitudes de manera adecuada.
- Elección del dispositivo mecánico más apropiado para un tipo de máquina sembradora.

En estas situaciones surgen **problemas ingenieriles** más o menos complejos que involucran diferentes aspectos (costos y cuestiones financieras, problemas técnicos específicos, factibilidad, etc). En muchos de estos problemas ingenieriles surgen preguntas como, por ejemplo: "¿cuáles son los motivos de devolución más frecuentes en el último mes?", "¿los motivos de las devoluciones de los clientes difieren según la zona?", "¿cuál es la proporción de barras que cumplen con las especificaciones definidas por el nuevo cliente en relación a la longitud?", "¿es frecuente que el caudal del río supere cierto valor en cualquier época del año?", "¿la ruta fue construida con el espesor adecuado?", "¿el sistema de medición que se utiliza es adecuado?"[2], etc. Estas preguntas constituyen problemas de naturaleza estadística, los cuales deben ser resueltos a partir de la realización de estudios estadísticos. Las conclusiones que se obtengan del estudio estadístico (conclusiones estadísticas) se deben considerar junto con los otros aspectos mencionados (conclusiones en contexto) para dar una solución integral al problema ingenieril.

[2]El caso de las mediciones es muy particular, por varias cuestiones que se irán desarrollando en los diferentes capítulos. Por ahora, es importante señalar que el estudio del procedimiento de medición puede ser un problema en sí mismo o estar vinculado a otro problema donde interese conocer el comportamiento de alguna característica de interés. Por ejemplo, al estudiar el comportamiento del espesor de una ruta, es bueno tener confianza en los valores obtenidos con el método de medición empleado. Ahora bien, ¿en qué casos se podrá confiar? En la Sección 1.5.3 se comienza a abordar este tema y se continua en capítulos posteriores.

Un **problema de naturaleza estadística** es una pregunta o conjunto de preguntas que se plantean en un marco de variabilidad. Las preguntas buscan indagar sobre el comportamiento de características o variables de interés, relativas a ciertas poblaciones. En el proceso de resolución de este tipo de problemas, se requieren datos.

Este texto se ocupa de la resolución de problemas de naturaleza estadística.

Para resolver el problema de naturaleza estadística (en adelante, el problema), el primer paso consiste en plantearlo en forma precisa, lo que implica definir, entre otras cosas, la/las característica/s de interés (el motivo de las devoluciones, las dimensiones de las piezas que son de interés para el nuevo cliente, el caudal del río, los resultados de las mediciones, el espesor de la ruta, etc.) y el conjunto de unidades para el cual se quieren obtener conclusiones.

El problema o pregunta a responder, puede estar asociado a un momento dado del proceso o al comportamiento del mismo a través del tiempo. En el primer caso, el tiempo no constituye un factor de interés, como por ejemplo en el estudio de los motivos de las devoluciones de los clientes habituales en el último mes. En el segundo caso, en cambio, el tiempo constituye un factor principal e interesa estudiar cómo se comporta el proceso a través de él. Puede ocurrir que ese comportamiento sea estable o no:

- El proceso se comporta de manera estable a través del tiempo si actúan siempre las mismas causas y lo hacen de manera similar. La variabilidad en la o las características en estudio se denomina **variabilidad natural**. En ese caso, las frecuencias con las que se presentan los distintos valores tienden a repetirse y el modelo matemático o distribución pone de manifiesto el **patrón de variabilidad** de dicha característica.
- El proceso no se comporta de manera estable si algunos factores actúan ocasional y fortuitamente en el proceso. La variabilidad se denomina **variabilidad asignable**. En este caso, el modelo se va modificando, por lo que esta variabilidad debe ser identificada y preferentemente eliminada.

En el caso del posible nuevo cliente, se requiere conocer el conjunto de valores que asume alguna característica de interés en las barras metálicas (por ejemplo, su longitud), para saber si se puede cumplir, actualmente y en el futuro, con lo que este pretende. Para esto es importante que el proceso se comporte de manera estable a través del tiempo. En el caso del diseño del puente, entre otras cosas, se requieren datos históricos del caudal del río para definir un modelo del comportamiento que permita, por ejemplo, hacer predicciones del mismo para distintos períodos del año. No se puede estabilizar ese proceso natural pero es importante conocer la variabilidad del caudal a través del tiempo y evaluarla para tomar decisiones.

En este texto, se abordan problemas en los cuales no se va a estudiar el comportamiento de las características de interés a través del tiempo, ya sea porque no interesa o porque dicho comportamiento se supone estable. En la resolución de estos problemas se requiere el estudio del comportamiento de alguna característica (modelo o distribución) y es necesario entonces llevar adelante una investigación empírica para obtener la información pertinente.



¿Qué significa contar con información pertinente?

Quienes trabajan en ingeniería tratan con datos en forma constante; pero estos no se convierten en información por sí solos. No se trata de tener grandes cantidades de datos que no respondan a ninguna pregunta de interés, ni de llenar planillas, registros y reportes porque sí. Mientras los datos pueden definirse como resultados, numéricos o no, asociados a un hecho, la información, que incluye datos, debe entenderse como la respuesta a preguntas previamente planteadas.

Ahora bien, ¿qué datos tomar para obtener la información necesaria?, ¿cuántos?, ¿cómo tomarlos y registrarlos adecuadamente?, ¿qué análisis realizar para arribar a conclusiones satisfactorias?

La **estadística** cumple un rol fundamental en la resolución de problemas de naturaleza estadística desde el planteo mismo de la pregunta inicial, y proporciona métodos para obtener, organizar y analizar datos que se convierten luego en información de utilidad, y herramientas para la toma de decisiones en presencia de variabilidad e incertidumbre.

Además de los métodos y procedimientos aportados por la estadística, el **pensamiento estadístico** aporta la manera de razonar y comprender las etapas que deben seguirse en la resolución del problema, entre las que se destacan la construcción y obtención de los datos necesarios, su análisis y la adecuada conexión entre las conclusiones estadísticas y el contexto en el cual estas se aplican.

1.4 Problemas de naturaleza estadística para trabajar

A continuación se presentan situaciones problemáticas con las que se trabajará a lo largo del libro.



Problema 1: *Una empresa metalúrgica de la provincia de Santa Fe fabrica distintos tipos de barras de acero utilizadas en la industria. Una automotriz de la región está interesada en adquirir uno de estos productos, teniendo como requisito que la longitud de las barras se encuentre dentro del intervalo (250 ± 0,6) mm.*
Con el objetivo de convertirse en proveedora de dicha automotriz, la gerencia de la empresa necesita determinar qué porcentaje de barras cumple con las especificaciones de longitud establecidas.



Problema 2: *Una distribuidora de madera para la industria de la construcción comercializa placas de gran tamaño y prepara lotes de 80 unidades para sus clientes. Uno de ellos tiene requisitos estrictos respecto a la calidad superficial de las placas y exige que la cantidad promedio* [3] *de defectos (poros, rayaduras, etc.) por placa en cada lote sea inferior a 1,2.*
Al momento de recibir un pedido por parte de este cliente, en la distribuidora cuentan con un lote ya preparado y desean saber si pueden enviárselo o no.



Problema 3: *Se diseñó un servidor neumático de aspiración para mejorar la siembra mecánica de granos de maíz. Este dispositivo reemplaza los componentes originales de selección de semillas (placas perforadas) de las sembradoras convencionales. Con este sistema, las semillas se depositan en ranuras mediante aspiración y luego se siembran a distancias más uniformes. Se espera que el uso de este servidor neumático mejore la efectividad de la siembra de maíz en dos aspectos críticos:*

[3] El promedio y otras medidas se desarrollarán en el Capítulo 2.

- *Reduciendo la variabilidad en las distancias entre semillas consecutivas.*
- *Reduciendo la proporción de semillas que caen a menos de 3 cm unas de otras, situación problemática ya que en estos casos solo germina una semilla y las demás se desperdician.*

Interesa comparar el desempeño de ambos tipos de sembradoras.



Problema 4: *En una compañia de servicios que comercializa productos en línea se ha producido un incremento de las devoluciones de pedidos por parte de sus clientes, situación que aumenta sus costos operativos y afecta negativamente su reputación como vendedora.*
Frente a esta situación, la gerencia propone identificar el motivo principal de cada devolución para detectar las causas más frecuentes e implementar posteriormente acciones correctivas que mejoren el servicio y reduzcan la pérdida de clientes.
En particular, la gerencia sospecha que una de las principales causas podría ser el retraso en las entregas debido a problemas en el transporte. Se ha establecido que, si esta causa representa más del 30 % de las devoluciones, se procederá a contratar a una nueva empresa de logística. Para las demás causas identificadas, se implementarán mejoras de acuerdo con los resultados del estudio.



Problema 5: *Un fabricante de cables eléctricos comercializa sus productos en rollos. Un cliente muy importante está interesado en adquirir rollos de 100 m de cable unipolar de 2,5 mm y exige que al menos el 90 % de ellos presenten dos imperfecciones o menos.*
El Departamento de Calidad quiere determinar si se cumplen estos requisitos.



Problema 6: *Un comitente solicitó la evaluación de un tramo de ruta de 25 km recientemente ejecutado, con el objetivo de verificar el cumplimiento de los criterios de calidad establecidos. En particular, desea evaluar el espesor de la capa de pavimento y la resistencia a la compresión del material utilizado.*

- *Sobre el espesor de la capa, el comitente exige que en promedio sea como mínimo de 22 cm y que no existan puntos con espesor inferior a 19 cm.*
- *Respecto a la resistencia a la compresión del material, el comitente admite como máximo un 5 % de puntos con resistencia menor o igual a 30 MPa.*

Además, el comitente desea analizar la variabilidad de ambas características (espesor y resistencia a la compresión) a lo largo del tramo evaluado.



Problema 7: *Una empresa utiliza un multímetro digital para medir tensión en corriente continua y resistencia eléctrica. Periódicamente, la empresa envía el equipo a un laboratorio especializado para su calibración.*
En dicho laboratorio realizan los ajustes de calibración correspondientes y luego, verifican el desempeño del multímetro. Para ello realizan 15 mediciones repetidas sobre una fuente de voltaje estable con un valor de referencia de 12 V DC y 15 mediciones repetidas sobre una resistencia patrón de 100 Ω, calibrada y trazable a un patrón nacional de referencia.
Con los resultados obtenidos, el laboratorio elabora un informe técnico de desempeño que envía a la empresa dueña del multímetro y que sirve como certificado de garantía del equipo por un período de 6 meses. [4]

Problema 8: *Una empresa está interesada en adquirir una gran cantidad de dispositivos electrónicos de cierto tipo, para construir circuitos especiales. Por razones de seguridad, cada dispositivo sería reemplazado al ocurrir su primera falla. Dado que la tarea de reemplazo es compleja, la empresa requiere que la mayoría de los dispositivos fallen por primera vez después de 150 h de funcionamiento.*
Antes de concretar la compra, desea verificar si se cumple este requisito.

Problema 9: *Una empresa fabrica productos de limpieza. Su producto estrella es un desengrasante para cocina que se comercializa en envases de 700 ml.*
Los ingenieros a cargo del proceso de producción necesitan analizar el comportamiento de la demanda trimestral de este producto durante los últimos 5 años, con el objetivo de planificar de manera óptima la producción del próximo año.

Una vez que se planteen claramente los problemas, ¿qué otras tareas se deberán realizar en el proceso de resolución?

1.5 El ciclo para la resolución de problemas de naturaleza estadística

La resolución de problemas de naturaleza estadística, es decir, de problemas en los que están involucradas la variabilidad y la incertidumbre que esta genera, constituye en sí misma un proceso que se inicia con una adecuada formulación o planteo del problema a resolver y culmina con la obtención de conclusiones que deben relacionarse al contexto. Las etapas del proceso pueden resumirse, por ejemplo, en el Ciclo Investigativo PPDAC, propuesto por Wild y Pfannkuch (1999), aunque muchos autores lo mencionan sin darle ese nombre específico. Las etapas se esquematizan en la Figura 1.5.

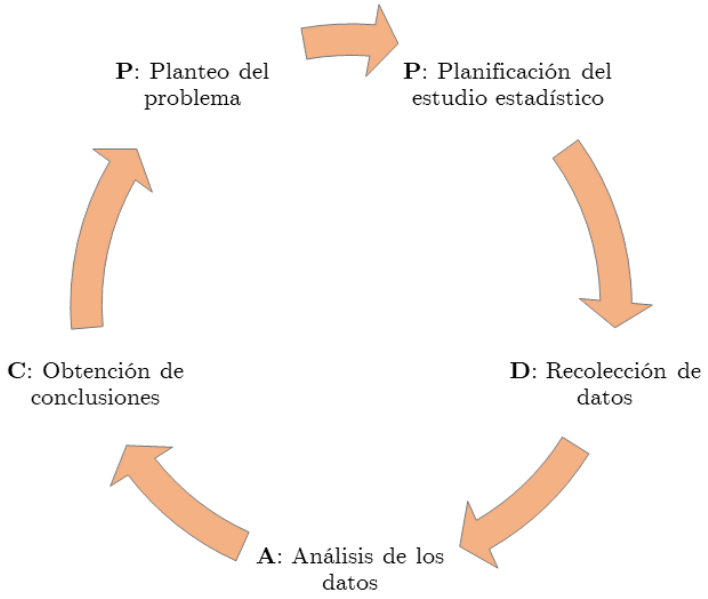


Figura 1.5 *El ciclo PPDAC para la resolución de problemas*

[4]La evaluación del desempeño de un equipo o método de medición se desarrollará en capítulos posteriores.

A continuación, se enuncian los principales conceptos estadísticos asociados al ciclo PPDAC y se describen las acciones que, generalmente, se llevan a cabo para cada etapa.

1.5.1 Etapa de Planteo del problema (P)

El problema se expresa a partir de una o más preguntas y por lo tanto, el planteo adecuado de las mismas es fundamental para el desarrollo de las etapas restantes y para dar finalmente una respuesta, es decir, para resolver el problema. Aunque se obtenga una buena respuesta, esta carece de valor si la pregunta fue mal formulada. En la formulación de las preguntas se requiere definir las **poblaciones** objeto de estudio y las **unidades** que las componen; así como las **variables** que se medirán y los **parámetros** de interés si correspondiere.

1. Conceptos estadísticos asociados

La **población** es el conjunto de todas las unidades, es decir, el universo respecto del cual se quiere obtener conclusiones o tomar decisiones. A cada una de las unidades que componen la población se la denomina **unidad elemental**. Las unidades pueden ser objetos o personas (unidades tangibles) o bien, períodos de tiempo, mediciones, etc. (unidades intangibles).

Según la cantidad de unidades que la componen, una población puede clasificarse en **finita** o **infinita**. En el caso de poblaciones finitas, el total de unidades elementales, o tamaño de la población, se simboliza con N.



Problema 1 (pág. 7): *Todas las barras de acero de cierto tipo que produce la empresa constituyen la población y cada una de las barras, una unidad elemental. Se trata de unidades tangibles.*
El tamaño de esta población se considera infinito ya que incluye todas las barras producidas y las que seguirán produciéndose mientras el proceso mantenga las mismas condiciones.



Siempre que se defina a la población como la totalidad de unidades de un proceso de producción, esta incluye a las unidades ya producidas, las que se están produciendo y las que se producirán en el futuro, bajo las mismas condiciones de operación. En este caso se considera que dicha población es infinita.



Problema 2 (pág. 7): *Todas las placas de madera de un lote constituyen la población y cada una de las placas, una unidad elemental. En este caso, se trata de unidades tangibles.*
Esta población es finita y su tamaño es $N = 80$.



Problema 7 (pág. 8): *Todas las mediciones que se pueden realizar, con cierto multímetro, de una resistencia conocida constituyen la población y cada una de las mediciones, una unidad elemental.*
Para este problema, hay una segunda población compuesta por todas las mediciones que se pueden realizar, con dicho multímetro, de una tensión conocida.
En ambos casos, se trata de unidades intangibles.
El tamaño de estas poblaciones se considera infinito.

En ocasión de definir la población, es muy importante determinar claramente los **criterios de inclusión**, es decir, el conjunto de condiciones que deben cumplir las unidades para formar parte de ella. Entre estos criterios deben estar cuestiones relativas al espacio y al tiempo: las unidades que componen la población deben estar asociadas a un determinado espacio (un sector de la empresa en la cual se producen, una localidad, etc.) y, también, a un determinado periodo de tiempo (en el que fueron producidas, comercializadas, etc.).



Problema 1 (pág. 7): *Suponga que la empresa produce las barras de interés en tres plantas de características similares, ubicadas en diferentes localidades de la provincia de Santa Fe. En ese caso, para definir adecuadamente a la población hay que revisar el objetivo. Pueden darse, por ejemplo, las siguientes situaciones:*

- *Si interesa conocer el comportamiento de las barras producidas en las tres plantas, la población está integrada por la totalidad de barras producidas por la empresa.*
- *Si interesa conocer el comportamiento de las barras producidas solo en la planta ubicada en Venado Tuerto, la población está integrada únicamente por las barras producidas por la empresa en dicha planta.*

Una situación análoga a la mencionada en relación a diferentes plantas, podría plantearse si hubiera barras producidas por diferentes turnos, máquinas, etc.



Problema 1 (pág. 7): *Si la población está compuesta por la totalidad de las barras, se entiende que se trata de las barras que se vienen produciendo (y de las que se seguirán produciendo si el proceso continúa con un comportamiento estable).*
Pero, considere la siguiente situación:
A partir del 1 de marzo del corriente año se introdujeron modificaciones importantes en el proceso de producción en las tres plantas e interesa conocer el comportamiento de las barras que se producen luego de la implementación de dichas modificaciones. Entonces, la población está formada por la totalidad de barras de cierto tipo que se producen en las tres plantas a partir del primer día de marzo del corriente año.

Un criterio de inclusión de alguna manera define **criterios de exclusión**, es decir, condiciones tales que, si las unidades las cumplen, impiden que estas formen parte de la población de interés.



Problema 1 (pág. 7): *Si solo interesan las barras que se producen en la planta de Venado Tuerto, se tiene que el criterio de inclusión es que la barra sea fabricada en dicha planta y el criterio de exclusión es que la barra sea fabricada en cualquiera de las otras dos plantas. En este caso, solo se explicita el criterio de inclusión y el de exclusión se define implícitamente; pero podría ocurrir lo contrario. Suponga que hubo problemas en la planta ubicada en Gálvez y, por el momento, no interesa estudiar barras que se producen en dicha planta. La población está compuesta entonces por las barras que se producen en la empresa, excepto las de la planta de Gálvez.*



Observe la importancia de definir a la población de acuerdo al objetivo planteado, considerando los criterios de exclusión/inclusión que correspondan.

Una **variable** es cualquier característica que puede tomar diferentes valores (números o categorías) en las unidades elementales.

Las variables se clasifican en cuantitativas o cualitativas, según los valores que pueden asumir.

- Las variables **cualitativas** o **categóricas** clasifican a las unidades elementales en categorías o niveles. Por ejemplo, "tipo de defecto", "condición de defectuosa de una pieza","calificación de una variedad de miel" y "preferencia de un consumidor en relación a una familia de productos".
- Las variables **cuantitativas** le asignan a cada unidad elemental un número. Estas, a su vez, se pueden clasificar en discretas y continuas.
  - Para las variables **discretas**, el conjunto de valores posibles es finito o infinito numerable y se asocia, generalmente, al conjunto $\mathbb{N}_0$. Son variables discretas "cantidad de poros en la superficie de una placa", "cantidad de piezas defectuosas en una caja de 20 unidades" y "cantidad de llamadas que ingresan a un conmutador en un período de 10 minutos".
  - Para las variables **continuas**, el conjunto de valores posibles es un intervalo o conjunto de intervalos de números reales. Son variables continuas "longitud de una pieza metálica", "volumen de lluvia caído en un mes", "dureza de una barra" y "temperatura de salida de un polímero en un instante".

Las variables pueden medirse utilizando diferentes escalas o niveles de medición, que se describen a continuación:

- **Nominal**: los valores que puede asumir la variable son diferentes entre sí pero no hay razón para ordenarlos, es decir, para pensar que uno es de menor o mayor jerarquía que otro. Por ejemplo, para la variable "tipo de defecto" se registra si se trata de un poro, una rayadura, una grieta, etc. o para la variable "preferencia del consumidor en relación a una familia de productos" se registra si el consumidor prefiere un producto dietético, vegano, sin gluten, etc. Ambas variables están medidas en escala nominal.
- **Ordinal**: los valores que puede asumir la variable son diferentes entre sí y, además, se pueden ordenar según alguna jerarquía; pero no se pueden evaluar las diferencias entre uno y otro. Por ejemplo, para la variable "calificación de una variedad de miel" se registra si esta es muy mala, mala, regular, buena o muy buena. Se trata entonces de una variable medida en escala ordinal.
- **De intervalo**: los valores que puede asumir la variable son diferentes entre sí, se pueden ordenar según alguna jerarquía y además se pueden hacer diferencias entre ellos; pero no cocientes ya que el valor 0 no indica ausencia de dicha característica. Por ejemplo, en el caso de la variable "temperatura" medida en grados centígrados, una temperatura de 0 grados no indica ausencia y por ese motivo, esta variable se considera medida en escala de intervalo.
- **De razón**: los valores que puede asumir la variable son diferentes entre sí, se pueden ordenar y además se pueden hacer diferencias y cocientes entre ellos, ya que el valor 0 sí indica ausencia de la característica. Por ejemplo, en el caso de variables como "longitud" o "cantidad de poros", el valor 0 indica ausencia de la característica que se mide y por lo tanto ambas variables están medidas en escala de razón.

En la Tabla 1.1 se resumen las operaciones que pueden realizarse con los datos obtenidos, según la escala de medición de la variable. Como se deduce de la cantidad de operaciones asociadas a cada nivel de medición y se observa en la Tabla 1.1, los niveles están ordenados de manera creciente y cada uno de ellos contempla operaciones de los anteriores.

Tabla 1.1 *Operaciones que pueden llevarse a cabo según el nivel de medición de la variable*

Nivel de medición	Diferenciar	Ordenar	Obtener diferencias	Obtener cocientes
Nominal	X			
Ordinal	X	X		
De intervalo	X	X	X	
De razón	X	X	X	X



Problema 1 (pág. 7): *Después de medir la longitud de las barras, puede interesar:*

- *si la barra cumple o no cumple con las especificaciones definidas por cierto cliente, en relación a su longitud. La variable se puede denominar "cumplimiento de las especificaciones" y en este caso se trata de una variable cualitativa, medida en escala nominal.*
- *si la barra es pequeña (mide menos de 240 mm), mediana (mide entre 240 mm y 260 mm) o grande (mide más de 260 mm). La variable se puede denominar "tamaño" y en este caso se trata de una variable cualitativa, medida en escala ordinal.*
- *el valor de su longitud, en mm. La variable se puede denominar "longitud" y en este caso se trata de una variable cuantitativa, medida en escala de razón (ya que el valor 0 no es un valor arbitrario sino que indica ausencia de longitud).*



Cuando se trata de variables cualitativas o categóricas, los valores que asume la variable son, en realidad, modalidades, categorías o niveles diferentes y no se expresan numéricamente. Puede ocurrir que a estas categorías se les asignan números (códigos) para facilitar la carga de los datos en alguna planilla; pero esto no convierte a la variable en cuantitativa.



Problema 4 (pág. 8): *En este problema se registra el "motivo principal de devolución" para un conjunto de devoluciones de pedidos realizados en línea. Las categorías para esa variable son: (1) Errores en la facturación, (2) Pedido equivocado, (3) Precio final no coincidente con el que figura en la página, (4) Producto dañado, (5) Producto mal embalado, (6) Producto vencido o en mal estado y (7) Retraso en la entrega por problemas del transporte. En la planilla de datos se registra, para cada devolución, un número entre 1 y 7, según el motivo principal de la misma; pero este hecho no convierte a la variable en cuantitativa; sigue siendo cualitativa, medida en escala nominal.*



La mayoría de las variables cuantitativas que se analizan en este texto están medidas en escala de razón. Interesa diferenciar si son discretas o continuas.



Problema 1 (pág. 7): *En este problema, la característica de interés es la longitud, variable cuantitativa continua, medida en escala de razón, por cuanto el valor 0 no es arbitrario sino que indica ausencia de longitud. Para la misma población de barras, podrían haberse observado otras variables. En la Tabla 1.2 se mencionan y clasifican algunas variables posibles.*

Tabla 1.2 *Clasificación de posibles variables a medir en barras de acero*

Variable	*Clasificación*	*Valores posibles*[5]	*Nivel de medición*
Cumplimiento de la especificación	*Cualitativa o categórica*	*Sí, No*	*Nominal*
Tipo de acero		*Al carbono, al silicio, al manganeso, etc.*	*Nominal*
Nivel de calidad		*Bajo, Medio, Alto*	*Ordinal*
Nº de imperfecciones	*Cuantitativa discreta*	*0, 1, 2, 3, ...*	*De razón*
Peso	*Cuantitativa continua*	*(0, +∞)*	*De razón*
Temperatura máxima que soporta		*(0, +∞)*	*De intervalo*
Longitud		*(0, +∞)*	*De razón*

No se deben confundir la cantidad de unidades elementales o tamaño del conjunto de los datos con la cantidad de niveles o valores que puede asumir la variable. Por ejemplo, la primera variable solo asume dos valores diferentes; mientras que la población en estudio está compuesta por todas las barras que se producen, que representan las unidades elementales.

En muchas ocasiones, la variable de interés no se puede medir directamente o bien, resulta muy costoso hacerlo. Entonces, se puede definir una **variable "sustituta"** que brinde la misma información (o similar). Por ejemplo, para saber si un conjunto de piezas fabricadas con cierta aleación tienen la dureza adecuada, se debería medir la dureza en cada una y luego indicar si esta es adecuada o no, según un criterio preestablecido. Pero los ensayos de dureza son muy costosos y en la empresa no están dispuestos a correr con ese costo. Un especialista sugiere que, cuando la dureza no es la adecuada, la pieza presenta grietas, lo cual es observable a simple vista. La variable **dureza de una pieza** cuyos resultados posibles son "adecuada" y "no adecuada", se sustituye entonces por la variable **presencia de grietas en una pieza** cuyas modalidades son "no presenta" (lo que indica dureza adecuada) y "sí presenta" (lo que indica dureza no adecuada).



Problema 9 (pág. 9): *En la empresa fabricante de productos de limpieza están interesados en conocer el comportamiento de la demanda trimestral de su producto estrella, durante los últimos 5 años, para planificar la producción para el próximo año.*
*La variable de interés, **demanda trimestral**, se refiere a la cantidad de unidades del producto demandadas por los clientes en cada trimestre. Esta variable es muy difícil de observar, ya que no se suelen registrar los casos en que un cliente demanda el producto y no lo consigue. Una variable sustituta puede ser **venta trimestral**, que se refiere a la cantidad de unidades del producto vendidas en cada trimestre. Sobre esta variable es mucho más simple obtener la información requerida.*

Haciendo una síntesis de lo visto hasta ahora, se puede decir que si la pregunta es "¿**cuáles** son los elementos bajo estudio?", surge en primer lugar el concepto de **unidad elemental** y luego el de **población**, como la totalidad de las unidades elementales, asociadas a un objetivo. Y si la pregunta es "¿**qué** características interesa estudiar en dichas unidades?", surge el concepto de **variable**.

La **población estadística** es el conjunto de todos los valores (números o categorías) que asume la variable de interés cuando se mide en cada unidad de la población.

En la definición de población estadística se consideran conjuntamente los conceptos de población y variable vistos anteriormente.

Problema 1 (pág. 7): *Para la población de barras de este problema, si la variable de interés es su longitud, la población estadística es el conjunto de las longitudes de dichas barras. Pero si la variable de interés es el cumplimiento de las especificaciones (se considera que una barra cumple con las especificaciones si su longitud es un valor entre 249,4 mm y 250,6 mm), la población estadística es el conjunto de las condiciones (en relación al cumplimiento de las especificaciones) de dichas barras. Es decir que al definir distintas variables para una misma población (en este caso, el conjunto de todas las barras producidas por ese proceso), surgen diferentes poblaciones estadísticas, las cuales pueden ser univariadas o multivariadas, según la cantidad de variables que interese estudiar simultáneamente. En la Figura 1.6 se presentan dos poblaciones estadísticas univariadas, asociadas a la población de barras producidas en la empresa metalúrgica, a modo de ejemplo.*

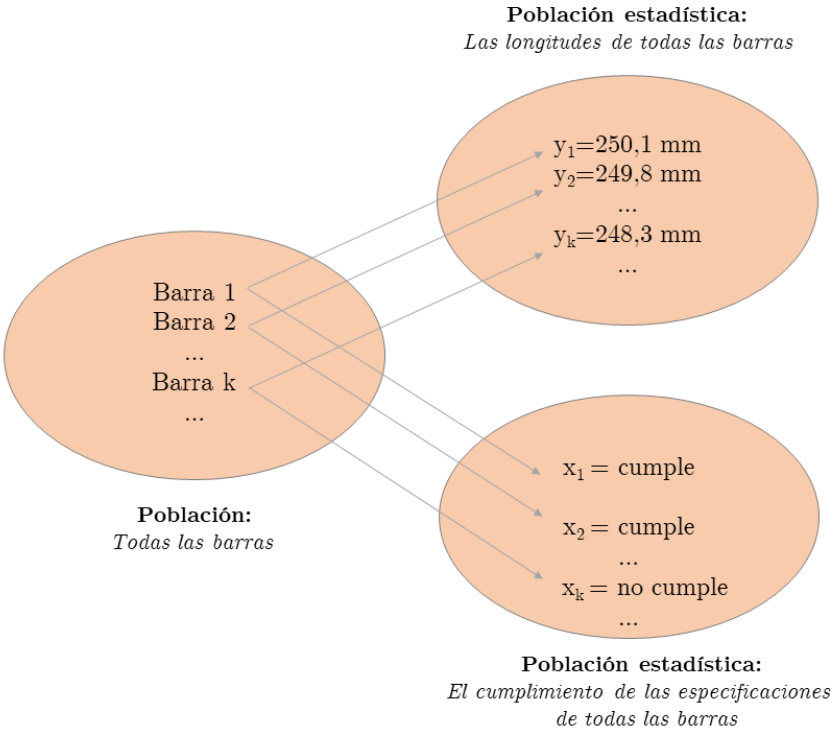


Figura 1.6 *Poblaciones estadísticas asociadas a la población de barras del Problema 1*

Cualquier medida que resuma información de la población se denomina **parámetro**.

Generalmente, los parámetros se simbolizan con letras griegas. Por ejemplo, la proporción poblacional se simboliza con la letra π, el promedio poblacional se simboliza con la letra μ y la desviación estándar poblacional se simboliza con σ [6].

[6] Estas y otras medidas de interés se van a definir en el Capítulo 2.



Problema 1 (pág. 7): *Todas las barras de acero producidas por la empresa constituyen la población y una medida que resume información de dicha población es la proporción de barras con longitudes entre 249,4 mm y 250,6 mm (π). También puede interesar la longitud promedio (μ) y la desviación estándar de la longitud (σ).*



Con lo visto hasta ahora, se pueden resolver las Actividades 1, 2 y 3-a, Sección 1.7.

2. Tareas que se realizan en la etapa de Planteo del problema

El primer paso para la resolución de un problema estadístico es plantearlo correctamente, es decir, formular las preguntas que se desean responder. En ellas debe estar claramente definido lo siguiente:

- la/s poblacion/es;
- la unidad de elemental;
- la/s variable/s;
- la/s población/es estadística/s y
- el/los parámetro/s de interés.

Observe que, en esta primera etapa, el problema se refiere a una población determinada que debe ser definida claramente, con los criterios de inclusión y/o exclusión que correspondan. Definidas la población y las variables de interés, así como los parámetros, el planteo del problema puede reescribirse en función de estos elementos. Es decir, que se realiza una traducción o modelado de la situación problemática. El objetivo del estudio estadístico que se llevará a cabo es dar respuesta a las preguntas formuladas al inicio, es decir, dar respuesta al problema.

El objetivo puede involucrar:

- el conocimiento del comportamiento de una característica en la población (modelo matemático o distribución);
- el conocimiento de alguna/s medida/s que resuman información acerca de la población estadística en estudio (parámetro).

El conocimiento del modelo matemático y/o de algún parámetro puede servir para la toma de decisiones sobre el proceso, si esto es lo que se busca.



Problema 1 (pág. 7): *La pregunta que se desea responder es:* ***¿Cuál es la proporción de barras de acero de cierto tipo que cumple con las especificaciones fijadas por la empresa automotriz en relación a su longitud?*** *Se define:*

- ***Población****: la totalidad de las barras de acero de cierto tipo que produce la empresa (población infinita).*
- ***Unidad elemental****: cada barra.*
- ***Variables****: longitud de una barra (en mm) (variable cuantitativa continua, medida en escala de razón) y cumplimiento de las especificaciones por parte de una barra (variable cualitativa, medida en escala nominal).*

- ***Poblaciones estadísticas**: las longitudes de todas las barras de acero de cierto tipo que produce la empresa y el cumplimiento de las especificaciones de todas las barras.*
- ***Parámetro de interés**: proporción de barras que cumplen con las especificaciones (π), es decir, proporción de barras con longitudes entre 249,4 mm y 250,6 mm.*

Dadas estas definiciones, el objetivo del estudio estadístico es conocer la proporción de barras de acero de cierto tipo que cumple con las especificaciones fijadas por la empresa automotriz en relación a su longitud.



Problema 2 (pág. 7): *La pregunta que se desea responder es: **¿Se puede enviar al cliente el lote de 80 placas de madera que está disponible en la distribuidora?** Se define:*

- ***Población**: todas las placas de madera que contiene el lote (población finita, N = 80).*
- ***Unidad elemental**: cada placa.*
- ***Variable**: Cantidad de defectos (poros, rayaduras, etc.) en la superficie de una placa (variable cuantitativa discreta, medida en escala de razón).*
- ***Población estadística**: las cantidades de defectos de todas las placas del lote.*
- ***Parámetro de interés**: cantidad promedio de defectos por placa (μ).*

Dadas estas definiciones, el objetivo del estudio estadístico es conocer la cantidad promedio de defectos por placa para el lote disponible y ver si esta es menor que 1,2 o no, para luego decidir, considerando información de contexto, si enviar o no ese lote al cliente.



Problema 3 (pág. 7): *La pregunta que se quiere responder es: **El uso del servidor neumático, ¿mejora el proceso de siembra de semillas de maíz?** Ahora bien, ¿qué quiere decir "mejora"?*
En este caso, se definen dos poblaciones:

- ***Población 1**: todos los pares de semillas de maíz que se podrían sembrar de manera consecutiva con una sembradora que no cuenta con un servidor neumático.*
  ***Población 2:** todos los pares de semillas de maíz que se podrían sembrar de manera consecutiva con una sembradora que sí cuenta con un servidor neumático.*
  Ambas poblaciones son conceptuales, porque se va a llevar a cabo un experimento.
- ***Unidad elemental**: cada par de semillas consecutivas (sembradas con y sin el servidor neumático).*
- ***Variable**: distancia entre un par de semillas consecutivas (en cm) (variable cuantitativa continua, medida en escala de razón). A partir de esta variable también se registra una variable cualitativa que indica si una distancia es menor que 3 cm o no.*
- ***Poblaciones estadísticas**: las distancias entre todos los pares de semillas que se podrían sembrar consecutivamente con y sin el servidor neumático.*
- ***Parámetros de interés**: distancia promedio (μ); proporción de pares de semillas consecutivas para los cuales la distancia es menor que 3 cm (π); desviación estándar de las distancias (σ).*
  Estos parámetros se definen para cada una de las poblaciones. Por ejemplo, interesan la distancia promedio cuando no se usa el servidor y la distancia promedio cuando sí se lo utiliza.

Suponga que se considera que el dispositivo mejora el proceso de siembra si, cuando este se usa se mantiene constante la distancia promedio pero se reducen la desviación estándar de la distancia entre semillas y la proporción de pares de semillas con distancia menor que 3 cm (comparadas con las medidas correspondientes al caso en el que el dispositivo no se usa).

> *Dadas estas definiciones, el objetivo del estudio estadístico es comparar las distancias promedio, las desviaciones estándar y las proporciones de distancias menores que 3 cm para la siembra hecha con y sin el dispositivo y decidir si conviene usarlo o no.*



Con lo visto hasta ahora, se puede resolver la Actividad 4, Sección 1.7.

1.5.2 Etapa de Planificación del estudio estadístico (P)

Para responder cualquiera de las preguntas que se planteen **se necesitan datos** y, por lo tanto, se debe decidir cómo llevar adelante el estudio estadístico para obtenerlos (observacional o experimental, poblacional o muestral, transversal o longitudinal) y cómo analizar estos datos posteriormente para obtener las conclusiones utilizando herramientas de análisis descriptivo y/o inferencial, según corresponda. Esto remite a cuestiones relativas a la etapa de Planificación del estudio estadístico.

Si el estudio es poblacional, solo se necesita planificar cómo obtener los datos de todas las unidades que conforman la población en estudio y cómo llevar adelante el análisis descriptivo de los mismos para pasar a las conclusiones. En cambio, si se planifica un estudio por muestreo, se debe decidir cómo seleccionar las unidades que conformarán la muestra, cómo obtener los datos de cada una de ellas, cómo realizar el análisis descriptivo de los mismos y qué herramientas de análisis inferencial utilizar para elaborar las conclusiones sobre la población en estudio.

En esta etapa hay muchas cuestiones a considerar, pero todas deben llevar a que el diseño de la investigación sea el más apropiado para la obtención de conclusiones en relación al problema planteado.

1. Conceptos estadísticos asociados

Estudios observacionales vs. estudios experimentales

Un **estudio observacional** es un estudio en el cual se registran los valores de algunas características de interés en las unidades, sin realizar modificaciones en la población o proceso, excepto las necesarias para obtener los datos requeridos. En este tipo de estudio no se intenta manipular ni modificar las unidades.

En todo estudio observacional, muestral o poblacional, las unidades elementales se denominan **unidades observacionales**.

Un **estudio experimental** o experimento propiamente dicho, es un estudio en el cual se hacen cambios en los niveles de algunas variables (factores), es decir, se hacen modificaciones en algunas unidades de la población o en el proceso, con el objetivo de analizar si dichas modificaciones influyen/afectan a los valores de algunas características de interés (variables de respuesta). En todo estudio experimental, las unidades elementales se denominan **unidades experimentales**.

Los experimentos constituyen la mejor manera de valorar el efecto de uno o más factores sobre alguna variable de interés ya que, para analizar la respuesta a un cambio, se trata de controlar todos los factores que presumiblemente influyen sobre ese cambio.



Problema 3 (pág. 7): *En este caso interesa estudiar el efecto de un factor, el dispositivo neumático, con dos niveles "se usa", "no se usa", sobre la variable de interés que es la distancia entre un par de semillas consecutivas. Esto da lugar finalmente, a dos poblaciones estadísticas que se deben comparar: las distancias entre pares de semillas consecutivas cuando se usa el dispositivo y las distancias entre pares de semillas consecutivas cuando no se usa.*
Suponga que, para evaluar si el servidor neumático mejora efectivamente la siembra, se diseñó el siguiente experimento: Se seleccionaron dos lotes de campo con características similares (tamaño, tipo de suelo, humedad, etc.). En uno se realizó la siembra con una sembradora convencional y en el otro con una sembradora equipada con el servidor neumático.
El experimento se realizó bajo las siguientes condiciones controladas: Las sembradoras tenían características similares (modelo, antiguedad, etc.) y fueron programadas para trabajar de la misma manera. La asignación de cada sembradora a cada lote se determinó mediante sorteo. Se prepararon dos bolsas con la misma cantidad de semillas (medida en peso) del mismo tipo y luego se asignaron aleatoriamente a cada sembradora. La siembra en ambos lotes se realizó el mismo día y bajo las mismas condiciones climáticas.
Como se puede observar, algunos factores que podían influir en los resultados del experimento se mantuvieron constantes: las características de los dos lotes de campo, de las dos máquinas sembradoras y de las semillas que se utilizarán, las cuestiones climáticas, etc. Respecto de otros factores que podrían influir, se dejó que el azar "los controle de alguna manera", decidiendo por sorteo qué bolsa de semillas se sembró con cada máquina y qué máquina se utilizó en cada lote.

Estudios poblacionales versus estudios por muestreo

De acuerdo a la cantidad de datos, los estudios observacionales pueden ser muestrales o poblacionales, según se estudie a un subconjunto o a toda la población (censo). Los experimentos siempre se realizan con un número relativamente pequeño de unidades, que se consideran la muestra en estudio. En ese tipo de estudios, la población está compuesta por las infinitas unidades en las que se podría probar el efecto de los factores de interés y, por lo tanto, no tiene existencia física, se trata de una población conceptual.

En un **estudio poblacional** se estudian todas las unidades de una población (estudio exhaustivo o censo). En un **estudio por muestreo** se analiza a un subconjunto de unidades de una población.

Los estudios poblacionales no siempre son posibles de llevar a cabo, ya sea porque la población es infinita o porque los ensayos que deben hacerse para medir la característica en estudio son destructivos o muy costosos. También puede ser porque el proceso (de producción de las unidades o de recolección de los datos) sea muy lento y se demore mucho tiempo en obtener todas las observaciones.

Acerca de las muestras

Una **muestra** es un subconjunto de elementos de la población bajo estudio. Su tamaño o cantidad de elementos en este texto se simboliza con n.

Existen diferentes maneras de seleccionar una muestra; pero es fundamental tener presente que solo a partir de **muestras aleatorias** o **probabilísticas** se pueden extender los resultados a toda la población de manera válida.

¿Qué es una muestra aleatoria o probabilística? La idea básica de una muestra de este tipo es que cada unidad de la población tenga una posibilidad o chance conocida de ser seleccionada para integrar la misma. Dentro de las muestras aleatorias existen diferentes métodos de muestreo (simple al azar, sistemático, por conglomerados, estratificado, multietápico) que se aplican según las características de la población que se desee estudiar. En este texto, se considerará solamente el muestreo simple al azar, que se definirá específicamente en el Capítulo 7.

Además del método, es también importante elegir adecuadamente el tamaño de la muestra ya que ambos influyen en la calidad de las conclusiones que se obtendrán.

En las **muestras por conveniencia**, en cambio, las unidades se incorporan porque se consiguen fácilmente, por voluntad del participante, etc. Estas muestras pueden resultar sesgadas y no representar adecuadamente el patrón de variabilidad de la población en estudio.

Estudios según el rol del factor tiempo

Existen estudios en los cuales interesa conocer el comportamiento de las variables a lo largo del tiempo (en diferentes períodos de tiempo). En cambio, existen otros estudios en los cuales interesa conocer el comportamiento de las variables en un único período de tiempo (en ese caso puede ser que las observaciones se realicen en un mismo momento o en distintos momentos, pero suponiendo que el tiempo transcurrido entre ellos no constituye un factor importante para el objetivo del estudio).



Problema 9 (pág. 9): *En este caso interesa estudiar el comportamiento de la demanda trimestral de desengrasantes durante los últimos 5 años, con el objetivo de describirlo, pero fundamentalmente de predecir valores futuros. Se trata de un estudio en el cual interesa tener en cuenta el tiempo, en este caso, considerando períodos de un trimestre.*

Además del tipo de estudio y de las cuestiones relacionadas con ellos, en la etapa de Planificación del estudio estadístico se deben definir, en la medida de lo posible, los análisis que se llevarán a cabo una vez recolectados los datos. En cualquiera de los casos en los que se trabaje con un subconjunto de unidades de la población (ya sean estudios observacionales o experimentales), en la etapa de Análisis de los datos se requerirá la aplicación de herramientas inferenciales para generalizar los resultados a toda la población, a la que siempre remiten los objetivos del estudio.

Por último se pueden planificar cuestiones relativas a la comunicación de los resultados del estudio estadístico.

2. Tareas que se realizan en la etapa de Planificación del estudio estadístico

En esta etapa, lo primero que se debe decidir es el tipo de estudio que se llevará a cabo:

- ¿Será un estudio observacional o experimental?
- Si se decide que sea observacional, ¿se trabajará con toda la población o se tomará una muestra?
- Si se decide tomar una muestra, ¿de qué tipo y tamaño será?
- Si se decide llevar a cabo un experimento, ¿cómo se diseñará dicho experimento?, ¿con cuántas unidades se experimentará?

Además de seleccionar el tipo de estudio, hay muchas más preguntas a contestar en esta etapa. A continuación se presentan algunas de ellas:

- ¿Cómo se van a medir las variables de interés? ¿Con qué métodos o instrumentos?
- ¿Qué análisis se llevarán a cabo? ¿Será suficiente con el análisis descriptivo? ¿Se requerirán herramientas de inferencia estadística? ¿Cuáles?
- ¿Qué riesgos máximos se admiten en el caso de tener que inferir las conclusiones de la muestra a la población?

Las dos primeras preguntas son comunes a todos los estudios, mientras que la última es propia de los estudios en los que se trabaja con una muestra. Esto incluye a los estudios experimentales, en los que nunca se realizan relevamientos exhaustivos.



Problema 1 (pág. 7): *Como la población (todas las barras de cierto tipo) es un conjunto infinito, se debe decidir cuántas barras se extraerán del proceso (n), cuándo y cómo se registrará su longitud. En este caso caben algunas consideraciones:*

- *¿Será lo mismo extraer 30 o 100 barras, para inferir las conclusiones a toda la población?*
- *¿Cómo se elegirán las barras que van a constituir la muestra? ¿Se tomarán de cualquier turno de trabajo?, ¿de cualquier día de la semana?*
- *¿Cómo se analizarán los datos?*

Cualquiera sea la opción elegida, solo se podrán obtener medidas de resumen de la muestra (estadísticos) y luego habrá que elaborar conclusiones para toda la población. Entonces, ¿con qué riesgo de error se obtendrán dichas conclusiones?



Problema 2 (pág. 7): *En este caso se cuenta con una población finita y, por lo tanto, puede llevarse a cabo un estudio poblacional para todo el lote o también un estudio por muestreo. En el primer caso, se pueden calcular los parámetros de interés y concluir, sin riesgo de inferencias erróneas.*



Problema 3 (pág. 7): *Dado que se trata de un experimento y las poblaciones no tienen existencia física, se deben decidir cuántas semillas de maíz se van a sembrar utilizando el servidor neumático y sin utilizarlo. También se debe pensar en los factores que se pueden mantener constantes, en los mecanismos de asignación aleatoria (sorteo), etc. Es importante definir además cómo se va a medir la distancia entre cada par de semillas consecutivas. Una vez obtenidos los valores se pueden calcular, para ambas muestras, las distancias promedio, las desviaciones estándar y las proporciones de distancias nulas, que son medidas muestrales (estadísticos). Luego habrá que compararlas y elaborar conclusiones respecto de ambas poblaciones, aplicando herramientas de inferencia estadística, de modo que se debe decidir con qué riesgo de error se obtendrán las mismas.*



Es muy importante tener en cuenta que los datos solamente son útiles si su obtención se realizó a partir de una adecuada planificación del estudio estadístico.



Con lo visto hasta ahora, se pueden resolver las Actividades 3-b, 5 y 6, Sección 1.7.

1.5.3 Etapa de Recolección de datos (D)

Planificado el estudio estadístico, la tercera etapa consiste en la recolección de los datos y en el volcado de los mismos en una planilla de registro. Es muy importante tomar todas las precauciones para **obtener datos de calidad**, ya que ellos constituyen la materia prima fundamental para la obtención de conclusiones adecuadas. Datos de mala calidad o datos "basura", producen malas conclusiones, por más que se apliquen herramientas de análisis muy sofisticadas.

1. Conceptos estadísticos asociados

La **planilla de registro** es el instrumento en el cual se vuelcan los datos recabados para cada unidad elemental a medida que se obtienen.

Esta planilla es de gran ayuda en la etapa de Recolección de datos, ya que en ella constan todas las características que van a observarse en las unidades. Puede definirse en soporte papel o digital.

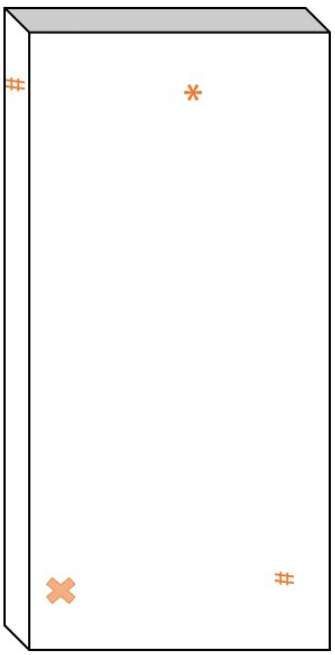
En todos los casos, es deseable que incluya instrucciones que orienten el trabajo de las personas a cargo de dicha tarea. En el caso de una planilla digital, además, es conveniente incorporar controles que eviten posibles errores, por ejemplo, mediante la programación de alertas cuando se ingrese un valor fuera del rango definido para alguna variable.

En ciertas situaciones, puede ocurrir también que se diseñe un mecanismo que permita la generación automática de los datos durante el proceso.

En la Figura 1.7 se presentan algunas planillas de registro a modo de ejemplo.

CONTROL DE DEFECTOS								
Identificación:								
Máquina: Lote: Hoja de ruta:				Fecha: Línea: Operario:				
Total placas revisadas:								
Defectos:								
Tipo:								**Total:**
Poro (X)								
Rayadura (*)								
Grieta (#)								
Otros								
Total general:								
Observaciones:								

(a) *Planilla para el control de defectos en placas de madera (Problema 2).*



(b) *Planilla para la localización de defectos en una placa de madera (Problema 2).*

Figura 1.7 *Ejemplos de planillas de registro utilizadas en formato papel*

Esta planilla puede ser también el formulario de una encuesta con preguntas abiertas y cerradas, a través de las cuales se recolecte la información de cada unidad.

La **trazabilidad** es la posibilidad de identificar, rastrear y recuperar las unidades de las cuales se extrajeron los datos, a través de códigos especialmente elaborados. También permite identificar características de interés en dichas unidades (en qué horario fue producida, con qué máquina o lote de materia prima, con qué equipo se realizó la medición, etc.).

Garantizar la trazabilidad de los datos es de gran utilidad, especialmente cuando aparecen datos erróneos, dudosos o atípicos, ya que permite recuperar las unidades elementales de las cuales se extrajeron y eventualmente repetir las mediciones. Los códigos que identifiquen a las unidades deben registrarse en la planilla.



En muchas ocasiones, el proceso de recolección de datos requiere la realización de mediciones con ciertos procedimientos, que pueden incluir el uso de equipos. Es importante chequear periódicamente estos procedimientos para obtener datos de calidad.

Las mediciones y sus errores

Las mediciones cumplen un papel fundamental en la resolución de problemas de naturaleza estadística, ya que en la mayoría de los estudios que se llevan a cabo para resolver este tipo de problemas, se deben realizar observaciones que implican hacer mediciones con algún procedimiento [7].

Cualquiera sea el procedimiento que se emplee, este puede presentar errores y, en general, los valores medidos son algo diferentes de los valores reales. De hecho, es posible que ninguna medición sea igual que la magnitud real que se busca medir.

La presencia de errores a causa de los procedimientos de medición hace que estos constituyan uno de los factores que puede introducir variabilidad en la característica de interés (Sección 1.2, Figura 1.4) y, por lo tanto, es importante evaluar su desempeño antes de emplearlos en el proceso de recolección de los datos.

El estudio del desempeño de un procedimiento de medición constituye un problema en sí mismo y también puede estar asociado a cualquier otro problema en el cual se requiera realizar mediciones con dicho procedimiento. Si el desempeño no es el adecuado, las mediciones que se obtengan al aplicarlo no serán confiables e indirectamente pueden conducir a conclusiones erróneas sobre el problema a resolver.

En el Capítulo 3 se describe el comportamiento de las mediciones y de los errores y se definen algunas propiedades que permiten evaluar el desempeño de los procedimientos de medición.

[6] La expresión “procedimiento de medición” involucra no solo al equipo que se utiliza para realizar las mediciones sino también al personal que lo utiliza, las formas de utilizarlo, etc.



Problema 7 (pág. 8): *Para este problema, el objetivo del estudio estadístico es evaluar el desempeño de un procedimiento de medición realizado con un multímetro, a partir de la realización de mediciones repetidas de una tensión y de una resistencia eléctrica conocidas. La evaluación de los procedimientos de medición es clave para contar con datos de calidad.*

La **planilla de volcado** es el instrumento en el cual se vuelca la información contenida en todas las planillas de registro.

La planilla de volcado se define en formato digital y se completa una vez finalizada la etapa de recolección de los datos. Es recomendable que tenga la siguiente estructura:

1	x_1	y_1	$\cdots$	$\cdots$	$\cdots$	z_1
2	x_2	y_2	$\cdots$	$\cdots$	$\cdots$	z_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\boldsymbol{n}$ **(o** $\boldsymbol{N}$**)**	x_n	y_n	$\cdots$	$\cdots$	$\cdots$	z_n

En la primera columna de la planilla de volcado se ubican los individuos o unidades elementales analizadas. Cada fila contiene información de un individuo, de modo que la cantidad de filas coincide con el tamaño de la población (N) o de la muestra (n), según corresponda. En las columnas restantes se ubican los valores de las variables estudiadas. El número de variables consideradas en el estudio se simboliza con p, y cada una de ellas se corresponde con una columna.

2. Tareas que se realizan en la etapa de Recolección de datos

En esta etapa se recolectan los datos, es decir, se seleccionan las unidades elementales a estudiar y en cada una de ellas se observan las variables de interés. Entre las tareas que deben realizarse, se pueden mencionar las siguientes:

- redactar un protocolo de recolección de datos, que incluya planillas de registro apropiadas e instructivos para completarlas;
- utilizar esos instructivos para capacitar a todas las personas que vayan a realizar las mediciones o a recabar los datos a través de observaciones, encuestas, etc.;
- hacer una prueba piloto de los métodos de recolección de los datos, por si aparecen cuestiones dudosas, situaciones no previstas, etc.;
- chequear los procedimientos de medición;
- supervisar la tarea de recolección de los datos, así como la tarea de carga de los mismos en planillas electrónicas;
- garantizar, cuando sea posible, la trazabilidad de los datos y
- volcar toda la información recolectada en las planillas de registro, en una planilla de volcado en la que se presenten todos las unidades analizadas en el estudio y para cada una, la información de las variables relevadas.



Con lo visto hasta ahora, se puede resolver la Actividad 3-c, Sección 1.7.

1.5.4 Etapa de Análisis de los datos (A)

A partir de la información contenida en la planilla de volcado, debe hacerse un **tratamiento primario de los datos**, es decir, validarlos, recodificarlos si es necesario, chequear su calidad, etc. Pasado este tratamiento, se lleva a cabo el **análisis descriptivo** y, si corresponde, el **análisis inferencial**, con riesgos asociados.

1. Conceptos estadísticos asociados

Se denomina **tratamiento primario** al análisis que se realiza sobre los datos para chequear su calidad, previo al análisis descriptivo o inferencial.

En el tratamiento primario se pueden detectar datos faltantes, valores que no se corresponden con los posibles de la variable o códigos mal ingresados, entre otros. Si la trazabilidad está garantizada, se pueden recuperar las unidades correspondientes y corregir los errores observados. También pueden utilizarse técnicas de imputación en caso de haber datos faltantes.

El **análisis descriptivo de los datos** consiste en la aplicación de herramientas (tablas, gráficos, indicadores) para resumir o presentar un conjunto de datos, sean estos de una muestra o de una población finita.

Si el estudio es poblacional, las herramientas de análisis descriptivo permiten obtener conclusiones definitivas [8] y dar respuesta al problema planteado, ya que se cuenta con información de todas las unidades de la población y se pueden obtener los valores de los parámetros de interés. En cambio, en el caso de estudios observacionales por muestreo o de estudios experimentales, las conclusiones obtenidas deben inferirse a la población a través de herramientas de análisis inferencial, basadas en información proveniente de una muestra y en el cálculo de estadísticos.

Se denomina **estadístico** a cualquier medida calculada a partir de una muestra que resume información de la misma (por ejemplo, el promedio muestral, la proporción muestral de unidades que cumplen cierta condición o el valor más frecuente en la muestra). Los estadísticos también se distinguen de los parámetros en su notación.

El **análisis inferencial de los datos** consiste en la aplicación de herramientas (intervalos de confianza, pruebas de hipótesis) que permiten extender las conclusiones de una muestra hacia la población, con riesgos controlados. Estas herramientas se apoyan en la Teoría de la Probabilidad.

2. Tareas que se realizan en la etapa de Análisis de los datos

Una vez recolectados los datos y chequeada su calidad, se aplican herramientas de análisis descriptivo, es decir, se construyen tablas y gráficos y se obtienen indicadores o medidas de resumen (estadísticos o parámetros, según corresponda).

[8] El término “definitivas” aplicado a las conclusiones obtenidas del análisis descriptivo en el caso de estudios poblacionales indica que no hace falta aplicar herramientas inferenciales.

- Si se pudo estudiar a toda la población, con este análisis es suficiente para obtener las conclusiones y dar respuesta al problema de naturaleza estadística.
- Si se estudió a un subconjunto de la población, se deben aplicar además herramientas de análisis inferencial para generalizar las conclusiones obtenidas a toda la población, corriendo cierto riesgo de obtener conclusiones erróneas.



Problema 1 (pág. 7): *Para este problema, relativo a las barras metálicas, una vez decidido el tamaño de la muestra y extraídas las barras, se registran sus longitudes. Con ellas se realiza un análisis descriptivo y, finalmente, se obtiene la proporción muestral de barras que satisfacen las especificaciones. Esta medida es un estadístico, ya que se obtiene a partir de una muestra y, entonces, deberán aplicarse herramientas inferenciales para dar respuesta al objetivo en estudio, con algún riesgo de error.*



Problema 2 (pág. 7): *Para este problema, relativo al lote de placas de madera, se pueden plantear dos situaciones diferentes. Si se estudia a las 80 placas del lote,* ***no hace falta aplicar herramientas inferenciales puesto que ya se calcula de manera exacta el valor del parámetro, que en este caso es el número promedio de defectos por placa*** *(μ). Si se decide seleccionar una muestra de placas del lote, una vez registrados los valores de las cantidades de defectos en la superficie de las placas seleccionadas en la muestra, se realiza un análisis descriptivo de la información obtenida y luego, un análisis inferencial.*



Problema 3 (pág. 7): *Para este problema, una vez decidido cuántas semillas se van a sembrar, se lleva a cabo el experimento y se observan y registran para cada máquina (la que usa el dispositivo neumático y la que no) las distancias entre los pares de semillas consecutivas. Después de hacer un análisis descriptivo se obtienen los estadísticos correspondientes y se aplican herramientas inferenciales para dar respuesta al objetivo en estudio, con algún riesgo de error.*



Con lo visto hasta ahora, se puede resolver la Actividad 3-d, Sección 1.7.

1.5.5 Etapa de Obtención de conclusiones (C)

1. Conceptos estadísticos asociados

En esta etapa se da finalmente respuesta al problema de naturaleza estadística planteado, a partir de los resultados obtenidos en la etapa anterior.

Conclusión estadística es la interpretación de los resultados que arrojan las herramientas estadísticas aplicadas.

Conclusión en contexto es la vinculación de la conclusión estadística con el problema ingenieril en el que se originó el problema de naturaleza estadística.

La **validez externa** de un estudio se refiere a la posibilidad de generalizar las conclusiones del mismo a la población definida en el planteo del problema y también a otras poblaciones que se consideren similares.

Considerar la validez externa de las conclusiones lleva implícitas dos cuestiones importantes:

- La necesidad de tener claramente definida a la población de referencia a la hora de elaborar las conclusiones.
- La posibilidad de considerar las características de esta población y pensar en otras poblaciones con características similares. En este caso, se debe pensar: "lo que se encontró en este estudio, ¿a qué otras poblaciones se puede generalizar?"



Problema 1 (pág. 7): *Si solo se estudiaron barras de acero producidas en la planta de Venado Tuerto, las conclusiones que se obtengan se pueden generalizar a la producción de dicha planta y no a la producción de toda la empresa, a menos que se conozca por experiencia que en el resto de las plantas se trabaja de manera similar (máquinas similares, mismos proveedores de materia prima, etc.) y se pueda pensar que se producen barras de características similares.*

2. Tareas que se realizan en la etapa de Obtención de conclusiones

Una vez analizados los datos con las herramientas correspondientes, deben obtenerse las conclusiones estadísticas y luego considerarlas junto con información de contexto para dar una respuesta al problema.

Como se mencionara para la etapa de Análisis de los datos, si se trata de un estudio poblacional, alcanza con el análisis descriptivo y el cálculo de los parámetros de interés para obtener las conclusiones estadísticas. Si se trata de un estudio observacional por muestreo o de un estudio experimental, además del análisis descriptivo y del cálculo de los estadísticos correspondientes, se deben aplicar herramientas inferenciales que permitan generalizar las conclusiones estadísticas a la población bajo estudio. Estas conclusiones pueden ser erróneas, ya que no se trabaja con la totalidad de las observaciones (errores propios del muestreo) pero los riesgos de que esto ocurra están controlados por quien lleva a cabo la investigación, que los define en la etapa de Planificación del estudio estadístico.

Existe otro tipo de errores (errores ajenos al muestreo) que se presentan tanto en los estudios poblacionales como en los muestrales, como por ejemplo, un dato mal tomado, una variable mal medida o mal registrada en la planilla, un cálculo mal hecho, etc. Para estos errores no se pueden cuantificar los riesgos, pero se debe trabajar cuidadosamente en todas las etapas del ciclo PPDAC para evitarlos.

En esta etapa es importante:

- Interpretar los resultados de las herramientas aplicadas en términos estadísticos, comprendiendo exactamente el alcance y las limitaciones de las mismas.
- Interpretar dichos resultados en el contexto del problema, dando respuesta al objetivo planteado.
- Considerar la validez externa de las conclusiones.



Con lo visto hasta ahora, se pueden resolver las Actividades 3-e, 7, 8, 9 y 10, Sección 1.7.

En la Figura 1.8 se destacan los principales conceptos estadísticos asociados a cada una de las etapas del ciclo de resolución de problemas. En la Tabla 1.3 se resumen las etapas seguidas para los Problemas 1 y 2.

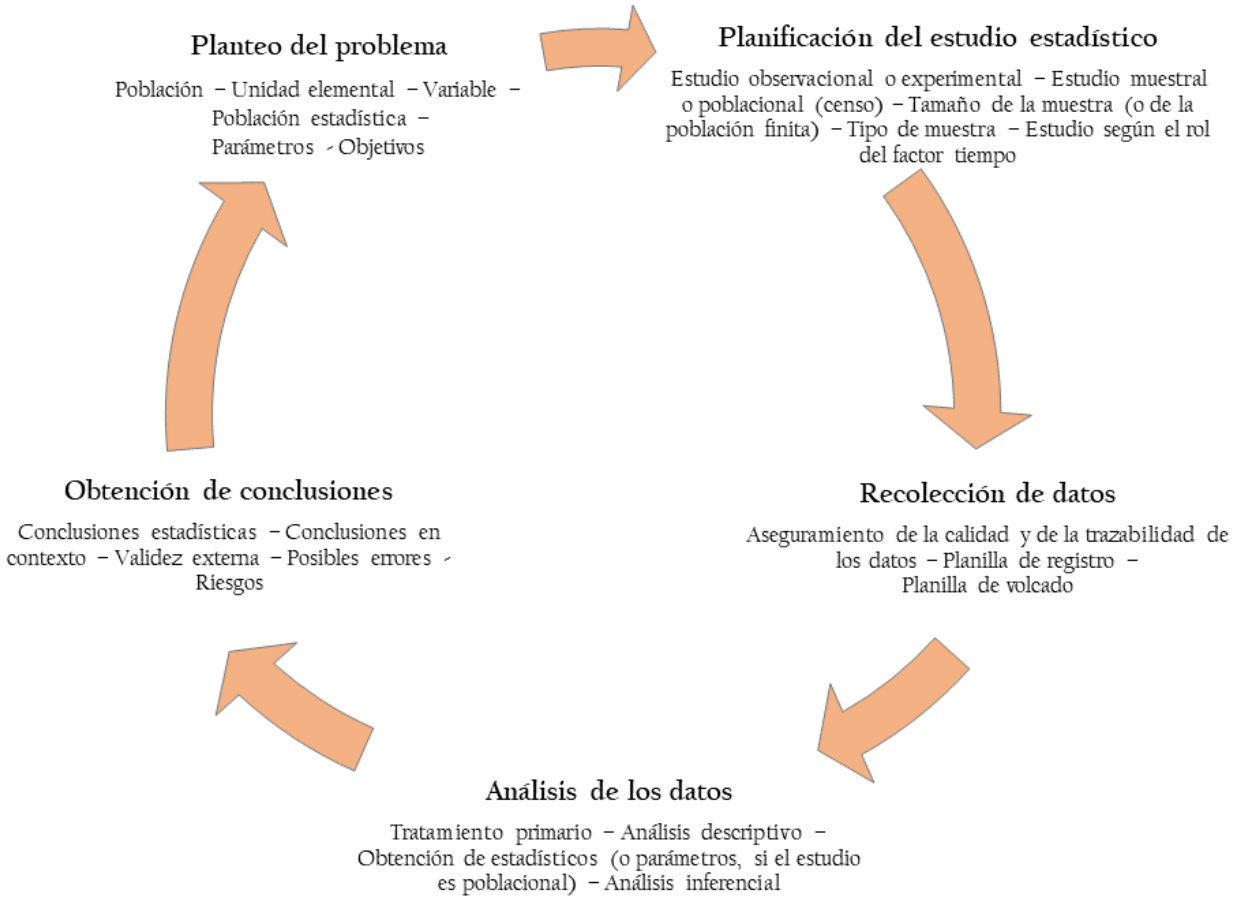


Figura 1.8 *Etapas del ciclo PPDAC para estudios poblacionales y muestrales (o experimentales). Conceptos estadísticos asociados a cada una*

Tabla 1.3 *Etapas del ciclo PPDAC para los Problemas 1 y 2*

Etapa	Problema 1 - Estudio observacional muestral (no interesa el factor tiempo)	Problema 2 - Estudio observacional poblacional [8] (no interesa el factor tiempo)
P: Planteo del problema	***Las barras de acero de cierto tipo, ¿cumplen con los requerimientos de la automotriz en relación a su longitud? ¿Cuánto vale la proporción de barras que satisfacen las especificaciones en relación a la longitud?***	***¿Se puede enviar el lote disponible al cliente? ¿La cantidad promedio de defectos por placa, en el lote disponible, es menor a 1,2?***
	Población: todas las barras de acero de cierto tipo que se producen en la empresa (pob. infinita). **Unidad elemental**: cada barra **Variables**: longitud (en mm) y cumplimiento de las especificaciones (con categorías "cumple", "no cumple") . **Población estadística**: las longitudes y el cumplimiento de especificaciones de todas las barras de acero de cierto tipo que se producen en la empresa. **Parámetro de interés**: proporción de barras que cumplen las especificaciones (π). **Objetivo en términos estadísticos**: Conocer el valor de π.	**Población**: la totalidad de placas de madera del lote (población finita $N = 80$). **Unidad elemental**: cada placa de madera. **Variable**: cantidad de defectos. **Población estadística**: las cantidades de defectos de las 80 placas de madera del lote. **Parámetro de interés**: cantidad promedio de defectos por placa (μ). **Objetivo en términos estadísticos**: Conocer si $\mu < 1{,}2$ defectos por placa.
P: Planificación del estudio estadístico (Algunas tareas)	Se decide realizar un estudio observacional de tipo muestral. Se establece el tamaño de la muestra y la forma de selección de las unidades. Se define el método de medición[9] de las variables. Se planifican los análisis que se llevarán a cabo, el software a emplear y los niveles de riesgo tolerados asociados a las herramientas inferenciales	Se decide realizar un estudio observacional de tipo poblacional. Dado que, en este caso, se va a utilizar información previamente recolectada, se planifican los análisis que se llevarán a cabo y el software a emplear.
D: Recolección de datos	Se seleccionan las n barras que integrarán la muestra y se registran sus longitudes. Se cargan los datos en la planilla de volcado. Se controlan errores en la toma y en la carga de los datos.	Se registran las 80 cantidades de defectos. Se cargan los datos en la planilla de volcado. Se controlan errores en la carga de los datos.
A: Análisis de los datos	Se realiza el análisis descriptivo de la muestra obtenida. Se calcula algún estadístico que dé buena información acerca del parámetro de interés (π). Se realiza el análisis inferencial para obtener posibles valores del parámetro con riesgo de error controlado.	Se realiza el análisis descriptivo de la población estadística obtenida. Se calcula el valor del parámetro de interés (μ).
C: Obtención de conclusiones	Se elaboran las conclusiones estadísticas sobre el parámetro π, obteniendo un conjunto de valores posibles. En función de estos resultados y de la información de contexto disponible se decide si las barras pueden venderse de inmediato a la automotriz o si es necesario realizar previamente modificaciones en el proceso, etc.	Se elaboran las conclusiones estadísticas sobre el parámetro μ, determinando si su valor es inferior a 1,2. En función de este resultado y de la información de contexto disponible se decide si el lote puede enviarse directamente al cliente o si es necesario realizar previamente ajustes antes de su despacho, etc.

[8] Para este problema, se supone en este caso que se decidió estudiar a todas las placas del lote.

[9] Es importante chequear el procedimiento de medición antes de utilizarlo para medir la longitud de cada barra

Herramientas básicas para la mejora de los procesos

Las **"7 herramientas básicas de la calidad"** o **"7 herramientas básicas de Ishikawa"**[10] constituyen un conjunto de técnicas de gestión (**diagrama de espina de pescado**) y estadísticas (**planilla de volcado, histograma, diagrama de Pareto, diagrama de dispersión, estratificación y gráficas de control**) identificadas por su autor como las más útiles en la solución de la mayoría de los problemas relacionados con la calidad de los procesos.

Estas herramientas poseen la ventaja de ser conceptualmente simples, visualmente intuitivas y de fácil aplicación en contextos industriales y de ingeniería, permitiendo organizar información, identificar causas, detectar patrones y fundamentar la toma de decisiones basada en datos. En conjunto, constituyen un soporte operativo para las distintas etapas del ciclo PPDAC, especialmente en la recolección, organización y análisis de la información necesaria para comprender y resolver problemas reales.

En este capítulo se presentaron el **diagrama de espina de pescado** (Página 4) y la **planilla de volcado** (Página 24), herramientas orientadas principalmente a la identificación y estructuración de posibles causas y al registro sistemático de datos.

En el próximo capítulo se desarrollan el **diagrama de Pareto** y el **histograma**, dos herramientas fundamentales para el análisis descriptivo y la priorización de problemas, ampliamente utilizadas en el control y la mejora de procesos industriales.

La aplicación sistemática de estas y otras herramientas estadísticas permite transformar datos en información útil para el análisis, control y mejora continua de los procesos.

[10] Ver nota al pie en la Página 4.

1.6 Síntesis

Un problema de naturaleza estadística se caracteriza por la presencia de variabilidad e incertidumbre. En su resolución deben seguirse distintas etapas (Planteo del problema, Planificación del estudio estadístico, Recolección de datos, Análisis de los datos, Obtención de conclusiones), cada una de las cuales tiene su complejidad e importancia.



Las tareas que se llevan a cabo para dar respuesta al problema no necesariamente se desarrollan de manera secuencial o lineal: muchas veces, a partir del trabajo realizado en una de las etapas surge la necesidad de revisar y corregir tareas llevadas a cabo en etapas anteriores. Es clave, entonces, tener la mente abierta, usar la imaginación y la lógica y sobre todo, ser críticos en relación a lo realizado en cada paso, poniendo especial atención a los procedimientos y conceptos estadísticos involucrados.

Se destacan algunos puntos:

- El planteo del problema debe estar bien delimitado y los objetivos deben ser claros.

- La planificación del estudio estadístico debe ser completa para facilitar la recolección de los datos y el análisis de los mismos. Preferentemente, se deben considerar en esta etapa los riesgos máximos de error que se admiten en las conclusiones.
- La recolección de datos debe ser supervisada para garantizar su calidad y minimizar la presencia de errores no deseados.
- Los procedimientos de medición que se utilicen deben ser previamente evaluados.
- El análisis de los datos debe hacerse con herramientas adecuadas que garanticen la obtención de conclusiones válidas y acordes a los objetivos propuestos.
- Las conclusiones estadísticas deben luego asociarse a otras cuestiones de contexto para dar respuesta al problema ingenieril.

1.7 Actividades propuestas

1. Piense en un problema ingenieril y en un problema de naturaleza estadística asociado al mismo.
   a) Defina brevemente el problema ingenieril y plantee el problema de naturaleza estadística. ¿Qué preguntas desearía responder?
   b) Comente qué información cuantitativa requeriría. Defina población y variables de interés.
   c) Para alguna de las variables definidas, construya un diagrama de espina de pescado mencionando algunos factores que pueden causar variabilidad en la misma.

2. Seleccione uno de los problemas de naturaleza estadística presentados en la Sección 1.4 (página 7) y piense en un problema ingenieril que lo haya originado. Descríbalo brevemente.

3. Lea atentamente los enunciados de los 9 problemas de la Sección 1.4 y complete una tabla semejante a la Tabla 1.3, relativa a los Problemas 1 y 2, para los problemas restantes.
   Tenga en cuenta los siguientes puntos:
   a) Respecto del planteo del problema, defina:
      - Población (indique además si es infinita o finita y en ese último caso, mencione su tamaño) y unidad elemental.
      - Variable/s y su clasificación. Escala de medición.
      - Población estadística.
      - Parámetro/s de interés.
      - Objetivo del estudio estadístico.
   b) Respecto a la planificación del estudio estadístico que se llevará a cabo para dar respuesta al problema, responda a las siguientes preguntas. Fundamente su respuesta.
      - ¿Para cuál o cuáles de los problemas, se puede realizar un estudio exhaustivo (censo) y para cuál o cuáles se puede realizar un estudio por muestreo?
      - ¿Cuál o cuáles pueden ser estudios observacionales? ¿y experimentales?
      - ¿En cuál o cuáles de los estudios interesa conocer el comportamiento de la variable a través del tiempo?, ¿en cuáles no?
   c) Respecto a la recolección de datos:
      - ¿Cómo llevaría adelante la obtención de los datos requeridos en cada uno de los problemas?
      - Diseñe una planilla de recolección acorde.
   d) Sobre el análisis de los datos, indique, en cada caso, si corresponde realizar un análisis descriptivo y/o inferencial.
   e) En relación a la obtención de conclusiones, comente sobre la validez externa de las mismas, según cómo haya diseñado la recolección de datos.

4. Suponga que para el Problema 4 (Sección 1.4), se analizaron los motivos principales de las 350 devoluciones de productos comprados en línea, ocurridas durante el último semestre del año anterior y se encontró que aproximadamente el 33 % de las devoluciones analizadas se debía al retraso en la entrega.
   a) Defina un objetivo para el cual el conjunto de datos analizados corresponda a una muestra y otro para el cual corresponda a una población.
   b) ¿En cuál de los casos el valor obtenido correspondería a un parámetro?

5. ¿Qué alternativas tiene a la hora de planificar el estudio estadístico asociado al Problema 2, Sección 1.4? Explique y comente cuestiones relativas al análisis de los datos y a las conclusiones en cada una.

6. Para los siguientes enunciados, relacionados con algunos de los problemas de la Sección 1.4, indique si los valores mencionados corresponden a estadísticos o a parámetros. Justifique según el tipo de estudio que Ud. considere que se llevó a cabo.
   a) La resistencia promedio a la compresión de algunos puntos analizados en el tramo de ruta es 30,96 MPa (Problema 6).
   b) La cantidad promedio de defectos de todas las placas del lote disponible es 1,65 (Problema 2).
   c) La proporción de barras que no cumplen con las especificaciones, entre las analizadas, es 0,05 (Problema 1).
   d) Se consideraron 45 dispositivos y, como mínimo, transcurrieron 4 horas hasta que ocurrió la primera falla en uno de ellos (Problema 8).

7. Para los siguientes enunciados, relacionados con algunos de los problemas de la Sección 1.4, comente sobre la validez externa de las conclusiones. ¿A qué población se pueden generalizar?
   a) Suponga que en el Problema 1 interesa conocer la proporción de barras que cumplen con las especificaciones en relación a la longitud y para ello se seleccionó una muestra de 120 barras producidas en la planta Rosario durante el turno mañana.
   b) Suponga que los 45 dispositivos analizados en el marco del Problema 8, se seleccionaron al azar del total de dispositivos de cierto tipo.

8. En una empresa alimenticia interesa evaluar, en un momento dado, la calidad del proceso de producción de la nueva línea de galletitas horneadas en un momento dado. Formule el problema en términos estadísticos, defina posibles objetivos, variable/s y parámetro/s de interés.

9. Dé un ejemplo de un proceso, y para dicho proceso:
   a) Identifique una característica de interés, asociada a una salida (producto, servicio u otra).
   b) Enuncie los factores que puedan estar influenciando en la variabilidad de dicha característica. Construya un diagrama de espina de pescado.
   c) Plantee un problema que se pueda presentar en el proceso y, en relación al mismo, defina un objetivo de estudio estadístico que se llevaría a cabo. Luego:
      1) Defina población, unidad elemental, variable y parámetros de interés.
      2) ¿De qué tipo es la población? Mencione posibles criterios de inclusión/exclusión.
      3) Clasifique a la variable de interés. Mencione la escala de medición.
      4) Reescriba el objetivo en términos estadísticos.

10. Reflexione y responda:
    a) ¿Qué características tienen los problemas de naturaleza estadística que los distinguen de otro tipo de problemas?
    b) ¿Qué información brinda el conocimiento del "modelo del comportamiento" de una variable?
    c) ¿Qué importancia tiene definir con claridad el objetivo del estudio y la población asociada?
    d) ¿Qué condiciones debe cumplir la población para poder hacer un estudio poblacional? Comente.

e) ¿Por qué cree que son tan frecuentes los estudios muestrales?
f) ¿Qué diferencia un estudio observacional de uno experimental?
g) ¿En un estudio experimental, se puede relevar a toda la población?
h) ¿Qué condiciones impondría a la extracción de una muestra si quisiera luego realizar una inferencia estadística?
i) ¿Qué importancia tiene garantizar la calidad de los datos? ¿Cómo lo haría?
j) ¿Cómo reconoce si un grupo de datos corresponde a una población o a una muestra?
k) ¿Por qué se llama "conclusiones preliminares" a las obtenidas del análisis de una muestra?
l) ¿Por qué no se puede tomar una decisión solo con el cálculo de un estadístico?
m) ¿Qué importancia tiene la validez externa en la elaboración de las conclusiones?

1.8 Estadística con R

1.8.1 Introducción

R es un entorno y lenguaje de programación con un enfoque al análisis estadístico [2]. Sus comienzos se remontan a los años 1993-1996 y fue creado por **R**oss Ihaka and **R**obert Gentleman (Universidad de Auckland, Nueva Zelanda), basándose en el lenguaje S creado por John Chambers en 1976 para Bell Labs (parte de AT&T, desarrolladores de C).

R se distribuye gratuitamente bajo los terminos de la GNU General Public Licence [3] y se encuentra disponible en cualquier plataforma de sistema (Windows, OS X, Linux). El mantenimiento y desarrollo de R es realizado por el *R Development Core Team* [2], un equipo de especialistas en ciencias computacionales y estadística provenientes de diferentes instituciones y lugares alrededor del mundo. La versión de R mantenida por este equipo es conocida como *base* y como su nombre indica, es sobre aquella que se crean otras implementaciones de R, así como los *paquetes* que expanden su funcionalidad. Esto se realiza en forma colaborativa, por la cual quien lo desee puede publicar *paquetes*, es decir, conjuntos de funciones programadas con la misión de realizar una tarea específica. Algunas características de R son:

- Manejo y almacenamiento efectivo de los datos.
- Un conjunto de operadores para la realización de cálculos con matrices.
- Una gran colección de herramientas para el análisis de datos.
- Utilidades gráficas para la visualización de datos.
- Un lenguaje de programación bien desarrollado que incluye saltos condicionales, bucles, funciones recursivas, utilidades para la entrada y salida de datos, etc.



El lenguaje de programación R se integra bien con otros lenguajes de programación como C, C++ o Fortran para tareas de análisis de datos computacionalmente intensivas (alto consumo de recursos como CPU y RAM). Además, puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python.



Habitualmente, R se utiliza mediante RStudio [4], un entorno de desarrollo integrado. RStudio es un programa para manejar R y utilizarlo de manera más cómoda en algunos aspectos. Se trata de una máscara para visualizar el software que tiene como principales ventajas el orden y la visualización de los procesos que son llevados a cabo con R, todo de manera simultánea.

1.8.2 Instalación

Para instalar R hay que dirigirse a la colección de paquetes en CRAN (del inglés, *Comprehensive R Archive Network*) [5]. CRAN es un repositorio global donde se puede acceder al software y a los paquetes que la comunidad produce. La instalación comienza entrando en `https://cran.r-project.org/` donde se debe seleccionar el link correcto de descarga, dependiendo del sistema operativo.

A continuación, se podrá instalar RStudio. Para ello, se debe ingresar a `https://posit.co/products/open-source/rstudio/` y seguir los pasos para instalar Rstudio Desktop, de acuerdo al sistema operativo de la computadora donde se utilizará.

Tanto para R como para RStudio, se descargarán archivos instaladores que se deberán ejecutar, siguiendo los pasos, para su instalación completa.

1.8.3 Entorno

Cuando se ingresa a RStudio por primera vez, se observa una pantalla dividida en tres paneles, como se muestra en la Figura 1.9.

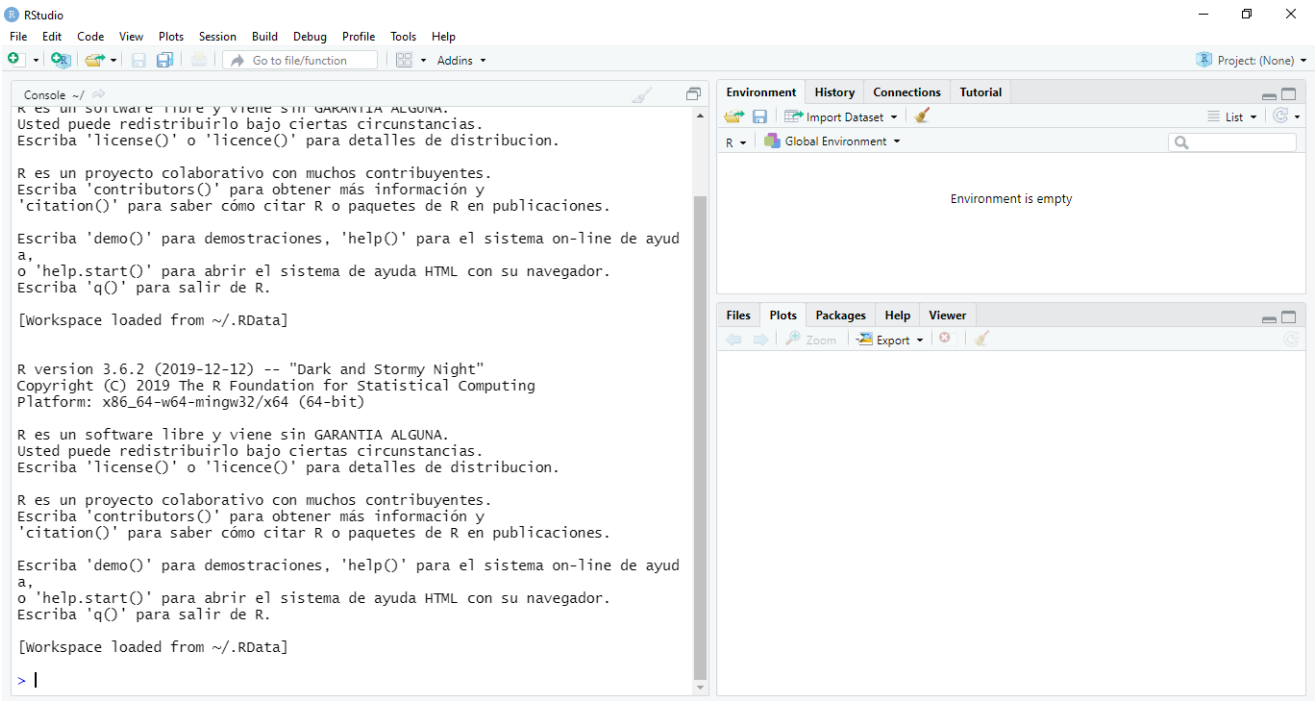


Figura 1.9 *Pantalla de RStudio con tres paneles iniciales*

- **Panel izquierdo**: En este panel se encuentra la *Consola*, que es la ventana que se comunica con el motor de R. A través de ella se envían las instrucciones o sentencias a R y se visualizan los resultados. Es el *cerebro* de R.
- **Panel superior derecho**: Panel con las pestañas *Environment*, *History*, *Connections* y *Tutorial*.
  - **Environment** (ambiente): Muestra todos los objetos que R tiene en la memoria, como conjuntos de datos, modelos, resultados, etc.
  - **History**(historial): Lista todas las instrucciones que R ha corrido.
  - **Connections**(conexiones: Muestra todas las conexiones activas a bases de datos remotas, establecidas por ejemplo mediante servidores SQL (esta pestaña puede no aparecer, dependiendo de la versión del programa que esté instalada).
  - **Tutorial**: Brinda tutoriales sobre RStudio pero para ello, previamente se debe instalar el paquete `"learnr"`.
- **Panel inferior derecho**: Panel con las pestañas *Files*, *Plots*, *Packages*, *Help* y *Viewer*. Aquí se pueden observar los archivos de trabajo, los gráficos generados, el listado de paquetes disponibles, ayuda sobre distintas funciones, entre otras funcionalidades.

Si bien en la Consola se pueden escribir y ejecutar instrucciones o sentencias, una alternativa es escribir todas las instrucciones en un lugar separado y luego *correrlas* en la consola cuando se pretende que R las ejecute.

Al guardar dichas instrucciones, se crea un *script* (un archivo de código o sentencias), que almacena todas las instrucciones utilizadas para realizar el análisis de datos en un archivo permanente, reproducible, comentado, compartible y multi-plataforma. RStudio provee para este fin un editor de texto incorporado. Para iniciar un nuevo script, se debe utilizar el botón indicado con un ícono + en la parte izquierda superior. En la Figura 1.10 se muestra la pantalla con el panel de *script* incluido, el cual se ubica en el panel superior izquierdo; mientras que la *Consola* queda en el panel inferior izquierdo:

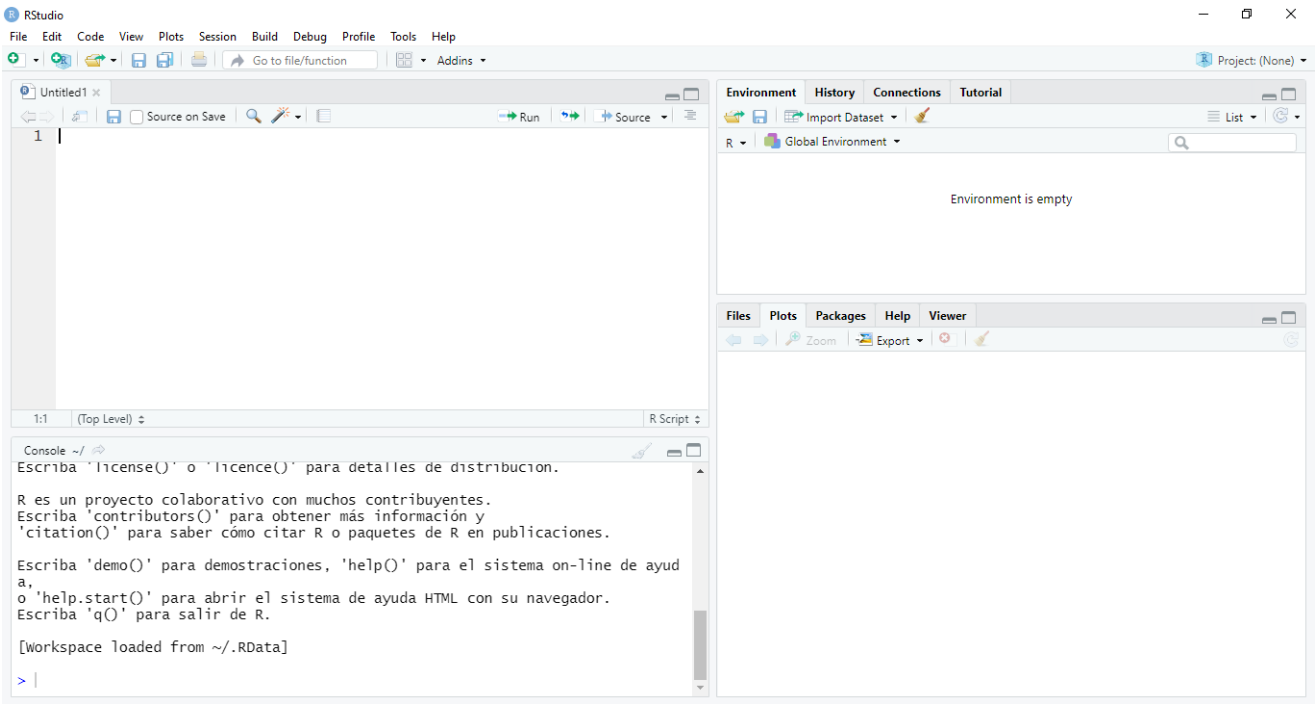


Figura 1.10 *Pantalla de RStudio con panel para script*

Es importante recordar guardar el script generado si se quiere usar posteriormente. Este se guarda como cualquier archivo, en una carpeta a elección (`File -> Save`, o directamente `Ctrl+S`) con un nombre que resulte informativo. Los archivos así creados reciben la extensión .R.

Para correr el script, se seleccionan todas las líneas que se pretende que R analice y ejecute. Luego, se hace click en el botón `Run`, o directamente `Ctrl+Enter`. Sin tener líneas del Script seleccionadas, se puede correr solamente la línea en la que actualmente está posando el cursor, haciendo también `Ctrl+Enter`.

1.8.4 Trabajo con paquetes

1. Instalación de paquetes

Al instalar R se incluyen múltiples funciones básicas que pueden servir para la importación de datos, la realización de transformaciones, el ajuste y la evaluación de modelos estadísticos, las representaciones gráficas, etc. Sin embargo, la enorme potencia de R deriva de su capacidad de incorporar en cualquier momento nuevas funciones capaces de realizar nuevas tareas específicas.

Un paquete (*package*) es una colección de funciones, datos y código R que se almacenan en una carpeta conforme a una estructura bien definida, fácilmente accesible para R.

Al instalar R se incorporan por defecto numerosos paquetes. Se puede ver una lista de los paquetes que actualmente están instalados en un ordenador ejecutando `library()` o bien ingresando en la pestaña *Packages* en el panel inferior derecho.

2. Uso de paquetes

Es importante distinguir entre tener un paquete instalado en el ordenador y tenerlo cargado en memoria. Tenerlo instalado en el ordenador significa simplemente que en algún momento se ha bajado de internet y se ha copiado en algún directorio en el que R lo puede localizar. Tenerlo cargado en memoria significa que, durante la actual sesión de trabajo, R ha leído el contenido del paquete y ha incorporado las funciones que contiene a su espacio de trabajo, de tal forma que tales funciones pueden ya ser invocadas y ejecutadas.

Si se observa el listado de paquetes en la pestaña *Packages* se ve que en el mismo hay algunos que están tildados y otros que no. Solo los paquetes tildados están actualmente cargados en memoria. El resto está simplemente guardado en algún directorio, sin que sus funciones estén aún disponibles para ser ejecutadas desde R. Se puede ver un listado completo de los paquetes actualmente cargados en memoria mediante la ejecución de `search()`.

Para poder utilizar un paquete específico en R, el mismo debe estar instalado, ya sea porque se incluye junto con la instalación del software o porque se realiza manualmente. Hay dos formas principales de hacer esto:

- Usar el comando `install.packages()` e introducir el nombre del paquete a instalar entre comillas, por ejemplo `install.packages("ggplot2")`.
- Dentro de la pestaña *Packages*, hacer click en el botón `Install`. Se abre una ventana asistente, que permite indicar el nombre del paquete que se desea instalar, habitualmente desde el repositorio CRAN. Finalmente, se pulsa `Install` y comienza la descarga.

En ambos casos, R se conecta a alguno de los repositorios de CRAN en internet, descarga el archivo que contiene el paquete, lo descomprime y lo instala en el directorio de paquetes por defecto.

Para activar o utilizar un paquete, se debe ejecutar el comando `library()` e introducir el nombre del paquete a utilizar entre comillas, por ejemplo `library("ggplot2")`. También se puede ir a la pestaña *Packages* de RStudio, buscar el paquete a emplear y hacer click en el cuadro que se encuentra a la izquierda del nombre. Es recomendable utilizar la primera opción y dejar las sentencias `library()` escritas en el script si se va a realizar el mismo análisis posteriormente.

Todos los paquetes incluyen obligatoriamente manuales de ayuda con las instrucciones para utilizar sus funciones (se ven en la pestaña *Help*). Algunos, además, vienen con viñetas, que son tutoriales donde los autores muestran las funcionalidades de sus paquetes de forma amigable y detallada, con ejemplos. Es posible explorar las viñetas de un paquete con la función `browseVignettes()` e introducir el nombre del paquete de interés entre comillas, por ejemplo, `browseVignettes("ggplot2")`.

1.8.5 Importación de datos

En principio, es importante destacar que, para el trabajo con R, es esencial seguir los principios de *datos ordenados* [6], un conjunto de reglas claras para organizar datos de manera consistente, de modo que permitan su fácil lectura y análisis.

El principio fundamental de los datos ordenados es tener una única unidad por fila y una única variable por columna (conocido también como *formato largo*), de modo que cada celda (cruce entre fila y columna) contenga un solo valor, tal como se ejemplifica en la Figura 1.11.

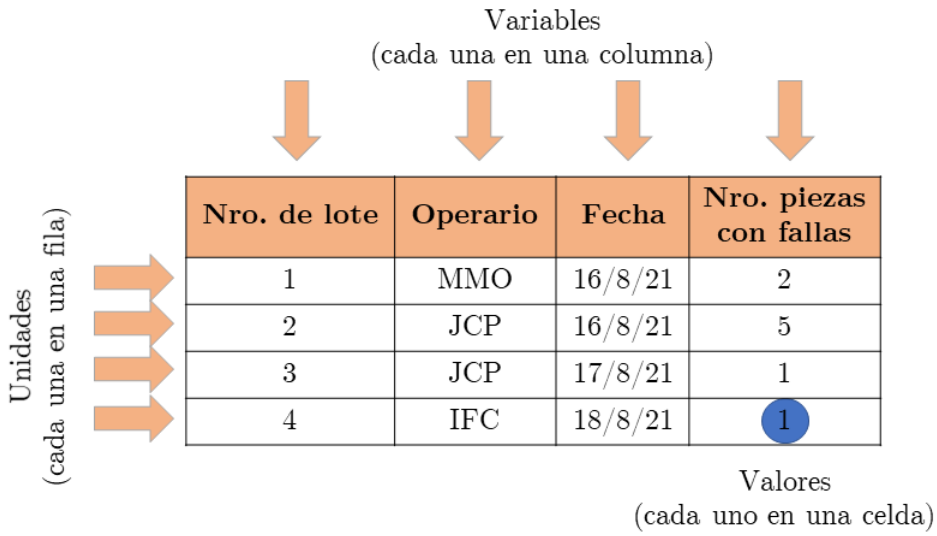


Figura 1.11 *Ejemplo de conjunto de datos ordenados*

Si bien existen formas de generar conjuntos de datos utilizando RStudio, generalmente los mismos se encuentran en formatos de archivos asociados a otros software, como puede ser Microsoft Excel u otros softwares estadísticos como SPSS, SAS o Stata, entre otros. R permite la importación de diversos formatos de archivos.

1. Importación de datos desde Excel

El paquete `readxl` [7] forma parte del paquete `tidyverse` [8], creado por Hadley Wickham (científico jefe en RStudio) y su equipo. Permite leer con R archivos que se encuentran en formato `.xls` o `.xlsx`. Como primer paso, es necesario instalarlo. Para ello, basta con ejecutar la sentencia `install.package("readxl")`.

Una vez que el paquete se encuentra instalado, solo será necesario ejecutar `library("readxl")` para hacerlo disponible para importar datos.

Por otra parte, es necesario conocer la *ruta* del archivo de Excel. La ruta, o *path* en inglés, es la dirección o ubicación donde está el archivo en la computadora. Si se ejecuta la sentencia `file.choose()` se abre una ventana que permite buscar el archivo de interés. Al pulsar `Open`, se obtiene la ruta correspondiente en la consola de trabajo. Luego:

- Si el archivo de Excel contiene una sola *hoja*, se ejecuta `datos <- read_excel()` con la ruta del archivo entre comillas, por ejemplo:

`datos <- read_excel("C:/Usuario/Juan/Tarea/Datos.xlsx")`
La primera palabra indica el nombre que tendrá el conjunto de datos importados dentro de R. La flecha invertida indica que, bajo ese nombre, se guardará el archivo que se indica a continuación.

- Si el archivo de Excel contiene varias *hojas*, es necesario indicar cuál de ellas utilizar, agregando la opción `sheet=` con el nombre de la hoja entre comillas, por ejemplo:
`datos <- read_excel("C:/Usuario/Juan/Tarea/Datos.xlsx", sheet="Hoja1")`
- En los casos anteriores, se supone que los datos se encuentran en la primera celda (A1) del archivo. En caso de no ser así, es posible indicar el rango de celdas a importar. Si los datos se encuentran, por ejemplo, en las columnas C, D y E y en las filas 7 a 13 de la Hoja 2 del archivo, entonces:
```
datos <- read_excel("C:/Usuario/Juan/Tarea/Datos.xlsx",
sheet="Hoja2", range="C7:E13")
```

Una vez importado, el conjunto de datos (*dataset*) figura en la pestaña *Environment* junto con todas las otras cosas que están en la memoria de R en ese momento. Si se hace click sobre el dataset en ese panel, se abrirá una ventana mostrándolo. Lo mismo se consigue con la función `View("datos")`.

Un siguiente paso razonable es asegurarse que los datos que R acaba de leer realmente son los deseados. Algunas cosas que se pueden chequear:

- Cantidad de filas.
- Cantidad y nombres correctos para las variables.
- El tipo de variables es correcto (por ejemplo, R reconoció variables numéricas como tales y no como caracteres).
- Variables cualitativas tienen el número correcto de categorías o niveles.

Los datos cargados se pueden explorar con algunas funciones básicas de R:

- `names("datos")` da los nombres asignados a cada columna (variable).
- `head("datos")` devuelve las primeras seis líneas del dataset.
- `tail("datos")` devuelve las últimas seis líneas del dataset.
- `dim("datos")` devuelve el número de filas y columnas.
- `str("datos")` describe toda la estructura del dataset, combinando toda la información anterior.
- `summary("datos")` es una función que se puede aplicar sobre distintos tipos de objetos. Cuando se la usa sobre un dataset, muestra un pequeño resumen de cada variable.

Cabe destacar que las funciones anteriormente mencionadas utilizan como ejemplo un dataset denominado *datos*, pero dicho nombre, que aparece entre comillas, debe adaptarse al del conjunto de datos que se quiere explorar.

2. Importación de datos en texto plano

Un archivo de *texto plano*, es un archivo informático que contiene únicamente texto formado por caracteres que son legibles por humanos, careciendo de cualquier tipo de formato tipográfico (negritas, subrayado, tipo de letra). Esta simplicidad permite que una gran variedad de programas pueda leer y editar ese contenido. Las distintas columnas (variables) deben indicarse con algún tipo de separador, por ejemplo, una coma o un espacio. Lo más frecuente es encontrar archivos en formato delimitado

por tabulaciones (`.txt`) o por coma (`.csv`, *comma separated values*).

Para leer archivos de este tipo, es recomendable instalar y utilizar el paquete `readr` [9]. Una vez cargado, provee distintas opciones:

- `read_csv()`: Para leer archivos con coma como separador.
- `read_csv2()`: Para leer archivos con punto y coma como separador.
- `read_tsv()`: Para leer archivos con tabulador como separador.
- `read_delim(,sep = "-")`: Para leer archivos con separador distintos como puede ser un guión medio.

Luego de la opción adecuada, se indica entre paréntesis y con comillas la ruta del archivo a leer. Por ejemplo, si el archivo se encuentra en formato `.csv` separado por coma, se ejecuta:

```
datos <- read_csv("C:/Usuario/Juan/Tarea/Datos.csv")
```



2. Análisis descriptivo de un conjunto de datos

2.1 Introducción

Frente a un problema de naturaleza estadística, lo primero que se debe realizar es plantearlo adecuadamente y luego seguir adelante con las restantes etapas del Ciclo PPDAC hasta arribar a las conclusiones y considerarlas en el contexto de ese problema. Durante la etapa de Recolección de datos (Etapa D), se seleccionan las unidades a estudiar y se observan en ellas las variables de interés, para luego completar la planilla de volcado, la cual se presenta en el Capítulo 1. Una vez iniciada la etapa de Análisis de los datos (Etapa A), concluir sobre el comportamiento del conjunto de datos a partir de observar la planilla tal y como se presenta es una tarea compleja. Sin embargo, la aplicación de herramientas de análisis descriptivo permite presentar los datos de manera de revelar sus aspectos más salientes y obtener información de utilidad. Para dar sentido a cualquier conjunto de datos hay que ordenarlos, agruparlos, construir tablas, representarlos gráficamente y obtener indicadores (promedio, desviación estándar y proporciones, entre otros). Todas estas tareas corresponden al **análisis descriptivo de los datos**. Previamente es necesario efectuar un **tratamiento primario** de los mismos, que incluye evaluar su calidad, identificar posibles errores en la medición de variables, en la codificación o en la carga, así como considerar la existencia de datos faltantes. En este sentido, garantizar la trazabilidad de los datos resulta de gran utilidad.

Cuando los datos provienen de un censo o estudio poblacional, luego de aplicar herramientas del análisis descriptivo (Etapa A) se pasa directamente a la obtención de conclusiones (Etapa C). En el caso de contar con una muestra, los datos están incompletos, en el sentido que no se cuenta con información de toda la población y, por ese motivo, a los resultados obtenidos del análisis descriptivo se los debe tomar como preliminares. El análisis se completa con herramientas inferenciales para generalizar las conclusiones a la población de referencia (Etapa C). Estas herramientas, que también se asocian al análisis de los datos (Etapa A), se presentan en los Capítulos 8 y 9.

En este capítulo se presentan las principales herramientas de **análisis descriptivo univariado**[1] tanto para variables cualitativas o categóricas como para variables cuantitativas continuas o discretas. El conjunto de datos que se analiza puede corresponder a una muestra o a una población finita.

[1] Se denomina *univariado* porque estudia el comportamiento de una única variable.

Los objetivos de este capítulo son:

- Presentar diferentes tablas y gráficos.
- Definir las medidas de resumen (indicadores) asociadas a la localización y a la variabilidad del conjunto de datos. Enunciar los conceptos y las principales propiedades de cada una, y presentar la forma de obtenerlas e interpretarlas.

2.2 Organización de los datos

Observando la planilla de volcado, es decir, los datos en bruto, es difícil dar respuesta a las preguntas planteadas. Como se manifestó en la sección anterior, para obtener información a partir de los datos, hay que ordenarlos y organizarlos en tablas y/o gráficos. La forma más sencilla de tabularlos es considerar los diferentes valores o conjuntos de valores (clases) que resultaron de medir la variable en las unidades estudiadas y contar cuántas veces se presenta cada uno de ellos (frecuencia absoluta), para construir una tabla que se denomina **tabla de distribución de frecuencias absolutas** (Tabla 2.3).

Tabla 2.1 *Tabla de distribución de frecuencias absolutas*

Variable (clase)	Conteo	Cantidad de unidades (frecuencia absoluta)
c_1	///// ///// /////	n_1
c_2	///// ///// ///	n_2
c_3	///// //	n_3
...	...	...
c_k	/////	n_k
Total		$\boldsymbol{n}$

Una **clase** es un conjunto de unidades que comparten características comunes. La j-ésima clase se representa mediante c_j.
El **número total de clases** consideradas se denota por k. Cuando cada clase coincide con un valor de la variable, k es el número de valores distintos que asume dicha variable.

En el caso de variables *cualitativas* o *categóricas*, una clase puede estar formada por una única categoría o nivel, o bien por un conjunto de categorías o niveles.



Problema 4 (pág. 8): *Para la variable Motivo principal de devolución, las clases son: c_1: "Error en la facturación", c_2: "Pedido equivocado", c_3: "Precio no coincidente", etc. Pero también podrían agruparse de manera diferente: c_1: "Error en la facturación o precio no coincidente", c_2: "Pedido equivocado", etc.*

En el caso de variables *cuantitativas discretas*, una clase puede estar formada por un único valor o por un conjunto de valores (esta última opción solo se utiliza si el número de valores distintos de la variable es grande).



Problema 5 (pág. 8): *Para la variable Cantidad de imperfecciones, las clases son: c_1: 0, c_2: 1, c_3: 2, c_4: 3,c_5: 4.*

En el caso de variables *cuantitativas continuas*, una clase está formada por un intervalo de valores de la variable. En este caso, definir las clases implica decidir cuántos intervalos se van a considerar y determinar su amplitud. Para que no se presenten dudas al clasificar a un elemento como perteneciente a uno u otro intervalo, existen dos maneras de construirlos:

1. Definiendo intervalos semicerrados en alguno de sus extremos (intervalos a límites nominales).
2. Definiendo intervalos abiertos pero considerando una cifra decimal adicional en los extremos de dichos intervalos (intervalos a límites reales).



Problema 1 (pág. 7): *Se mide la variable Longitud, en mm, con una cifra decimal significativa. Se pueden definir las clases de la siguiente manera:*

1. *Intervalos a límites nominales:* $c_1 : (249,2\ ;\ 249,4]$, $c_2 : (249,4\ ;\ 249,6]$, $c_3 : (249,6\ ;\ 249,8]$, *etc.*
2. *Intervalos a límites reales:* $c_1 : (249,25\ ;\ 249,45)$, $c_2 : (249,45\ ;\ 249,65)$, $c_3 : (249,65\ ;\ 249,85)$, *etc.*

Observaciones importantes:

1. Si se considera al conjunto C, definido por la totalidad de valores que resultaron de medir la variable en las unidades estudiadas, las clases constituyen subconjuntos del mismo y deben cumplir con las siguientes propiedades:
   a) $c_i \cap c_j = \emptyset$, para todo $i \neq j$
   b) $\cup_{i=1}^{k} c_i = C$
   La propiedad *a)* indica que no puede haber una unidad elemental que pertenezca a dos clases simultáneamente (*clases mutuamente excluyentes*). En el caso de variables cuantitativas agrupadas en intervalos, esto se puede lograr construyendo intervalos a límites nominales o a límites reales, como se presentó en párrafos anteriores. En el caso de variables cualitativas, la propiedad se cumple definiendo claramente el criterio de clasificación de los elementos de las unidades.
   La propiedad *b)* indica que no deben quedar unidades sin ser clasificadas (*clases exhaustivas*) es decir, cada unidad bajo estudio debe pertenecer a alguna de las clases.
2. Independientemente del tipo de variable que se trate, se dice que una unidad pertenece a una clase cualquiera c_j cuando presenta cualquier valor de la variable incluido en ella.

La **frecuencia absoluta** de la clase c_j se define como la cantidad de unidades pertenecientes a dicha clase en el total de unidades analizadas. Surge directamente del conteo y se representa mediante n_j.

A partir de las frecuencias absolutas, se pueden obtener otras frecuencias, que se presentan y definen en la Tabla 2.2 y así completar la **tabla de distribución de frecuencias**.

Tabla 2.2 *Frecuencias absolutas, relativas y acumuladas para la clase c_j*

	Frecuencia absoluta n_j	**Frecuencia relativa** f_j (***)	**Frecuencia absoluta acumulada** N_j(*)	**Frecuencia relativa acumulada** F_j(*) (***)
¿Cómo se define?	Es la cantidad de unidades elementales que se clasifican en c_j.	Es la proporción de unidades elementales clasificadas en c_j.	Es la cantidad de unidades elementales que se clasifican en c_j y en todas las clases que contienen valores menores.	Es la proporción de unidades elementales clasificadas en c_j y en todas las clases que contienen valores menores.
¿Cómo se obtiene?	Se obtiene directamente del conteo de las unidades en cada clase.	Se obtiene dividiendo la frecuencia absoluta por la cantidad total de datos (n o N)(**) $f_j = \frac{n_j}{n}$	Se obtiene sumando las frecuencias absolutas correspondientes a las clases c_i, con $i \leq j$. $N_j = \sum_{i=1}^{j} n_i$	Se obtiene dividiendo la frecuencia absoluta acumulada por el total de datos (n o N)(**) $F_j = \frac{N_j}{n}$.
¿Qué condiciones / propiedades debe cumplir?	La suma de las frecuencias absolutas coincide con el tamaño de la muestra o la población, según corresponda.	Ninguna frecuencia relativa es menor que 0. La suma de las frecuencias relativas vale 1.	Las frecuencias absolutas acumuladas constituyen una función no decreciente. El valor de N_k coincide con el tamaño de la muestra o de la población, según corresponda.	Las frecuencias relativas acumuladas constituyen una función no decreciente. El valor de F_k es 1.

(*) Las frecuencias acumuladas solo tienen sentido si el nivel de medición de la variable es ordinal o superior.

(**) Recuerde que n representa al total de observaciones obtenidas, cuando se analiza una muestra o se lleva a cabo un experimento. Si se trata de una población finita, dicho total se representa con N.

(***) Los valores de las frecuencias relativas (f_j y F_j) pueden multiplicarse por 100 % y expresarse en porcentaje.

2.2.1 Distribuciones de frecuencias para variables cualitativas o categóricas. Representación tabular y gráfica

Recuerde que cuando se cuenta con una variable *cualitativa* o *categórica*, las clases pueden estar formadas por una o más categorías, según interese. Entonces, el número de clases k coincide con el número de categorías consideradas o es menor. Para cada clase se obtienen las frecuencias absolutas y, a partir de ellas, las frecuencias relativas definidas en la Tabla 2.2. Las frecuencias acumuladas solo se obtienen si la variable está medida en escala ordinal.

La información relativa a las diferentes clases y las frecuencias correspondientes se puede presentar en forma tabular o gráfica: se puede construir una *tabla de distribución de frecuencias*, un *gráfico de barras* o uno *de sectores* para las frecuencias no acumuladas (n_j o f_j) y, en algunos casos, un *diagrama de Pareto*.



Problema 4 (pág. 8): *En este problema, la población está constituida por todas las devoluciones de productos vendidos en línea que recibe la empresa.* ***La variable de interés es el motivo principal de devolución, variable cualitativa medida en escala nominal.***
Entre los objetivos del estudio se encuentran identificar los motivos de devolución más frecuentes y determinar la proporción de devoluciones ocasionadas por retrasos en la entrega.
Para cumplir con estos objetivos, se analizaron las 350 devoluciones correspondientes al último semestre del año anterior, las cuales pueden considerarse una muestra del total de devoluciones de la empresa. Una parte de la planilla de registro se presenta en la Tabla 2.3.

Tabla 2.3 *Planilla de registro de los motivos principales de las devoluciones analizadas*

Fecha	***Responsable***	***Devolución N°***	***Motivo principal***	***Código***
02/07/25	*JLK*	*01*	*Producto dañado*	*(4)*
03/07/25	*JLK*	*02*	*Producto dañado*	*(4)*
...	*...*	*...*	*...*	*...*
27/12/25	*MLC*	*350*	*Pedido equivocado*	*(2)*

A continuación se presenta una parte de la planilla de volcado. En el archivo https://sites.google.com/view/libro-estadistica se encuentra la planilla completa.

Tabla 2.4 *Planilla de volcado de los motivos principales de las devoluciones analizadas*

Devolución N°	***Código del motivo principal***
1	*4*
2	*4*
...	*...*
350	*2*

En la Tabla 2.5 se presenta la distribución de frecuencias para el motivo principal de devolución. Se destaca el "retraso en la entrega por problemas de transporte", con un 32,86 % de las devoluciones analizadas. Le siguen, en orden decreciente, el "producto dañado" con casi un 22 % de las devoluciones, el "pedido equivocado" con un 16,57 % y los "errores en la facturación" con un 11,43 %. Los restantes motivos representan, cada uno, menos del 10 % de estas devoluciones.

Tabla 2.5 *Tabla de distribución de frecuencias para los motivos principales de las 350 devoluciones analizadas*

Motivo principal de devolución	***Cantidad de devoluciones*** (n_j)	***Proporción de devoluciones*** (f_j)
Error en la facturación (c_1)	*40* (n_1)	*0,1143* (f_1)
Pedido equivocado (c_2)	*58* (n_2)	*0,1657* (f_2)
Precio no coincidente (c_3)	*25* (n_3)	*0,0714* (f_3)
Producto dañado (c_4)	*76* (n_4)	*0,2171* (f_4)
Producto mal embalado (c_5)	*15* (n_5)	*0,0429* (f_5)
Producto vencido o en mal estado (c_6)	*21* (n_6)	*0,0600* (f_6)
Retraso en la entrega (c_7)	*115* (n_7)	*0,3286* (f_7)
Total	*350* (n)	*1*

> *Dado que la variable Motivo principal de devolución está medida en escala nominal, no corresponde obtener frecuencias acumuladas.*

A continuación se interpreta, a modo de ejemplo, la fila correspondiente a la tercera clase, en términos del problema:
c_3: Precio no coincidente, n_3: 25 devoluciones y f_3: 0,0714 (7,14 % de las devoluciones).
En 25 de las devoluciones analizadas, el motivo principal fue que no coincidía el precio final del producto con el que figuraba en la página. En términos relativos, en una proporción de 0,0714 de las devoluciones analizadas (el 7,14 % de las devoluciones, llevado a porcentaje) el motivo principal fue la no coincidencia del precio con el que figuraba en la página.

1. Gráfico de barras

La distribución de frecuencias absolutas o relativas correspondiente a una variable cualitativa se puede representar a través de un **gráfico de barras simples**. A cada clase se le asigna una barra cuya altura es proporcional a la frecuencia. Todas las barras tienen el mismo ancho y color y están separadas entre sí por el mismo espacio. La escala del eje donde se representan las frecuencias debe comenzar en 0 para evitar que las interpretaciones se distorsionen. En el caso que el número de clases sea pequeño, se puede reemplazar el gráfico de barras por el de sectores.

2. Gráfico de sectores

El **gráfico de sectores** presenta la composición de un total en partes. Para construirlo, se divide al círculo en tantos sectores como clases se quieran representar y se halla el ángulo central de cada sector de tal manera que este sea proporcional a la frecuencia de cada una de las clases. Este gráfico solo es útil si el número de clases k es pequeño.

3. Diagrama de Pareto

Este diagrama es un tipo especial de gráfico de barras verticales donde las respuestas categorizadas se presentan en orden de frecuencia descendente y se combinan con un polígono acumulativo. En su construcción clásica, el eje de ordenadas de la izquierda contiene las frecuencias absolutas o relativas y el eje de la derecha, contiene las frecuencias relativas acumuladas, generalmente expresadas como %. En el eje de abscisas están las distintas clases, en orden decreciente según sus frecuencias.



Problema 4 (pág. 8): *En la Figura 2.1 se presenta la distribución de frecuencias relativas correspondiente al Problema 4 a través de un gráfico de sectores y de uno de barras.*

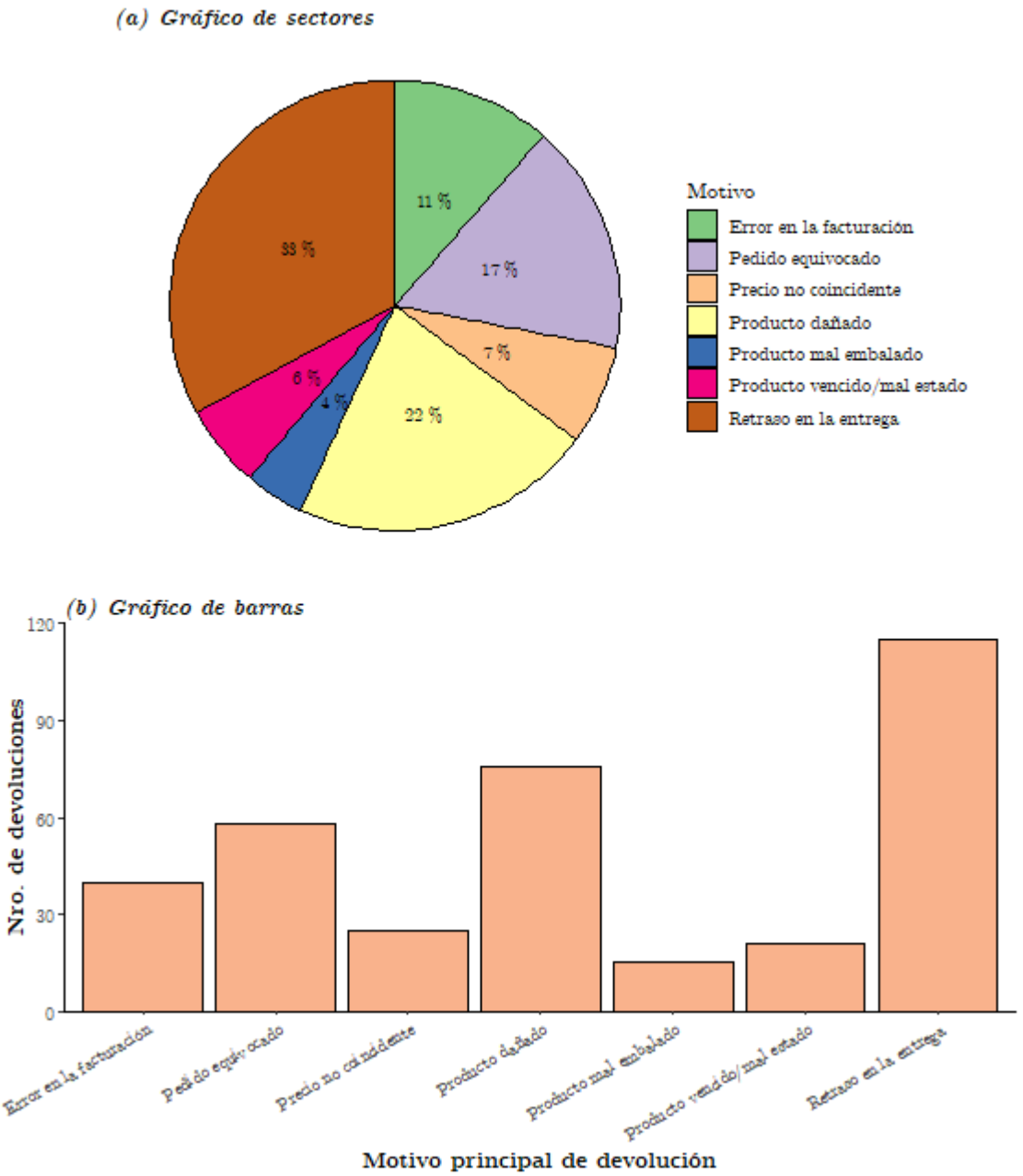


Figura 2.1 *Gráfico de sectores y gráfico de barras de los motivos principales de las devoluciones analizadas*

> *En este caso, dado que el número de categorías no es pequeño, resulta más simple comparar visualmente las alturas de las barras que las áreas que forman los sectores.*

En la Figura 2.2 se presenta un ***diagrama de Pareto*** *construido con la información sobre los motivos principales de las devoluciones en las compras por Internet.*
De los gráficos correspondientes (Figuras 2.1 y 2.2) se observa que el motivo de devolución más frecuente entre las devoluciones analizadas fue el "retraso en la entrega por problemas de transporte", con más del 30 %, seguido por "producto dañado" con aproximadamente un 22 %, "pedido equivocado" con aproximadamente un 17 % y "errores en la facturación" con un 11 %. Del Diagrama de Pareto se observa además que los tres primeros motivos acumulan el 71,1 % de las devoluciones analizadas y, si se consideran también los errores en la facturación, acumulan el 82,6 %.

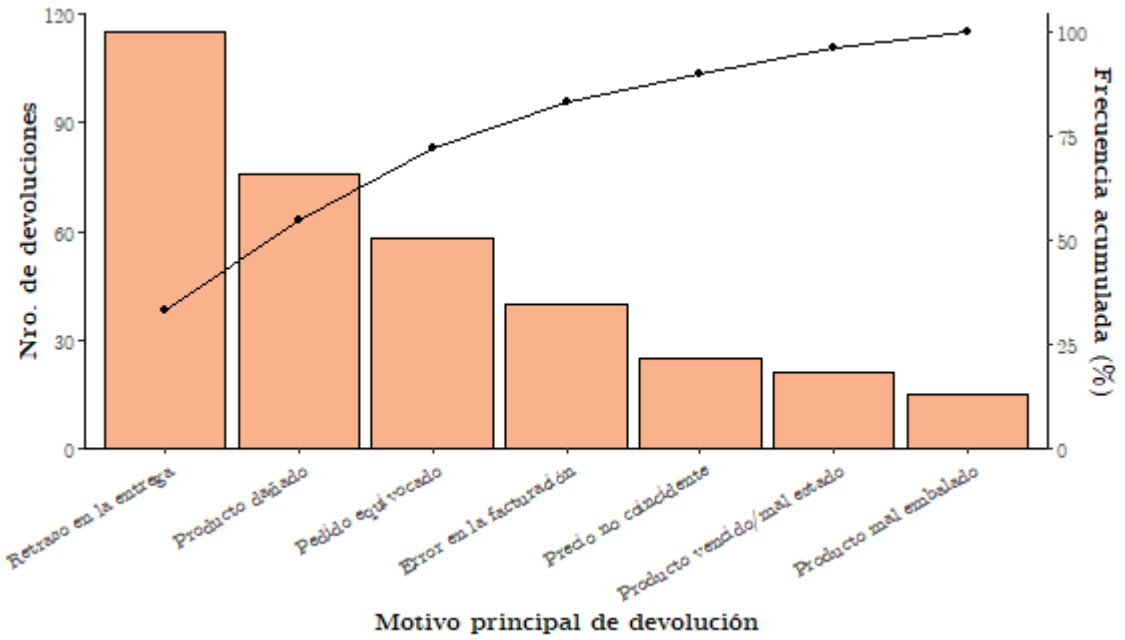


Figura 2.2 *Diagrama de Pareto de los motivos principales de las devoluciones analizadas*

El **diagrama de Pareto** es una de las **"7 herramientas básicas de la calidad"** mencionadas al finalizar el Capítulo 1 (página 30).

El principio que está detrás de este gráfico consiste en distinguir los "pocos vitales" de los "muchos triviales", lo que permite enfocar la atención en aquellas categorías o causas que tienen mayor impacto sobre el problema analizado.

De este modo, el diagrama de Pareto constituye una herramienta muy útil para la priorización de problemas, la toma de decisiones y la planificación de acciones correctivas o de mejora.



En la construcción del diagrama de Pareto, la importancia de cada clase puede venir dada por su frecuencia, como en el caso de la Figura 2.1, o por el costo o el tiempo asociado a cada una, entre otros.

En el Ejemplo 2.1 se presentan dos diagramas de Pareto construidos con diferente información.

Ejemplo 2.1 En una empresa metalúrgica que fabrica turbinas para máquinas agrícolas se realizó un relevamiento de los problemas detectados durante el año anterior, con el objeto de planificar acciones correctivas y mejoras en los procesos involucrados. Se encontraron 30 rodetes defectuosos, 470 piezas mal pintadas (se desprende la pintura), 70 piezas con problemas de torque, 40 piezas con óxido, 32 piezas con el galvanizado en mal estado y 120 piezas con presencia de poros, lo que totaliza 762 problemas. Con esa información se construyó el diagrama de Pareto que se presenta en la Figura 2.3.

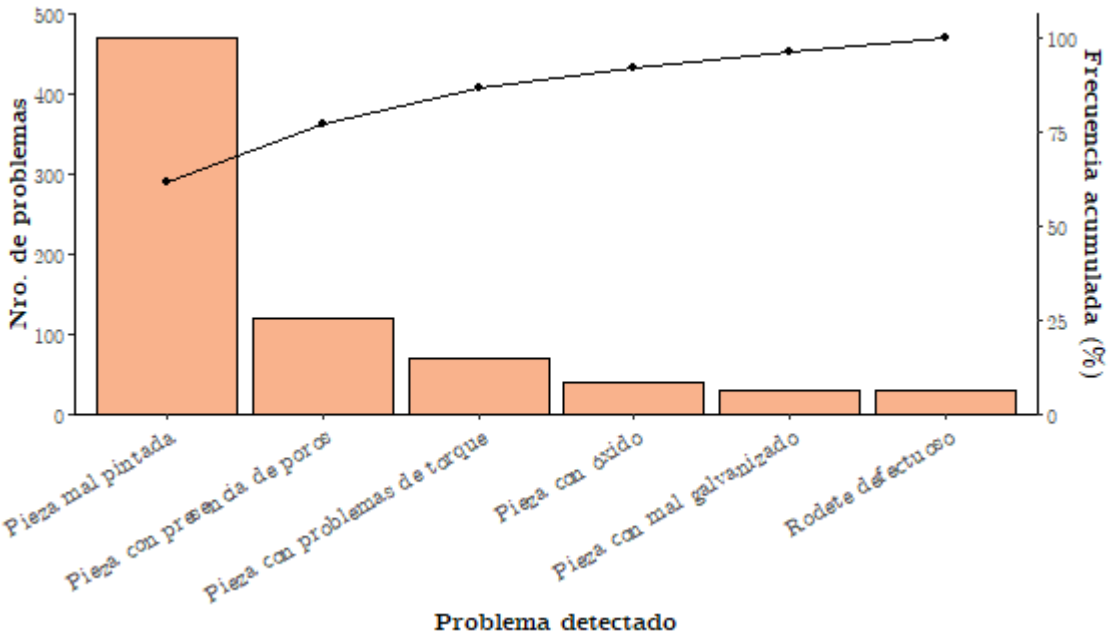


Figura 2.3 *Diagrama de Pareto de los problemas detectados en la fabricación de turbinas durante el año anterior, según su frecuencia de aparición*

Del diagrama se observa que el principal problema a corregir se asocia al proceso de pintura de las piezas, el cual representa aproximadamente el 62 % de los problemas observados durante el año anterior. En orden de importancia, según su frecuencia de aparición, le siguen las piezas con presencia de poros, que representan aproximadamente el 16 % y las piezas con problemas de torque, que representan aproximadamente el 9 %. Estos tres problemas representan más del 80 % de todos los detectados durante ese año y resultaría razonable diseñar estrategias para solucionarlos y para prevenirlos en años sucesivos.

Una de las ingenieras de la empresa no quedó conforme con los resultados de este primer análisis, ya que consideró que los costos asociados a cada una de los problemas detectados era muy diferente. Solicitó entonces al Departamento de Costos de la empresa que le brindara un valor aproximado para el costo unitario de cada uno de estos problemas (ya sea el costo que implican las acciones correctivas o el costo de la pérdida del material, en caso que las piezas o las turbinas deban desecharse). En la Tabla 2.6 se presenta la información obtenida a partir del pedido de la ingeniera.

Tabla 2.6 *Problemas detectados en la fabricación de turbinas durante el año anterior y costos asociados*

Problema	Frecuencia de aparición	Costo unitario (en $)	Costo total (en $)
Rodete defectuoso	30	380000	11400000
Pieza mal pintada	470	18500	8697000
Pieza con problemas de torque	70	1000	70000
Pieza con óxido	40	30000	1200000
Pieza con mal galvanizado	32	70000	2240000
Pieza con poros	120	67000	8040000
Total	**762**	–	**31645000**

Observación: el costo de un rodete defectuoso es muy elevado ya que este no se puede reprocesar, debiendo descartarse la pieza completa.

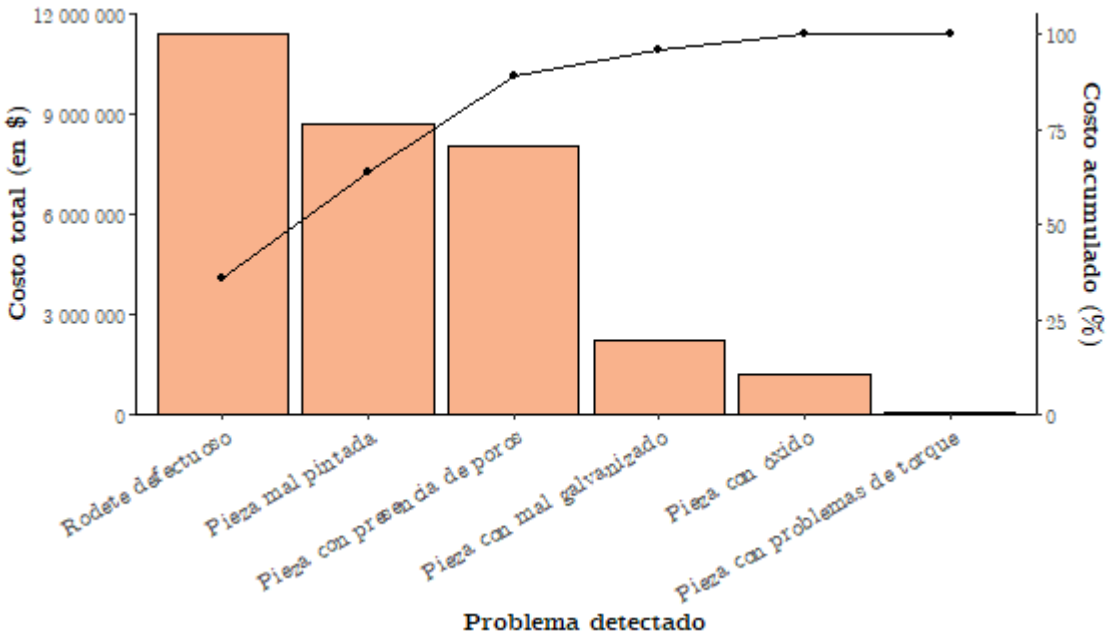


Figura 2.4 *Diagrama de Pareto de los problemas detectados en la fabricación de turbinas durante el año anterior, según los costos asociados*

En la Figura 2.4 se construye el diagrama de Pareto considerando el costo total aproximado de cada problema durante el año anterior. Con esta información se observa que el rodete defectuoso constituye el principal problema a resolver ya que, si bien solo representa aproximadamente el 4 % del total de problemas observados, en relación a los costos representa el 36 %. Le siguen, en orden de importancia en relación a los costos, el problema de la pintura, que representa el 27 % del costo total y el problema de las piezas con presencia de poros, que representan el 25 %. Estos tres problemas representan casi el 90 % del costo total de los problemas presentados el año anterior.

Observe que los problemas de la pintura y de la presencia de poros también fueron detectados en el diagrama de Pareto original, debido a su alta frecuencia de aparición.

La ingeniera y su equipo a cargo decidieron comenzar a trabajar en diseñar estrategias para reducir los rodetes defectuosos, los problemas de pintura y los problemas de las piezas con poros, buscando reducir los costos de la empresa y las quejas de los clientes.



Con lo visto hasta ahora, se pueden resolver las Actividades 1 y 2, Sección 2.5.

2.2.2 Distribuciones de frecuencias para variables cuantitativas discretas. Representación tabular y gráfica

Cuando se cuenta con una variable *cuantitativa discreta*, las clases generalmente están formadas por un único valor de la variable, aunque pueden asociarse a intervalos si la variable asume un número considerable de valores [2]. En el primer caso, k coincide con el número de valores distintos de la variable y, en el segundo, es menor. Para cada clase se obtienen las frecuencias absolutas y a partir de ellas, las frecuencias relativas y las acumuladas definidas en la Tabla 2.2.

[2] En este último caso, el número de intervalos o clases se elige buscando representar de la mejor manera el patrón de variabilidad de los datos.

La información relativa a las diferentes clases y las frecuencias correspondientes se puede presentar en forma tabular o gráfica: se puede construir una *tabla de distribución de frecuencias*, un *gráfico de bastones* para las frecuencias no acumuladas (n_j o f_j) y un *gráfico escalonado* para las frecuencias acumuladas (N_j o F_j).



Problema 5 (pág. 8): *En este problema, la población está constituida por todos los rollos de 100 m de cable unipolar de 2,5 mm fabricados por la empresa.* ***La variable de interés es Cantidad de imperfecciones por rollo, variable cuantitativa discreta medida de razón.***
Entre los objetivos del estudio se encuentra conocer la proporción de rollos de este tipo que presentan al menos 2 imperfecciones y determinar si esta medida vale al menos 0,90.
Para cumplir con estos objetivos, se analizaron 50 rollos seleccionados al azar entre los producidos en el primer bimestre del corriente año. Para cada uno de ellos, se registró la cantidad de imperfecciones, y_i. Una parte de la planilla de registro se presenta en la Tabla 2.7.

Tabla 2.7 *Planilla de registro de las cantidades de imperfecciones en los rollos analizados*

Fecha de producción	***Responsable***[3]	***Rollo Nro.***	***Cantidad de imperfecciones***	***Observaciones***
02/01/26	*MEF*	*01*	*0*	
02/01/26	*PC*	*02*	*2*	*Problemas en máquina 1*
...	*...*	*...*	*...*	*...*
27/02/26	*MEF*	*50*	*3*	*Problemas en el almacenamiento*

A continuación se presenta una parte de la planilla de volcado. En el archivo https://sites.google.com/view/libro-estadistica se encuentra la planilla completa.

Tabla 2.8 *Planilla de volcado de las cantidades de imperfecciones en los rollos analizados*

Rollo Nro.	***Cantidad de imperfecciones***
01	*0*
02	*2*
...	*...*
50	*3*

En la Tabla 2.9 se presentan las frecuencias absolutas, relativas y acumuladas correspondientes a las cantidades de imperfecciones por rollo, en una muestra de n = 50 rollos.

Tabla 2.9 *Tabla de distribución de frecuencias para las cantidades de imperfecciones de los 50 rollos analizados*

Cantidad de imperfecciones (y_j)	***Cantidad de rollos*** (n_j)	***Proporción de rollos*** (f_j)	***Cantidad acumulada de rollos*** (N_j)	***Proporción acumulada de rollos*** (F_j)
0 (c_1)	*23* (n_1)	*0,46* (f_1)	*23* (N_1)	*0,46* (F_1)
1 (c_2)	*17* (n_2)	*0,34* (f_2)	*40* (N_2)	*0,80* (F_2)
2 (c_3)	*7* (n_3)	*0,14* (f_3)	*47* (N_3)	*0,94* (F_3)
3 (c_4)	*1* (n_4)	*0,02* (f_4)	*48* (N_4)	*0,96* (F_4)
4 (c_5)	*2* (n_5)	*0,04* (f_5)	*50* (N_5)	*1,00* (F_5)
Total	*50* (n)	*1*	–	–

A continuación se interpreta, a modo de ejemplo, la fila correspondiente a la tercera clase, en términos del problema. c_3: 2 imperfecciones por rollo, n_3: 7 rollos, f_3: 0,14 (14 % de los rollos). En 7 de los rollos considerados, se observaron 2 imperfecciones. En términos relativos, en una proporción de 0,14 de los rollos (el 14 % de los rollos, llevado a porcentaje) se observaron 2 imperfecciones.
N_3: 47 rollos, F_3: 0,94 (94 % de los rollos). En 47 de los rollos considerados, se observaron 2 imperfecciones o menos. En términos relativos, en una proporción de 0,94 de los rollos (el 94 % de los rollos, llevado a porcentaje) se observaron como máximo 2 imperfecciones.

[3] Se refiere a la persona encargada de revisar el rollo y de registrar la cantidad de imperfecciones y cualquier otra observación relevante.

1. Gráfico de bastones

En este gráfico, los diferentes valores de la variable, y_j, se presentan en el eje de abscisas y sus frecuencias (n_j o f_j) en el eje de ordenadas. Para cada valor se levanta una línea vertical (bastón) de altura igual a su frecuencia.

2. Gráfico escalonado

En este gráfico, los diferentes valores de la variable, y_j, se presentan en el eje de abscisas y sus frecuencias acumuladas (N_j o F_j) en el eje de ordenadas. Para cada valor y_j se señala un punto cuya ordenada es el valor de la frecuencia acumulada (absoluta o relativa) correspondiente y, a partir de este punto, se traza una línea horizontal hasta el siguiente valor que asume la variable (y_{j+1}). Dado que la variable es discreta, el aspecto del gráfico es como el de una función por tramos:

- en cada uno de los valores de la variable se presentan discontinuidades de salto;
- la magnitud del salto en cada valor de la variable coincide con la frecuencia (absoluta o relativa);
- en el intervalo $[y_j, y_{j+1})$, definido por dos valores consecutivos de la variable, la gráfica toma el valor de la frecuencia acumulada asociada al valor y_j.



Problema 5 (pág. 8): *En la Figura 2.5 se presentan las distribuciones de frecuencias relativas y relativas acumuladas correspondientes a las cantidades de imperfecciones por rollo.*

Del gráfico de bastones presentado en la Figura 2.5.a, se desprende que, aunque el número de imperfecciones varía entre 0 y 4, la mayor proporción de rollos presenta 0 o 1 imperfección (46 % y 34 % respectivamente). Las frecuencias asociadas a 2, 3 y 4 imperfecciones son menores (14 %, 2 % y 4 % respectivamente).

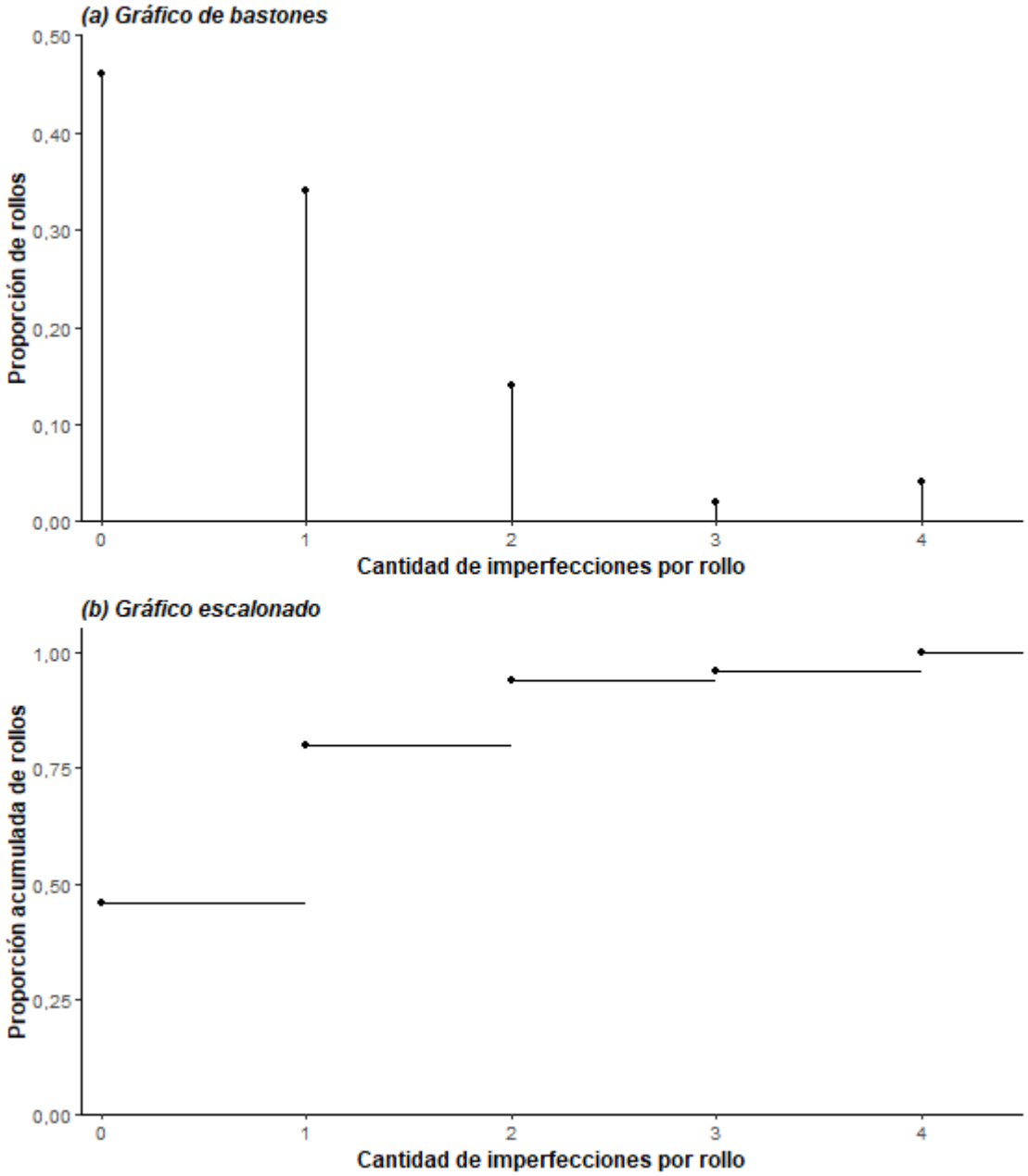


Figura 2.5 *Gráfico de bastones para frecuencias relativas y gráfico escalonado para frecuencias relativas acumuladas, correspondientes a las cantidades de imperfecciones para los rollos analizados*

¿Qué utilidad tiene la información de la quinta columna de la planilla de registro que se presenta en la Tabla 2.7, en el análisis de los datos?

Con lo visto hasta ahora, se pueden resolver las Actividades 3 y 4, Sección 2.5.

2.2.3 Distribuciones de frecuencias para variables cuantitativas continuas. Representación tabular y gráfica

Cuando se trata de variables *cuantitativas continuas*, cada clase está asociada a un intervalo de valores de la variable. Surgen en este caso dos preguntas relacionadas entre sí: *"¿cuántos intervalos considerar?"* y *"¿de qué amplitud deben ser estos intervalos?"*. La idea es lograr la mejor representación del patrón de variabilidad de los datos.

- En relación a la **cantidad de intervalos o clases** a considerar, no hay una única regla para determinarla. En general, esto depende de la cantidad de datos: cuantos más datos se hayan recolectado, se pueden definir más intervalos y estos pueden resultar de menor amplitud. Un criterio que se presenta en diferentes libros de texto es tomar tantos intervalos (o clases) como la raíz cuadrada de la cantidad de datos; pero puede tomarse otro criterio.
- En relación a la **amplitud de los intervalos**, generalmente es mejor que todas las clases tengan igual amplitud. Cuando no es así, se hace más difícil la construcción e interpretación de las tablas y gráficos correspondientes.
- En relación a los **extremos de los intervalos**, se busca que estos sean acotados; pero puede ocurrir que existan intervalos no acotados. Por ejemplo, si se considera la variable "% de contaminación" y el último intervalo de clase es "5 % o más" o bien, si se considera la variable "espesor de capa" en un proceso de cromado de piezas metálicas y el primer intervalo es "hasta 0,1 unidades". Cuando se cuenta únicamente con los datos ya organizados en una tabla de distribución de frecuencias y los intervalos no son acotados, se pierde información sobre el mínimo y/o el máximo valor de la variable y no es posible realizar operaciones que involucren la amplitud de los mismos o su punto medio.

Para determinar la amplitud de clase apropiada, se identifican los valores mínimo ($y_{\text{mín}}$) y máximo ($y_{\text{máx}}$) del conjunto de los datos y se define el número de clases deseado (k). Luego, se realiza el siguiente cociente para determinar la amplitud aproximada de clase:

Amplitud de clase = $(y_{\text{máx}} - y_{\text{mín}}) \;/\; k$.

El valor que se obtiene de la expresión anterior, se redondea buscando que la amplitud sea un número sencillo. Finalmente se definen los intervalos de clase de tal manera que el límite inferior del primer intervalo sea menor o igual que el valor mínimo y que el límite superior del último intervalo sea mayor o igual que el máximo. Este proceso se puede repetir para distintos valores de k, buscando la mejor distribución de frecuencias para resumir los datos. También pueden definirse intervalos de clase de diferente amplitud.

Definidas las clases, se obtienen las frecuencias absolutas correspondientes contando la cantidad de unidades que pertenecen a cada una y luego, el resto de las frecuencias presentadas en la Tabla 2.2.

La información relativa a las diferentes clases y las frecuencias correspondientes se puede presentar en forma tabular o gráfica: se puede construir una *tabla de distribución de frecuencias*, un *histograma* y un *polígono de frecuencias* para las frecuencias no acumuladas (n_j o f_j) y un *polígono acumulativo* para las frecuencias acumuladas (N_j o F_j).

Problema 1 (pág. 7): *Para este problema, la población está constituida por todas las barras metálicas de cierto tipo que se producen en la empresa metalúrgica.* ***La variable de interés es la longitud (mm), variable cuantitativa continua medida en escala de razón****.*
El objetivo del estudio es saber si, con el proceso tal y como está funcionando actualmente, pueden producir barras con longitudes dentro del intervalo (250 +/- 0,6) mm y conocer, al menos en forma aproximada, la proporción de barras que cumplen con ese requerimiento impuesto por la empresa automotriz. Suponga que para cumplir ese objetivo se selecciona una muestra aleatoria de n = 120 barras, tomadas del proceso estabilizado y en cada una de ellas se evalúa su longitud. Una parte de la planilla de registro se presenta en la Tabla 2.10.

Tabla 2.10 *Planilla de registro de las longitudes de las barras analizadas*

Barra N°	***Longitud(mm)***	***Medida por***	***Fecha de la medición***	***Observaciones***
001	*250,07*	*LDB*	*10/07/26*	*Barra medida sin problemas*
002	*250,34*	*LDB*	*10/07/26*	*Barra medida sin problemas*
...	*...*	*...*	*...*	*...*
120	*250,14*	*FJK*	*12/07/26*	*Barra medida sin problemas*

A continuación se presenta una parte de la planilla de volcado. En el archivo https://sites.google.com/view/libro-estadistica se encuentra la planilla completa.

Tabla 2.11 *Planilla de volcado de las longitudes de las barras analizadas*

Barra N°	***Longitud (mm)***
1	*250,07*
2	*250,34*
...	*...*
120	*250,14*

En la Tabla 2.12 se resume el proceso de definición de los intervalos de clase, para 5 valores diferentes de k. Luego, en la Tabla 2.13, se presentan 8 intervalos de clase de diferente amplitud propuestos para este conjunto de datos.

Tabla 2.12 *Proceso de definición de intervalos de clase para las longitudes de las barras analizadas*

Total de clases(k)	***Cálculo de la amplitud de clase (mm)***	***Amplitud elegida (mm)***	***Intervalos obtenidos (mm)***
4	*(250,71 - 249,29) / 4 = 0,355*	*0,4*	*(249,2 ; 249,6]* *(249,6 ; 250,0]* *(250,0 ; 250,4]* *(250,4 ; 250,8]*
8	*(250,71 - 249,29) / 8 = 0,1775*	*0,2*	*(249,2 ; 249,4]* *(249,4 ; 249,6]* *(249,6 ; 249,8]* *(249,8 ; 250,0]* *(250,0 ; 250,2]* *(250,2 ; 250,4]* *(250,4 ; 250,6]* *(250,6 ; 250,8]*
10	*(250,71 - 249,29) / 10 = 0,142*	*0,15*	*(249,25 ; 249,40]* *(249,40 ; 249,55]* *(249,55 ; 249,70]* *(249,70 ; 249,85]* *(249,85 ; 250,00]* *(250,00 ; 250,15]* *(250,15 ; 250,30]* *(250,30 ; 250,45]* *(250,45 ; 250,60]* *(250,60 ; 250,75]*
12	*(250,71 - 249,29) / 12 = 0,118*	*0,12*	*(249,28 ; 249,40]* *(249,40 ; 249,52]* *(249,52 ; 249,64]* *(249,64 ; 249,76]* *(249,76 ; 249,88]* *(249,88 ; 250,00]* *(250,00 ; 250,12]* *(250,12 ; 250,24]* *(250,24 ; 250,36]* *(250,36 ; 250,48]* *(250,48 ; 250,60]* *(250,60 ; 250,72]*
24	*(250,71 - 249,29) / 24 = 0,059*	*0,06*	*(249,28 ; 249,34]* *(249,34 ; 249,40]* *(249,40 ; 249,46]* *...* *(250,54 ; 250,60]* *(250,60 ; 250,66]* *(250,66 ; 250,72]*

Tabla 2.13 *Propuesta de intervalos de clase de diferente amplitud para las longitudes de las barras analizadas*

Intervalo de clase	***Amplitud***
(249,28 ; 249,52]	*0,24*
(249,52 ; 249,64]	*0,12*
(249,64 ; 249,76]	*0,12*
(249,76 ; 249,88]	*0,12*
(249,88 ; 250,00]	*0,12*
(250,00 ; 250,12]	*0,12*
(250,12 ; 250,36]	*0,24*
(250,36 ; 250,72]	*0,36*

Finalmente se decide trabajar con 8 clases (k = 8) de igual amplitud y se consideran los intervalos de amplitud 0,2 mm presentados en la Tabla 2.12. Observe que dichos intervalos son a límites nominales.

En la Tabla 2.14 se presenta la distribución de frecuencias para la longitud de barras metálicas. De la tabla se observa que las longitudes de las barras analizadas oscilan entre 249,2 mm y 250,8 mm. El intervalo que se presenta con mayor frecuencia corresponde a longitudes entre 250 mm y 250,2 mm (25 % de las barras). Le siguen en orden decreciente de frecuencia los intervalos (249,6 ; 249,8] mm y (249,8 ; 250] mm (aproximadamente un 20 % de las barras en cada uno) y el intervalo (250,2 ; 250,4] mm (17,5 % de las barras). El porcentaje de barras con longitudes fuera de especificaciones (menores que 249,4 mm o mayores que 250,6 mm) entre las analizadas es 2,5 % para el caso de longitudes menores y 2,5 % para el caso de longitudes mayores, totalizando un 5 %.

Tabla 2.14 *Tabla de distribución de frecuencias para las longitudes de las 120 barras analizadas*

Longitud (mm) (y_j)	***Nro. de barras*** (n_j)	***Proporción de barras*** (f_j)	***Nro. acumulado de barras*** (N_j)	***Proporción acumulada de barras*** (F_j)
$(249{,}2\ ;\ 249{,}4]\ (c_1)$	$3\ (n_1)$	$0{,}0250\ (f_1)$	$3\ (N_1)$	$0{,}0250\ (F_1)$
$(249{,}4\ ;\ 249{,}6]\ (c_2)$	$8\ (n_2)$	$0{,}0667\ (f_2)$	$11\ (N_2)$	$0{,}0917\ (F_2)$
$(249{,}6\ ;\ 249{,}8]\ (c_3)$	$26\ (n_3)$	$0{,}2166\ (f_3)$	$37\ (N_3)$	$0{,}3083\ (F_3)$
$(249{,}8\ ;\ 250{,}0]\ (c_4)$	$23\ (n_4)$	$0{,}1917\ (f_4)$	$60\ (N_4)$	$0{,}5000\ (F_4)$
$(250{,}0\ ;\ 250{,}2]\ (c_5)$	$30\ (n_5)$	$0{,}2500\ (f_5)$	$90\ (N_5)$	$0{,}7500\ (F_5)$
$(250{,}2\ ;\ 250{,}4]\ (c_6)$	$21\ (n_6)$	$0{,}1750\ (f_6)$	$111\ (N_6)$	$0{,}9250\ (F_6)$
$(250{,}4\ ;\ 250{,}6]\ (c_7)$	$6\ (n_7)$	$0{,}0500\ (f_7)$	$117\ (N_7)$	$0{,}9750\ (F_7)$
$(250{,}6\ ;\ 250{,}8]\ (c_8)$	$3\ (n_8)$	$0{,}0250\ (f_8)$	$120\ (N_8)$	$1\ (F_8)$
Total	$120\ (n)$	1	–	–

A continuación se interpreta, a modo de ejemplo, la fila correspondiente a la tercera clase, en términos del problema. c_3: (249,6 ; 249,8] mm, n_3: 26 barras, f_3: 0,2166 (21,66 % de las barras). 26 de las barras analizadas tienen longitudes entre 249,6 mm y 249,8 mm. En términos relativos, en una proporción de 0,2166 de las barras (el 21,66 % de las barras analizadas, llevado a porcentaje) la longitud de las mismas está entre 249,6 mm y 249,8 mm.

N_3: 37 barras, F_3: 0,3083 (30,83 % de las barras). 37 de las barras analizadas tienen longitudes de hasta 249,8 mm. En términos relativos en una proporción de 0,3083 de las barras (el 30,83 % de las barras analizadas, llevado a porcentaje) la longitud como máximo es 249,8 mm.

1. Histograma

En este gráfico, los intervalos de valores de la variable (es decir, las clases c_j) se presentan en el eje de abscisas y para cada uno se levanta una barra cuya área es proporcional a la frecuencia (n_j o f_j). Dado que la variable es continua, y que el extremo superior de un intervalo de clase coincide con el extremo inferior del intervalo siguiente, las barras no están separadas entre sí.

Si todos los intervalos de clase tienen la misma amplitud, las alturas de las barras resultan proporcionales a las frecuencias (absolutas o relativas) y el gráfico puede incluir al eje de ordenadas para facilitar su lectura, pudiendo realizarse comparaciones entre las clases. En cambio, si las amplitudes de los intervalos de clase son diferentes, las alturas deben ser calculadas para conservar la proporcionalidad entre el área de cada barra y la frecuencia que representa, y debe incluirse una escala de áreas en el gráfico.

Problema 1 (pág. 7): *En la Figura 2.6 se presentan los histogramas construidos para los diferentes conjuntos de clases presentados en las Tablas 2.12 y 2.13. La distribución de frecuencias correspondiente al segundo de ellos se presenta en la Tabla 2.14.*

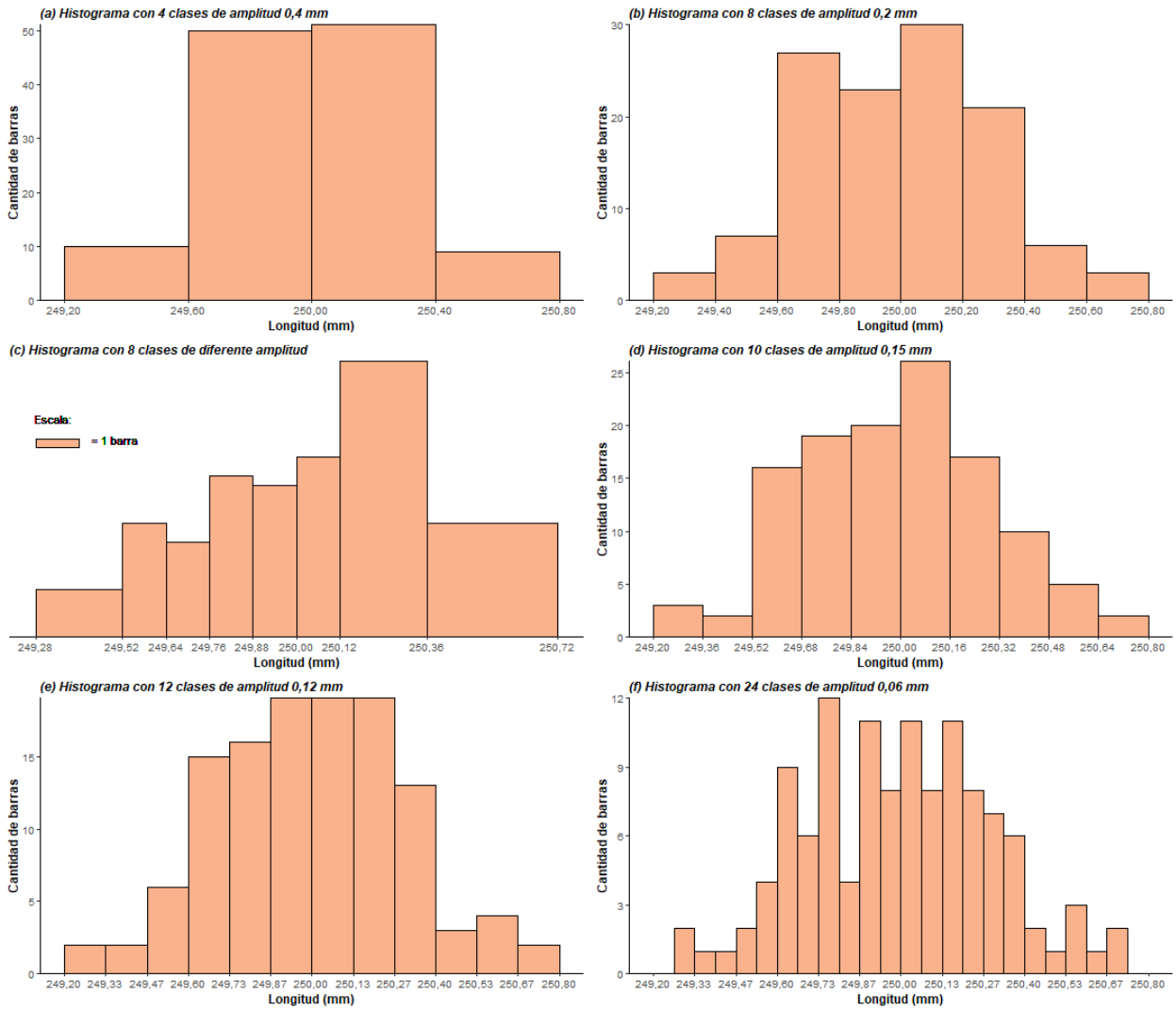


Figura 2.6 *Histogramas (construidos con diferentes conjuntos de clases) de las longitudes de las barras analizadas*

Como se aprecia en la Figura 2.6, la forma de un histograma depende del número de intervalos de clase que se consideren. Cuando se emplean pocos o demasiados intervalos de clase la visualización del histograma no ofrece buena información: en el primer caso (Figura 2.6.a) no se discrimina convenientemente la distribución de los datos y en el segundo caso (Figura 2.6.f), no se alcanza a lograr un patrón de la distribución de los mismos.

En el segundo histograma (Figura 2.6.b) se observa que la distribución de las longitudes de las barras es aproximadamente simétrica, con valores entre 249,2 mm y 250,8 mm. Más frecuentemente se presentan longitudes entre 249,6 mm y 250,4 mm; pero especialmente entre 250 mm y 250,2 mm. La proporción de barras con longitudes fuera de especificaciones (menores que 249,4 mm o mayores que 250,6 mm) es pequeña.
Cuando los intervalos son de menor amplitud (Figuras 2.6.c y 2.6.d) se puede hacer un análisis con más detalle. En todos los histogramas se observa que la distribución es aproximadamente simétrica.

2. Polígono de frecuencias

El **polígono de frecuencias** está referido a un sistema coordenado donde cada vértice tiene por abscisa al punto medio del intervalo de clase correspondiente y por ordenada a la altura de la barra. Para hallar los puntos de iniciación y finalización del polígono, se consideran dos intervalos adicionales de clase (uno anterior al primero y otro posterior al último), de igual amplitud que los restantes y de altura nula. De este modo ambos puntos quedan sobre el eje de abscisas. Por igualdad de triángulos se demuestra que el polígono así construido encierra la misma área que el histograma. Este gráfico muestra el comportamiento de la variable en estudio de manera más sencilla que el histograma.

3. Gráfico de frecuencias acumuladas o polígono acumulativo

Este gráfico está referido a un sistema coordenado donde cada vértice tiene por abscisa al límite superior del intervalo de clase correspondiente y por ordenada a la frecuencia acumulada del mismo. El punto de iniciación se señala en el límite inferior del primer intervalo y su altura es 0; mientras que el punto de finalización se señala en el límite superior del último intervalo y su altura es 1 si se trata de un polígono de frecuencias relativas acumuladas o el total de datos (n o N) si se trata de un polígono de frecuencias absolutas acumuladas.

Problema 1 (pág. 7): *En la Figura 2.7 se presentan el histograma con su polígono de frecuencias y el polígono acumulativo asociados a la distribución de frecuencias de las longitudes de las barras analizadas (Tabla 2.14).*

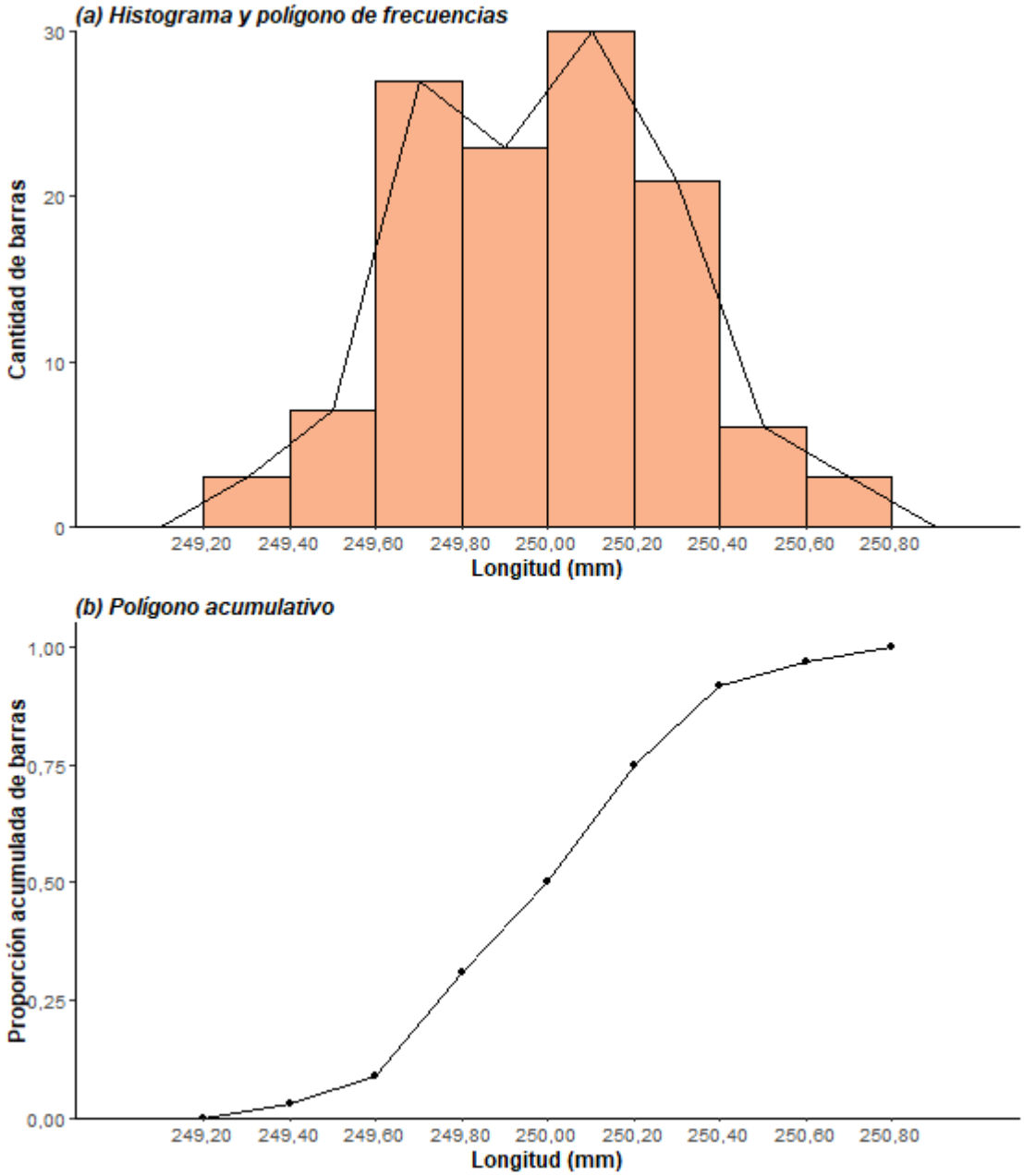


Figura 2.7 *Histograma (con polígono de frecuencias) y polígono acumulativo para frecuencias relativas de las longitudes de las barras analizadas*



Con lo visto hasta ahora, se pueden resolver las Actividades 5 y 6, Sección 2.5.

2.2.4 Otras formas de representar las distribuciones de frecuencias para variables cuantitativas (discretas o continuas)

No siempre es necesario calcular frecuencias para conocer el patrón de variabilidad de un conjunto de datos. Existen gráficos, como el *diagrama de puntos* y el *diagrama de tallo y hoja* que ponen de manifiesto ese patrón utilizando los propios datos. De su lectura se puede observar fácilmente:

- entre qué valores se encuentran las observaciones;
- dónde está centrado, aproximadamente, el conjunto de los datos;
- si hay valores alejados;
- la forma de la distribución (para el caso del diagrama de tallo y hoja se debe rotar el gráfico 90°)y
- el valor aproximado de algunas medidas de resumen que se presentan en la Sección 2.3.

Estos gráficos pueden utilizarse al inicio del análisis de los datos porque dan una idea rápida del patrón de variabilidad (distribución) de los mismos e incluso pueden orientar en la definición de los intervalos de clase, si se trata de variables continuas.

1. Diagrama de puntos

Este diagrama constituye una manera simple de representar un conjunto de datos correspondiente a una variable cuantitativa y poner de manifiesto el patrón de variabilidad de la misma. Para construirlo, cada uno de los valores observados se representa a través de un punto a lo largo de un eje que puede ser horizontal o vertical. Cuando hay más de una observación con el mismo valor observado de la variable, los puntos correspondientes se ubican uno arriba del otro (o al lado, si el eje es vertical).

Si la variable es cuantitativa discreta, el gráfico de puntos coincide con el gráfico de bastones. Si la variable es cuantitativa continua, se prefiere su empleo en vez del histograma cuando el número de datos es escaso.



Problema 1 (pág. 7): *En la Figura 2.8 se muestra el diagrama de puntos para las longitudes de las barras. En este caso, se grafican las longitudes observadas sin necesidad de agruparlas en intervalos de clase.*

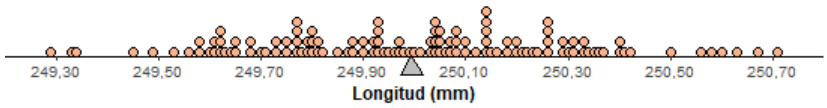


Figura 2.8 *Diagrama de puntos para las longitudes de las barras analizadas*



Problema 5 (pág. 8): *A continuación, se muestra el diagrama de puntos para las cantidades de imperfecciones por rollo. El mismo resulta muy similar al correspondiente diagrama de bastones (Figura 2.5.a).*



Figura 2.9 *Diagrama de puntos para las cantidades de imperfecciones de los rollos analizados*

2. Diagrama de tallo y hoja

Este diagrama también constituye una manera simple de resumir a un conjunto de datos correspondientes a variables cuantitativas discretas o continuas. Para construirlo, cada elemento del conjunto de los datos se divide en dos partes, denominadas "tallo" y "hoja" respectivamente. Luego se organizan los datos en líneas, de tal manera que cada una contenga todos los valores de las "hojas" correspondientes a un mismo "tallo", ordenadas de menor a mayor.

No existe una única forma de elegir tallos y hojas. Lo importante es que en el gráfico se ponga de manifiesto la forma de la distribución. Por ejemplo, si para una cierta variable los valores oscilan entre 40,8 y 50,3, puede considerarse como tallo a la parte entera y como hoja a la cifra decimal; o bien, si los datos son números enteros de 3 cifras, el tronco puede ser la cifra asociada a la centena y las hojas, las cifras asociadas a la decena y a la unidad. Incluso, un mismo tallo se puede dividir en dos o más partes, si es que tiene una gran cantidad de hojas incluidas en él.



Problema 1 (pág. 7): *En la Figura 2.10 se muestra el diagrama de tallo y hoja para las longitudes de las barras, en mm. Este gráfico presenta la información de las longitudes agrupadas en intervalos de amplitud 0,10 mm. Al rotarlo 90° a la izquierda, pone de manifiesto el comportamiento de la variable en estudio.*

```
The decimal point is 1 digit(s) to the left of the |

2492 | 9
2493 | 34
2494 | 59
2495 | 3688
2496 | 001222345588
2497 | 0133557777899
2498 | 0001125788
2499 | 002233334567789
2500 | 01334444555677888
2501 | 002444445688
2502 | 001234666699
2503 | 011334567
2504 | 0012
2505 | 068
2506 | 037
2507 | 1
```

Figura 2.10 *Diagrama de tallo y hoja para las longitudes de las barras analizadas (en mm). La coma decimal se encuentra una posición a la izquierda de la línea vertical*

Observación: En el diagrama, 2492\|9 representa un valor de longitud igual a 249,29 mm.



Problema 5 (pág. 8): *A continuación, se muestra el diagrama de tallo y hoja para las cantidades de imperfecciones por rollo. Este gráfico presenta la misma información que el diagrama de bastones (Figura 2.5.a) y que el diagrama de puntos (Figura 2.9). Al rotarlo 90° hacia la izquierda, pone de manifiesto el comportamiento de la variable en estudio. Note que, como los valores observados para la variable están compuestos por una única cifra, en este caso para las hojas se indica el valor 0.*

```
The decimal point is at the |

0 | 00000000000000000000000
1 | 00000000000000000
2 | 0000000
3 | 0
4 | 00
```

Figura 2.11 *Diagrama de tallo y hoja para las cantidades de imperfecciones de los rollos analizados. La coma decimal se encuentra en la posición de la línea vertical*

Observación: En el diagrama, 0\|0 representa 0 imperfecciones.



Con lo visto hasta ahora, se pueden resolver las Actividades 7 y 8, Sección 2.5.

3. Gráfico de series cronológicas o series de tiempo

El conjunto de valores de una variable, obtenidos a través del tiempo con cierta periodicidad, se denomina **serie cronológica** o **serie de tiempo** y puede representarse gráficamente como se observa en la Figura 2.12. En dichos gráficos el eje vertical se corresponde con la variable de interés y el eje horizontal denota el tiempo (horas, minutos, semanas, trimestres, días, etc.) o el orden en que se realizaron las observaciones o las mediciones.

Problema 9 (pág. 9): *En la empresa fabricante de productos de limpieza, están interesados en conocer el comportamiento de la demanda trimestral de desengrasante en los últimos 5 años para poder predecir valores futuros y planificar la producción para el próximo año. En principio no pueden obtener información de la demanda pero sí de las ventas trimestrales del producto* [4]*, en cantidad de unidades. La información se presenta en el gráfico de series cronológicas de la Figura 2.12.*

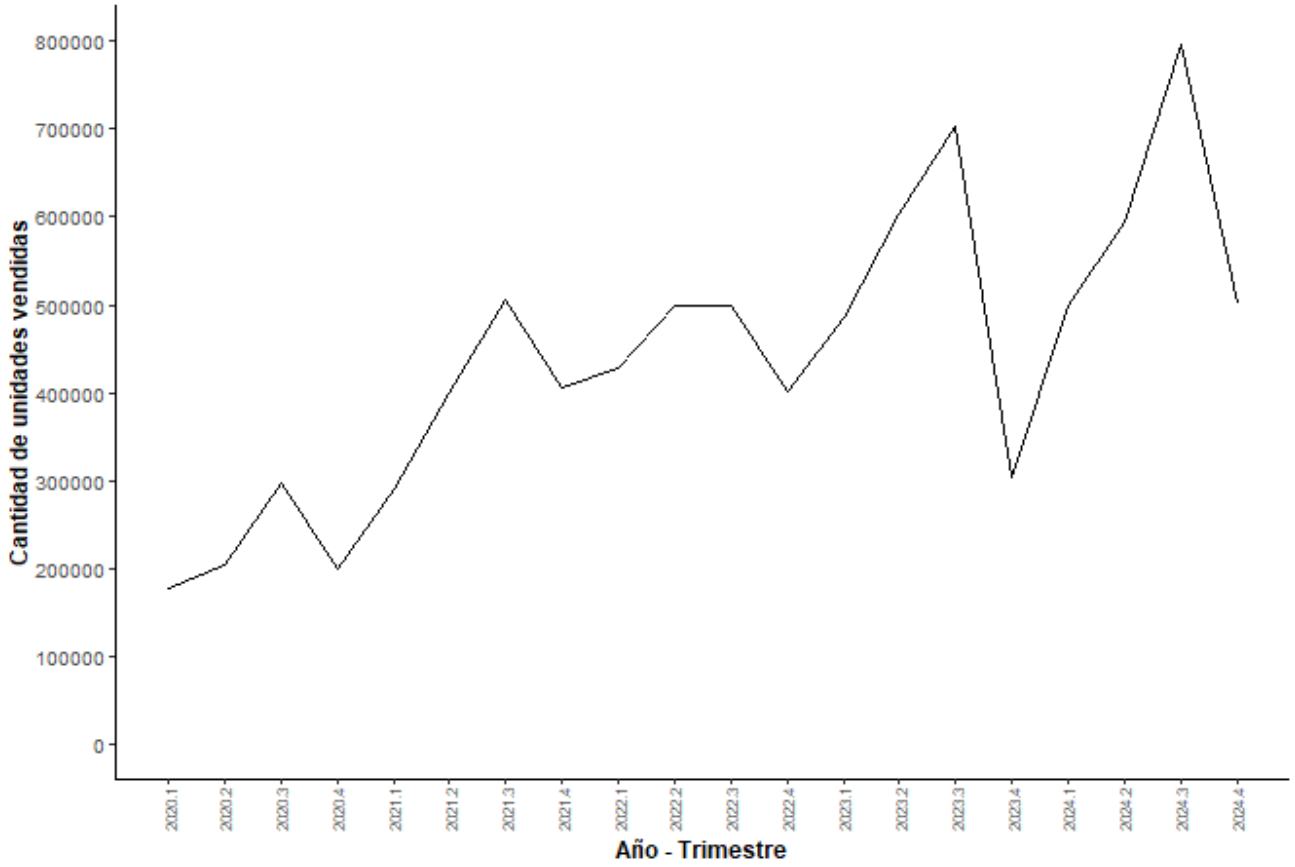


Figura 2.12 *Gráfica de series cronológicas para las ventas trimestrales de desengrasante. 2020-2024*

[4] La cantidad de unidades vendidas trimestralmente es la variable sustituta de la demanda trimestral.

Uno de los principales objetivos por los cuales se cuenta con una serie cronológica para una variable en particular es el de describir el comportamiento de dicha variable en el tiempo y hacer predicciones para períodos futuros. Esto último no se aborda en el presente libro.

La descripción del comportamiento de la variable a través del tiempo permite considerar si este es estable o no, debido a la presencia de tendencias, ciclos, cambios de nivel, etc. En la Figura 2.13 se muestran diferentes comportamientos que pueden presentarse a través del tiempo. En *(a)* se observa una tendencia creciente, en *(b)* se observa un comportamiento cíclico que se repite aproximadamente cada 6 períodos, en *(c)* se observa un cambio de nivel en la variable en estudio y en *(d)* un cambio en la variabilidad. Solo el gráfico *(e)* muestra un comportamiento aproximadamente estable.

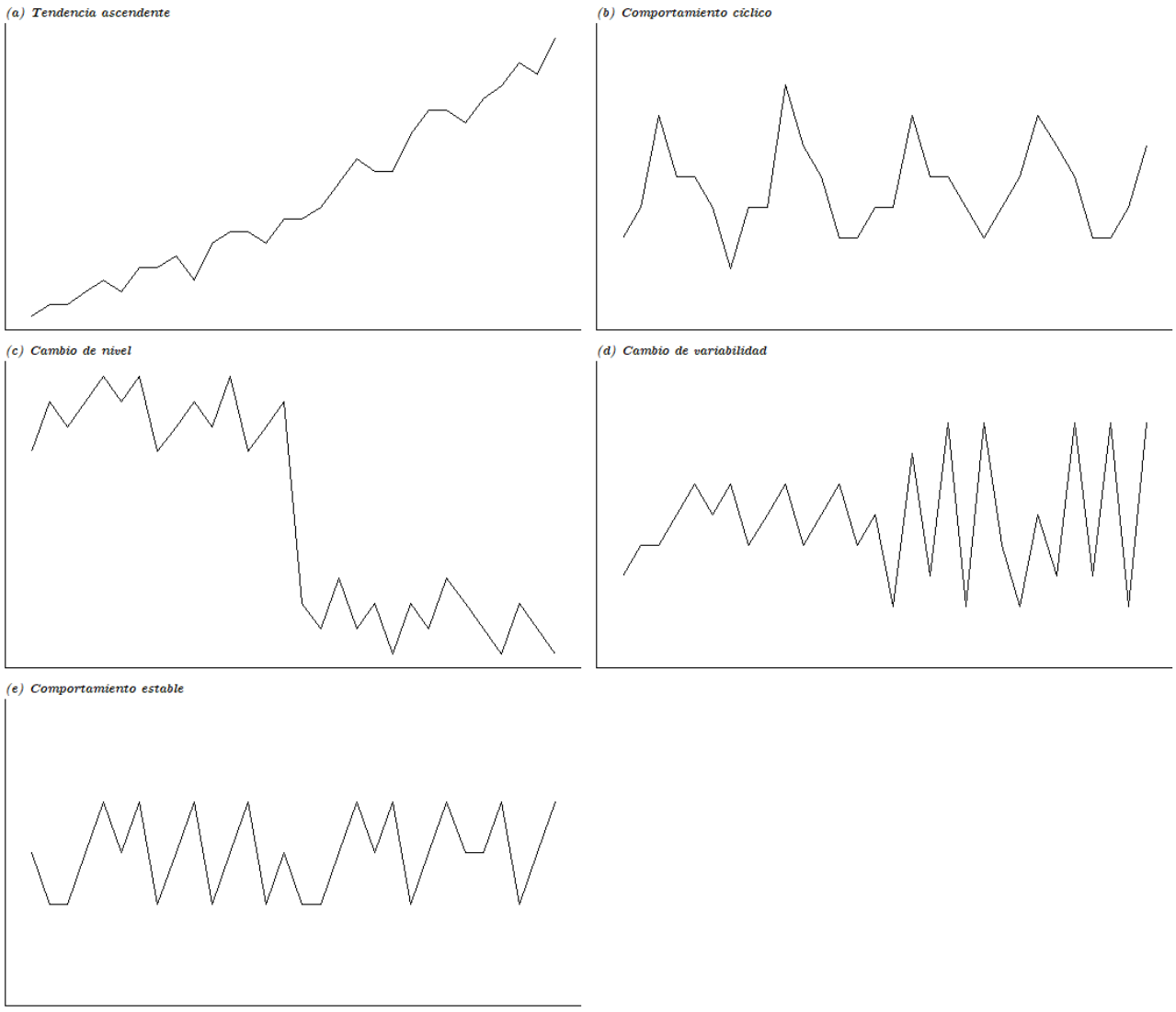


Figura 2.13 *Patrones que pueden observarse en un gráfico de series de tiempo*



Problema 9 (pág. 9): *En el gráfico de la Figura 2.12 se observa una tendencia creciente en la cantidad de unidades vendidas y un comportamiento estacional: en el período considerado se observa que en el cuarto trimestre de cada año las ventas disminuyen para luego aumentar en los tres trimestres restantes.*

Aunque en un estudio no interese el factor tiempo, siempre transcurre tiempo entre una observación y otra: los datos no se obtienen todos en un mismo instante. Por ese motivo, si para un conjunto de datos se cuenta con información del orden en el que las unidades fueron seleccionadas o en el que fueron medidas, antes de construir la distribución de frecuencias o realizar cualquier otro análisis con esos datos, es importante evaluar si el comportamiento de la variable es estable, analizando el gráfico de series cronológicas. Si dicho comportamiento no es estable (se presentan tendencias, cambios de nivel, ciclos, etc.) carece de sentido construir la tabla de distribución de frecuencias u obtener medidas de resumen para la totalidad de los datos recolectados.

En particular, cuando se cuenta con una muestra aleatoria simple de una población determinada y se conoce el orden en que se obtuvieron esas observaciones, el gráfico de series cronológicas correspondiente permite analizar el cumplimiento de propiedades asociadas a este tipo de muestras: si en verdad la muestra es aleatoria simple, no debería presentarse ningún patrón y el comportamiento de

las observaciones según el orden en el que fueron obtenidas debería ser estable. Esta temática se va a profundizar en los Capítulos 7, 8 y 9.

Problema 9 (pág. 9): *Si el gráfico de la Figura 2.12 hubiera mostrado un comportamiento estable, la información de las ventas trimestrales de desengrasante en el período 2020-2024 se podría considerar una muestra aleatoria simple. Pero esto no es así y por lo tanto no se pueden aplicar las herramientas que se presentan en los Capítulos 8 y 9 para realizar inferencias sobre parámetros como la cantidad promedio vendida por trimestre, entre otros.*
Existen técnicas que permiten predecir ventas futuras, con cierto margen de error, pero estas no se desarrollan en el presente libro.

Con lo visto hasta ahora, se pueden resolver las Actividades 9 y 10, Sección 2.5.

2.3 Medidas de resumen de una variable

En las secciones anteriores se proponen tablas y gráficos para analizar el patrón de variabilidad de alguna variable, a partir de un conjunto de datos. Ambas herramientas brindan abundante información sobre el comportamiento de la variable en estudio; pero hacen falta indicadores que resuman esa información numéricamente. Estos indicadores, que también reciben el nombre de "medidas de resumen", informan sobre las frecuencias relativas asociadas a distintos valores de la variable (proporciones), la localización del conjunto de los datos, la dispersión o variabilidad de los mismos y la forma de la distribución de frecuencias [5].

Para cada una de las medidas se presenta su concepto y sus principales propiedades. Asimismo se incluye la fórmula correspondiente para calcular su valor a partir de un conjunto de n observaciones $(y_1, y_2, \ldots, y_n)$, junto con la simbología utilizada para representar a dicho valor.

Recuerde que cuando estos indicadores o medidas de resumen se obtienen a partir de los valores de todas las unidades de la población se denominan **parámetros**; mientras que si se obtienen a partir de los valores de un subconjunto de n unidades, se denominan **estadísticos**, utilizándose en cada caso una notación diferente.

En este capítulo se presentan las medidas de resumen considerando, a modo de ejemplo, datos provenientes de una muestra dada y por ello, se utiliza principalmente la notación correspondiente a los estadísticos.

Para ejemplificar el cálculo de las medidas y su interpretación se trabaja con el conjunto de 120 observaciones, correspondientes a las longitudes de las barras seleccionadas (Problema 1, pág. 7). En este caso, la población está compuesta por todas las barras de cierto tipo que se producen en la empresa metalúrgica y la variable es: *longitud (mm)* (variable cuantitativa continua, medida en escala de razón). El tamaño de la muestra (n) es de 120 barras.

En el Apéndice, Sección 2.7.1 se presentan todas las medidas obtenidas a partir de los datos recolectados para la resolución de los Problemas 1, 4 y 5.

[5] En este libro no se van a definir indicadores o medidas que den cuenta sobre la forma de la distribución de los datos. Esta se va a analizar a través de la observación de los gráficos correspondientes.

2.3.1 Proporción de unidades que satisfacen cierta condición

Una medida útil en muchos problemas es la **proporción de unidades que satisfacen cierta condición**. Esta medida se puede obtener para cualquier tipo de variable, a partir de definir alguna condición de interés asociada a algún conjunto de valores de dicha variable.

Definida la condición, surgen las dos clases posibles:

- c_1: formada por el conjunto de valores de la variable que satisfacen la condición buscada,
- c_2: formada por el conjunto de valores de la variable que no la satisfacen.

La proporción buscada no es otra cosa que la frecuencia relativa asociada a la clase de interés, c_1 (Ver definición, fórmula y propiedades de la frecuencia relativa en Tabla 2.2).

En este libro, la proporción de unidades que satisfacen cierta condición en una muestra dada se representa mediante f_o.



Problema 1 (pág. 7): *Condición: Las barras no cumplen con lo pretendido por la industria automotriz, es decir, presentan longitudes menores que 249,4 mm o mayores que 250,6 mm.*
Proporción de barras cuya longitud no cumple con lo pretendido por la industria automotriz, en la muestra:

$f_o = (3+3) \,/\, 120 = 6 \,/\, 120 = 0{,}05.$

Entre las barras analizadas, un 5 % no cumplen con lo pretendido por la industria automotriz.



Problema 4 (pág. 8): *Condición: El motivo principal de la devolución es que el producto estaba mal embalado.*
Proporción de devoluciones debidas a que el producto estaba mal embalado, en la muestra:

$f_o = 15 \,/\, 350 = 0{,}043.$

Entre las devoluciones analizadas, el 4,3 % tuvo como motivo principal el estar mal embalado.

2.3.2 Medidas de localización o posición

Una característica importante de un conjunto de datos es su posición o localización en el eje de valores de la variable. Dentro de este grupo de medidas, se encuentran algunas como el mínimo, el máximo y algunos percentiles, que localizan al conjunto de los datos, y otras como el promedio o media (aritmética, geométrica, truncada), la mediana y la moda que localizan al centro de ese conjunto, desde distintos puntos de vista. A este último conjunto de medidas también se lo denomina **medidas de tendencia central**.

En resumen, las medidas de localización presentadas en esta sección pueden clasificarse en dos grupos:

- las de tendencia central, que intentan localizar al centro del conjunto de los datos;
- las de tendencia no central, que localizan a otros valores no centrales.

1. Valores mínimo y máximo

El **valor mínimo**, $y_{\text{mín}}$, es el menor valor observado en el conjunto de los datos.

El **valor máximo**, $y_{\text{máx}}$, es el mayor valor observado en el conjunto de los datos.

Ambas medidas se pueden definir para variables medidas en nivel ordinal o superior y se obtienen luego de ordenar a los datos y de observar cuáles son el menor y el mayor valor de ese conjunto, respectivamente.

El conocimiento de los valores mínimo y máximo de un conjunto de datos, permite localizar al conjunto en su totalidad y resulta de utilidad en muchos campos. Por ejemplo, en Hidrología, puede interesar la altura máxima de un río en cierto punto, o bien el nivel máximo de precipitaciones en cierta zona, con el objeto de planificar adecuadamente obras civiles para dicha zona. En Tecnología de los Materiales, puede interesar la resistencia mínima de cierto material, con el objeto de definir su uso para la construcción de tanques o autoclaves.

2. Percentiles

Se dice que p_α es el **percentil de orden** α de un conjunto de datos si la frecuencia acumulada relativa porcentual, obtenida hasta el valor p_α es $\alpha\,100\,\%$. En otras palabras, p_α es el valor de la variable que acumula el $\alpha\,100\,\%$ de las observaciones ordenadas.

Los percentiles [6] conforman un conjunto de medidas que se obtienen luego de ordenar a los datos de menor a mayor y por lo tanto se pueden calcular para variables medidas en escala ordinal o superior.

Existen diferentes algoritmos para obtener percentiles a partir de un conjunto ordenado de datos. A modo de ejemplo, se presenta uno de ellos:

Dado un valor de α se obtiene, en primer lugar, el orden del percentil correspondiente, de la siguiente manera: $(p_\alpha)^\circ = \alpha n$. La parte entera de dicho valor se simboliza con E.

- Si la parte decimal de $(p_\alpha)^\circ$ no es nula, $p_\alpha = y_{[E+1]}$
- Si la parte decimal de $(p_\alpha)^\circ$ es nula, $p_\alpha = \frac{1}{2}(y_{[E]} + y_{[E+1]})$

donde $y_{[i]}$ se refiere al valor de la variable ubicado en la posición i en el conjunto ordenado de datos.

Si los datos están organizados en una tabla de distribución de frecuencias y las clases están compuestas por un único valor de la variable (caso que ocurre generalmente para variables discretas), otra forma de obtener el valor del percentil de orden α es asignarle a p_α el valor de la variable correspondiente a la primera clase que acumula como mínimo el $\alpha\,100\,\%$ de los datos. Si las clases se corresponden con intervalos, se puede tomar como aproximación el punto medio de esa primera clase.

Los percentiles también pueden obtenerse a partir del gráfico escalonado (caso discreto) o del polígono acumulativo (caso continuo), como se observa en la Figura 2.14. En el caso de variables continuas, se obtiene un valor aproximado del percentil.

De acuerdo al porcentaje de observaciones que acumulan, algunos percentiles reciben nombres especiales y se simbolizan de manera diferente. Entre ellos se puede mencionar a los *cuartiles* y a la

[6] Los percentiles también se denominan cuantiles.

mediana. Los cuartiles son percentiles asociados a la división del conjunto ordenado en cuatro partes iguales y la mediana a la división del conjunto en dos.

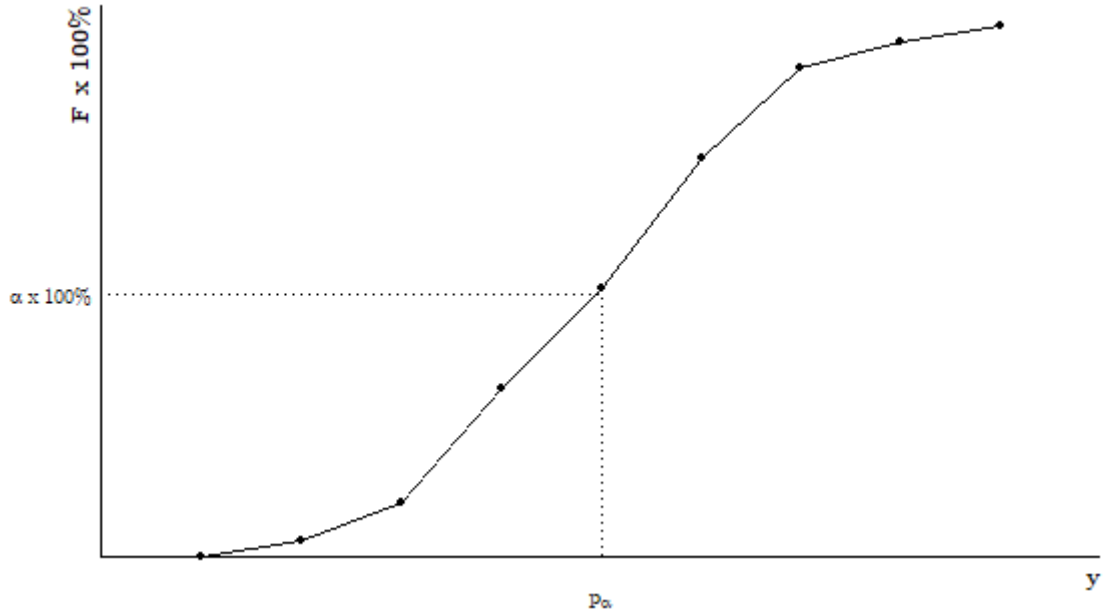


Figura 2.14 *Determinación aproximada del valor de un percentil a partir del polígono acumulativo*

La **mediana** ($p_{0,50}$, $\tilde{y}$ o q_2) es el valor de la variable que acumula el 50 % de las observaciones, ordenadas de menor a mayor, es decir, es el percentil 0,50 de la distribución.

El **primer cuartil** ($p_{0,25}$ o q_1) es el valor de la variable que acumula el 25 % de las observaciones, ordenadas de menor a mayor, es decir, es el percentil 0,25 de la distribución.

El **tercer cuartil** ($p_{0,75}$ o q_3) es el valor de la variable que acumula el 75 % de las observaciones, ordenadas de menor a mayor, es decir, es el percentil 0,75 de la distribución.

En la Figura 2.15 se presentan a modo de ejemplo el percentil de orden 0,10 ($p_{0,10}$) y el percentil de orden 0,25 ($p_{0,25}$ o primer cuartil, q_1), junto con el porcentaje de unidades de la muestra con valores menores o iguales que cada uno de ellos.



Figura 2.15 *Percentiles de orden 0,10 y 0,25 y porcentaje acumulado de observaciones*

Entre todos los percentiles, la mediana o percentil 0,50 es una medida de tendencia central ya que busca localizar el centro del conjunto de los datos. La principal característica de esta medida es que no se ve afectada por la presencia de valores extremos o atípicos. Se dice que es una **medida robusta** o **no sensible a valores atípicos**.

Problema 1 (pág. 7): *A continuación se obtienen e interpretan el mínimo, el máximo y algunos percentiles correspondientes a las longitudes de las barras analizadas.*
En la Tabla 2.15 se presentan las 120 observaciones correspondientes a la longitud de las barras analizadas, ordenadas de menor a mayor.

Tabla 2.15 *Longitudes de las barras analizadas, ordenadas de menor a mayor*

Orden	Longitud (mm)	Orden	Longitud (mm)	Orden	Longitud (mm)	Orden	Longitud (mm)
1	**249,29**	**31**	**249,77**	**61**	**250,01**	**91**	**250,21**
2	249,33	32	249,78	62	250,03	92	250,22
3	249,34	33	249,79	63	250,03	93	250,23
4	249,45	34	249,79	64	250,04	94	250,24
5	249,49	35	249,8	65	250,04	95	250,26
6	249,53	36	249,8	66	250,04	96	250,26
7	249,56	37	249,8	67	250,04	97	250,26
8	249,58	38	249,81	68	250,05	98	250,26
9	249,58	39	249,81	69	250,05	99	250,29
10	249,6	40	249,82	70	250,05	100	250,29
11	249,6	41	249,85	71	250,06	101	250,3
12	249,61	42	249,87	72	250,07	102	250,31
13	249,62	43	249,88	73	250,07	103	250,31
14	249,62	44	249,88	74	250,08	104	250,33
15	249,62	45	249,9	75	250,08	105	250,33
16	249,63	46	249,9	76	250,08	106	250,34
17	249,64	47	249,92	77	250,1	107	250,35
18	249,65	48	249,92	78	250,1	108	250,36
19	249,65	49	249,93	79	250,12	109	250,37
20	249,68	50	249,93	80	250,14	110	250,4
21	249,68	51	249,93	81	250,14	111	250,4
22	249,7	52	249,93	82	250,14	112	250,41
23	249,71	53	249,94	83	250,14	113	250,42
24	249,73	54	249,95	84	250,14	114	250,5
25	249,73	55	249,96	85	250,15	115	250,56
26	249,75	56	249,97	86	250,16	116	250,58
27	249,75	57	249,97	87	250,18	117	250,6
28	249,77	58	249,98	88	250,18	118	250,63
29	249,77	59	249,99	89	250,2	119	250,67
30	**249,77**	**60**	**250**	**90**	**250,2**	**120**	**250,71**

Mínimo y máximo
El valor mínimo ($y_{\text{mín}}$) es 249,29 mm y el máximo ($y_{\text{máx}}$) es 250,71 mm (sombreados en la Tabla 2.15). Es decir que en esa muestra de 120 barras, la menor de ellas medía 249,29 mm y la mayor medía 250,71 mm.

Cuartiles 1 y 3
Orden de $p_{0,25}$ o q_1: $(q_1)^\circ = 0,25 \cdot 120 = 30$. Como la parte decimal es nula, q_1 se obtiene como el promedio de los valores que están en las posiciones [30] y [31] (sombreados en la Tabla 2.15).
q_1 = (249,77 + 249,77) / 2 = 249,77.
Esto se interpreta de la siguiente manera: para la muestra obtenida, el 25 % de las barras miden 249,77 mm o menos y el 75 % restante, 249,77 mm o más.
Orden de $p_{0,75}$ o q_3: $(q_3)^\circ = 0,75 \cdot 120 = 90$. Como la parte decimal es nula, q_3 se obtiene como el promedio de los valores que están en las posiciones [90] y [91] (sombreados en la Tabla 2.15).
q_3 = (250,2 + 250,21) / 2 = 250,205.
Esto se interpreta de la siguiente manera: para la muestra obtenida, el 75 % de las barras miden 250,205 mm o menos y el 25 % restante, 250,205 mm o más.

Mediana
Orden de $p_{0,50}$, q_2 o $\tilde{y}$: $(p_{0,50})^\circ = 0,50 \cdot 120 = 60$. Como la parte decimal es nula, $p_{0,50}$ se obtiene como el promedio de los valores que están en las posiciones [60] y [61] (sombreados en la Tabla 2.15).
$p_{0,50}$ = (250,00 + 250,01) / 2 = 250,005.
Esto se interpreta de la siguiente manera: para la muestra obtenida, el 50 % de las barras miden 250,005 mm o menos y el 50 % restante, 250,005 mm o más.

3. Promedio o media aritmética

La **media aritmética**, $\bar{y}$, se define como la suma de todas las observaciones dividida por el total de observaciones.

Se trata de una medida que puede obtenerse para variables cuantitativas medidas en escala de intervalo o de razón.

El promedio se obtiene aplicando directamente la fórmula:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

donde $y_1, y_2, \ldots, y_n$ constituyen el conjunto de n observaciones y n es el tamaño de la muestra.

Si los datos están organizados en una tabla de distribución de frecuencias y cada clase se corresponde con un único valor de la variable, el promedio puede obtenerse de la siguiente manera:

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{k} y_j n_j = \sum_{j=1}^{k} y_j f_j$$

donde $y_1, y_2, \ldots, y_k$ son los k valores distintos que asume la variable en estudio; $n_1, n_2, \ldots, n_k$ son las frecuencias absolutas correspondientes; $f_1, f_2, \ldots, f_k$ son las frecuencias relativas correspondientes y n es el tamaño del conjunto de los datos.

Si los datos están organizados en una tabla de distribución de frecuencias, pero cada clase corresponde a un intervalo de valores de la variable, en la expresión anterior se reemplaza el valor de y_j por el punto medio del intervalo de clase, y'_j. Si solo se cuenta con la tabla de distribución de frecuencias y en ella hay intervalos no acotados, la media aritmética no se puede calcular.

El promedio es una medida que tiene las siguientes características:

- Toma en consideración toda la información (todos los valores entran en el cálculo) y por lo tanto es muy sensible a la influencia de los valores extremos o atípicos, como se observa en el Ejemplo 2.2. Esto puede ser una ventaja o desventaja, según el caso.
- Puede considerarse como el punto de equilibrio de la distribución o conjunto de datos como se muestra, por ejemplo, en la Figura 2.13.
- No coincide necesariamente con un valor de la variable.
- Es una medida útil para comparar dos o más distribuciones, solo si estas tienen formas semejantes.

Ejemplo 2.2 Suponga que se cuenta con los siguientes datos: 10; 12; 8; 15 y 13.

Si se quiere obtener el promedio o media aritmética, hay que sumar los 5 valores mencionados y dividir ese total por 5. En este caso, el promedio vale:

$$\bar{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = \frac{1}{5} \cdot 58 = 11,6$$

Suponga ahora que el valor 15 se reemplaza por el valor 35, que está alejado del conjunto original de valores. ¿Qué ocurre con el valor del promedio ahora? El nuevo promedio vale:

$$\bar{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = \frac{1}{5} \cdot 78 = 15,6$$

¿Qué hubiera pasado si en vez del promedio se hubiera obtenido la mediana o $p_{0,50}$?

Dado que se trata de $n = 5$ datos, el orden del percentil 0,50, $(p_{0,50})^\circ = 0,50 \cdot 5 = 2,5$. Como este valor tiene cifras decimales no nulas, $p_{0,50} = y_{[2+1]} = y_{[3]}$. Es decir que el percentil 0,50 es el tercer valor, en el conjunto ordenado de datos.

Buscando en el conjunto de datos ordenados: 8; 10; 12; 13; 15, se puede observar que $p_{0,50} = y_{[3]} = 12$.

Observe que, aunque el valor 15 se reemplace por el valor 35, la mediana conserva el mismo valor, ya que el tercer valor sigue siendo 12. Los dos conjuntos de datos se presentan en la Figura 2.16.

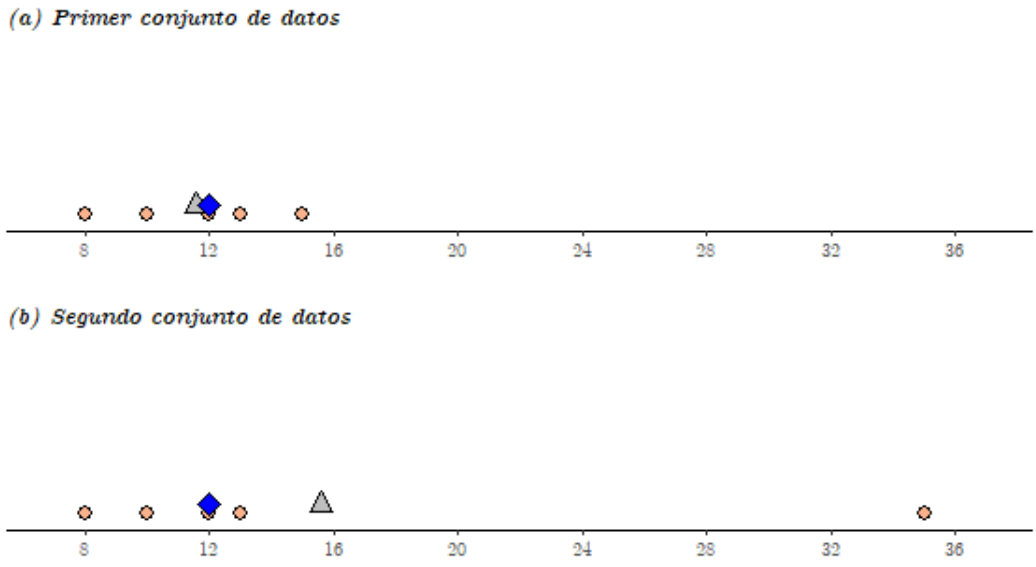


Figura 2.16 *Promedio (triángulo) y mediana (rombo) de los dos conjuntos de datos analizados en el Ejemplo 2.2*

En resumen, el promedio es una medida sensible a la presencia de valores extremos o demasiado alejados del conjunto, ya que su valor se ve afectado por ellos; mientras que la mediana no.

A partir de lo observado, si se calculan ambas medidas para un conjunto de datos y estas presentan valores muy diferentes, se trata de una indicación de la presencia de valores muy alejados o atípicos en ese conjunto.



Problema 1 (pág. 7): *En el diagrama de puntos de la Figura 2.17 se señala el valor del promedio o media aritmética, punto de equilibrio de la distribución.*

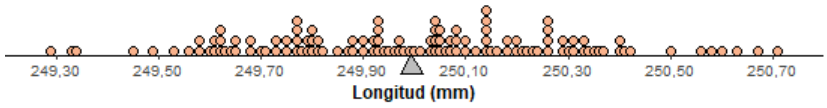


Figura 2.17 *La media aritmética o promedio es el punto de equilibrio de la distribución*

Para obtener el promedio o media aritmética de las longitudes de las barras analizadas, se aplica la primera expresión de la página 73, es decir, se suman los 120 valores obtenidos y se divide ese total por 120.

$$\bar{y} = \frac{1}{120}\sum_{i=1}^{120} y_i = \frac{1}{120}\cdot(249,29 + 249,33 + \ldots + 250,67 + 250,71) = \frac{1}{120}\cdot 29999,31 = 249,994$$

Este valor se puede interpretar de la siguiente manera: en promedio, las barras seleccionadas miden 249,994 mm.

Si solo se contara con las 120 observaciones organizadas en la tabla de distribución de frecuencias, una aproximación del promedio o media aritmética se obtiene aplicando la segunda expresión de la página 73. En la Tabla 2.16 se presenta el cálculo del numerador, el cual se divide por 120.

$$\bar{y} = \frac{1}{120}\sum_{j=1}^{8} y'_j n_j = \frac{1}{120}\cdot 29998{,}2 = 249{,}985$$

Tabla 2.16 *Cálculos requeridos para obtener el valor aproximado de la media aritmética de las longitudes de las barras analizadas*

Longitud (mm) (y_j)	***Punto medio del intervalo*** (y'_j) ***(mm)***	***Nro. de barras*** (n_j)	$y'_j \cdot n_j$ ***(mm)***
(249,2 ; 249,4]	*249,3*	*3*	*747,9*
(249,4 ; 249,6]	*249,5*	*8*	*1996*
(249,6 ; 249,8]	*249,7*	*26*	*6492,2*
(249,8 ; 250,0]	*249,9*	*23*	*5747,7*
(250,0 ; 250,2]	*250,1*	*30*	*7503*
(250,2 ; 250,4]	*250,3*	*21*	*5256,3*
(250,4 ; 250,6]	*250,5*	*6*	*1503*
(250,6 ; 250,8]	*250,7*	*3*	*752,1*
Total		***120 (n)***	***29998,2***

Si se tienen los promedios o medias aritméticas de varios subconjuntos de datos y se quiere obtener un promedio general, debe calcularse la **media o promedio ponderado**. Su expresión está dada por:

$$\bar{y}_p = \frac{\sum_{j=1}^{k} \bar{y}_j n_j}{\sum_{j=1}^{k} n_j}$$

donde k es la cantidad de subconjuntos, e $\bar{y}_j$ y n_j son la media y el número de elementos del subconjunto j ($j = 1, \ldots, k$), respectivamente.



Problema 1 (pág. 7): *Suponga que en la empresa metalúrgica se producen tres tipos de barras: las que posiblemente interesan a la industria automotriz (barras tipo A), barras roscadas (barras tipo B) y barras laminadas (barras tipo C). Además de la muestra de n = 120 barras para la industria automotriz, se tomó una muestra de n = 100 barras roscadas y una muestra de n = 80 barras laminadas y en cada caso se obtuvo la longitud promedio. La información se presenta en la Tabla 2.17.*

Tabla 2.17 *Longitud media de las barras analizadas, según tipo*

Tipo	***Nro. de barras seleccionadas*** (n_j)	***Longitud promedio (mm)*** $(\bar{y}_j)$
A	*120*	*249,994*
B	*100*	*200,103*
C	*80*	*212,315*

Para obtener la longitud promedio de las 300 barras seleccionadas se aplica la expresión de la página 76. El denominador corresponde a la suma de las cantidades de barras seleccionadas de cada tipo, es decir, al total de barras consideradas en el estudio, que en este caso es $n = 300$.

El numerador se obtiene como la suma de los productos entre la longitud promedio de cada tipo de barra y la cantidad de barras correspondientes $(\bar{y}_j)$. *De este modo se representa la suma de las longitudes de las 300 barras seleccionadas.*
Luego,

$$\bar{y}_p = \frac{\sum_{j=1}^{3} \bar{y}_j n_j}{\sum_{j=1}^{3} n_j} = \frac{249,994 \cdot 120 + 200,103 \cdot 100 + 212,315 \cdot 80}{120 + 100 + 80} = \frac{66994,78}{300} = 223,316.$$

Es decir que, en promedio, las 300 barras seleccionadas tienen una longitud de 223,316 mm.

4. Media truncada

La **media truncada de orden** p, $\bar{y}_{p\%}$, se define como la media aritmética obtenida luego de ordenar los datos de menor a mayor y de eliminar un $p\%$ de los valores de cada extremo.

Se trata de una medida que se puede calcular para variables cuantitativas medidas en escala de intervalo o de razón.

La fórmula de la media truncada es la misma que la de la media aritmética, luego de la eliminación del $p\%$ de los valores de cada extremo, lo cual implica contar previamente con los datos ordenados. Cabe aclarar que si la cantidad de valores a eliminar no coincide con un número entero, se pueden calcular dos medias truncadas y luego interpolar. Por ejemplo, si se quiere una media truncada 15 % y se cuenta con 10 datos, deberían eliminarse 1,5 datos a cada lado (lo cual es imposible). En ese caso, se puede obtener la media truncada 10 %, eliminando un valor a cada lado y la media truncada 20 %, eliminando dos valores a cada lado y promediar ambas medias para obtener la media truncada 15 %.

La media truncada es una modificación del promedio diseñada para evitar la influencia de valores atípicos y resulta, por lo tanto, una medida menos sensible y más robusta.

Problema 1 (pág. 7): *Si interesa calcular, por ejemplo, una media truncada del 10 % para las longitudes de las barras analizadas, se deben eliminar 12 valores en cada extremo, ya que se cuenta con un total de 120 observaciones. Una vez ordenados los datos, el 10 % truncado implica descartar las 12 observaciones más pequeñas y las 12 más grandes. Por lo tanto, la media truncada se calcula a partir de los valores comprendidos entre las posiciones 13 y 108 del conjunto ordenado (indicadas entre corchetes por su orden en la muestra), dividiendo la suma de estos valores por 96 (correspondiente a* 12024*).*
De la Tabla 2.15, se observa que el primer valor a considerar es $y_{[13]} = 249,62$ *y el último,* $y_{[108]} = 250,36$.

$$\bar{y}_{10\%} = \frac{1}{96}\sum_{i=[13]}^{[108]} y_{[i]} = \frac{1}{96}\cdot(249,62+\cdots+250,36) = \frac{1}{96}\cdot 23999 = 249,99$$

Este número se interpreta de la siguiente manera: eliminando el 10 % de barras con longitudes más bajas y el 10 % de barras con longitudes más altas, la longitud promedio de las barras analizadas es 249,99 mm.

5. Media geométrica

La **media geométrica**, $\bar{y}_G$, se define como la raíz enésima del producto de las n observaciones.

Es una medida que se puede calcular en variables cuantitativas medidas en escala de intervalo o razón, pero solo si la variable toma valores positivos.

La media geométrica se obtiene aplicando directamente la fórmula:

$$\bar{y}_G = \sqrt[n]{\prod_{i=1}^{n} y_i}$$

donde $y_1, y_2, \ldots, y_n$ constituyen el conjunto de n observaciones y n es el tamaño de la muestra.

Se trata de una medida menos sensible a los valores atípicos que la media aritmética y al igual que ella, considera a todos los valores en su cálculo; pero su interpretación es menos intuitiva. Además, si algún valor observado es nulo, el valor de la media geométrica se anula. Si bien su aplicación es menos frecuente que la media aritmética, resulta útil para promediar índices, como los de capacidad de procesos o tasas de crecimiento.

Problema 1 (pág. 7): *Para obtener la media geométrica de las longitudes de las barras analizadas (en mm), se multiplican los 120 valores obtenidos y se aplica luego la raíz de índice 120 a ese producto.*

$\bar{y}_G = \sqrt[120]{249,29\cdot 249,33\cdot 249,34\cdot\cdots\cdot 250,67\cdot 250,71} = 249,994.$

Este valor se puede interpretar de la siguiente manera: en promedio, las barras seleccionadas miden 249,994 mm.

6. Moda

Se denomina **moda**, *mo*, al valor de la variable que presenta un máximo relativo en las frecuencias no acumuladas.

La moda [7] es una medida de centrado que se puede obtener con variables cualitativas o cuantitativas, medidas en cualquier escala, y se interpreta como el valor de la variable que se presenta con mayor frecuencia. Para obtener la moda, es necesario identificar el o los valores que más se repiten en el conjunto de los datos.

Salvo excepciones, carece de sentido obtener la moda si se tienen datos sin agrupar, especialmente si se estudia una variable cuantitativa continua.

Si los datos están organizados en una tabla de distribución de frecuencias y las clases están formadas por un único valor, la moda es el valor y_h de la variable que se asocia a la máxima frecuencia no acumulada. Si las clases están formadas por un intervalo de valores, lo que se observa directamente es un **intervalo modal**, es decir, aquel que presenta la mayor frecuencia no acumulada. En estos casos, la moda no queda determinada de manera única como un valor puntual, ya que depende de la forma en que se hayan definido los intervalos.

La moda es una medida que presenta las siguientes características:

- es la medida de tendencia central utilizada para variables cualitativas;
- no toma en consideración toda la información disponible para su cálculo;
- en un conjunto puede existir una única moda (*distribución unimodal*) o bien más de una (*distribución bimodal* si se presentan dos, o *multimodal* si son más de dos modas);
- también es posible que no exista ninguna si todos los valores de la variable se dan con frecuencias similares.

En la Figura 2.18 se presentan dos distribuciones de frecuencias unimodales y una bimodal, la cual puede darse si en el mismo conjunto de datos se encuentran en realidad dos subconjuntos con diferentes características. Esto último podría ocurrir si, por ejemplo, en el Problema 1 se hubieran considerado las longitudes de barras producidas por el turno mañana y por el turno tarde y las barras de ambos turnos presentaran un comportamiento diferente en relación a la longitud.

[7] No existe un criterio unificado para simbolizar a la moda. En este texto se va a utilizar *mo* en el caso de muestras y Mo en el caso poblacional.

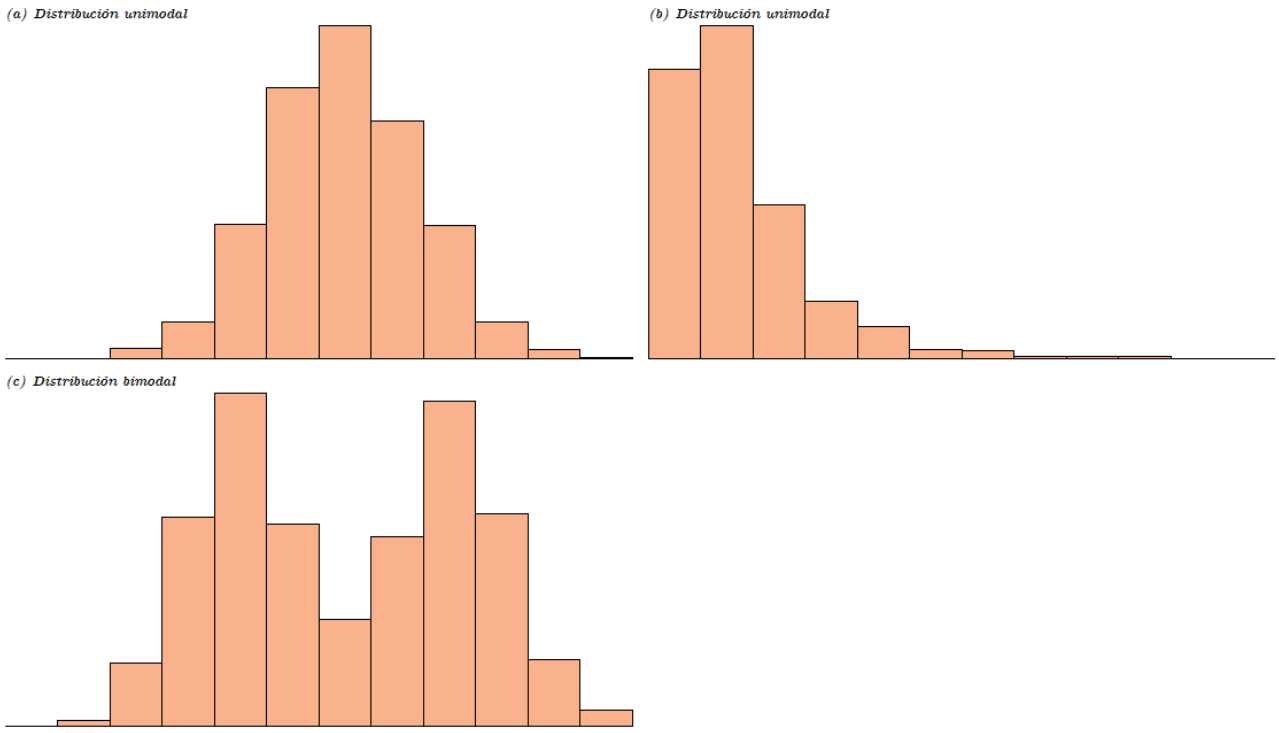


Figura 2.18 *Ejemplos de distribuciones de frecuencias unimodales y bimodales*



Problema 1 (pág. 7): *Como se observa en la Tabla 2.18, que recoge las frecuencias absolutas de la distribución presentada en la Tabla 2.14, en este caso, la máxima frecuencia absoluta es 30 y el intervalo asociado a dicha frecuencia es (250, 250,2], por lo que se podría decir que las barras presentan más frecuentemente longitudes comprendidas en dicho intervalo.*
Debe tenerse en cuenta que esta conclusión depende de la cantidad y amplitud de los intervalos seleccionados al construir la distribución de frecuencias.

Tabla 2.18 *Tabla de distribución de frecuencias absolutas para las longitudes de las barras analizadas*

Longitud (mm) (y_j)	***Nro. de barras*** (n_j)
(249,2 ; 249,4]	*3*
(249,4 ; 249,6]	*8*
(249,6 ; 249,8]	*26*
(249,8 ; 250,0]	*23*
(250,0 ; 250,2]	*30*
(250,2 ; 250,4]	*21*
(250,4 ; 250,6]	*6*
(250,6 ; 250,8]	*3*
Total	***120 (n)***



Con lo visto hasta ahora, se pueden resolver las Actividades 11 a 14, Sección 2.5.

Asimetría de la distribución y medidas de localización

Una distribución es **simétrica** cuando valores (o intervalos de valores) de la variable equidistantes de la media presentan frecuencias similares a ambos lados de la misma. En este caso la distribución de los valores a la izquierda de la media es similar a la distribución de los valores a la derecha de la misma. **En las distribuciones perfectamente simétricas la media y la mediana coinciden y si además, la distribución es unimodal, la moda coincide con ellas. También se cumple que la mediana es equidistante de los dos cuartiles y por lo tanto la diferencia entre la mediana y el tercer cuartil coincide con la diferencia entre la mediana y el primer cuartil**.

Cuando la distribución es **asimétrica a la derecha** o **positiva**, las observaciones están más concentradas a la izquierda de la media y más dispersas a su derecha, haciendo que la cola derecha de la distribución sea más larga. Los valores mayores son menos frecuentes que los menores, pero influyen en el cálculo del **promedio o media aritmética que generalmente resulta mayor que la mediana y también que la moda**.

Cuando la distribución es **asimétrica a la izquierda** o **negativa**, las observaciones están más concentradas a la derecha de la media y más dispersas a su izquierda, haciendo que la cola izquierda de la distribución sea más larga. Los valores menores son menos frecuentes que los mayores, pero influyen en el cálculo del **promedio o media aritmética que generalmente resulta menor que la mediana y también que la moda**.

En una distribución asimétrica a la izquierda o a la derecha, los cuartiles generalmente no resultan equidistantes de la mediana.

La primera información sobre la asimetría de una distribución surge de observar el gráfico de la misma. Es la forma más simple de identificar si un conjunto de datos presenta simetría en relación a la media o no.

En la Figura 2.19 se presenta un ejemplo de cada una de las tres situaciones mencionadas en relación a la simetría o asimetría. Observe que en el caso de la simetría (Figura 2.19.b), cuando se trabaja con muestras, esta se da generalmente en forma aproximada.

Existen además algunas medidas que definen la simetría o asimetría de una distribución y permiten establecer su grado de asimetría, las cuales no se abordan en este texto.

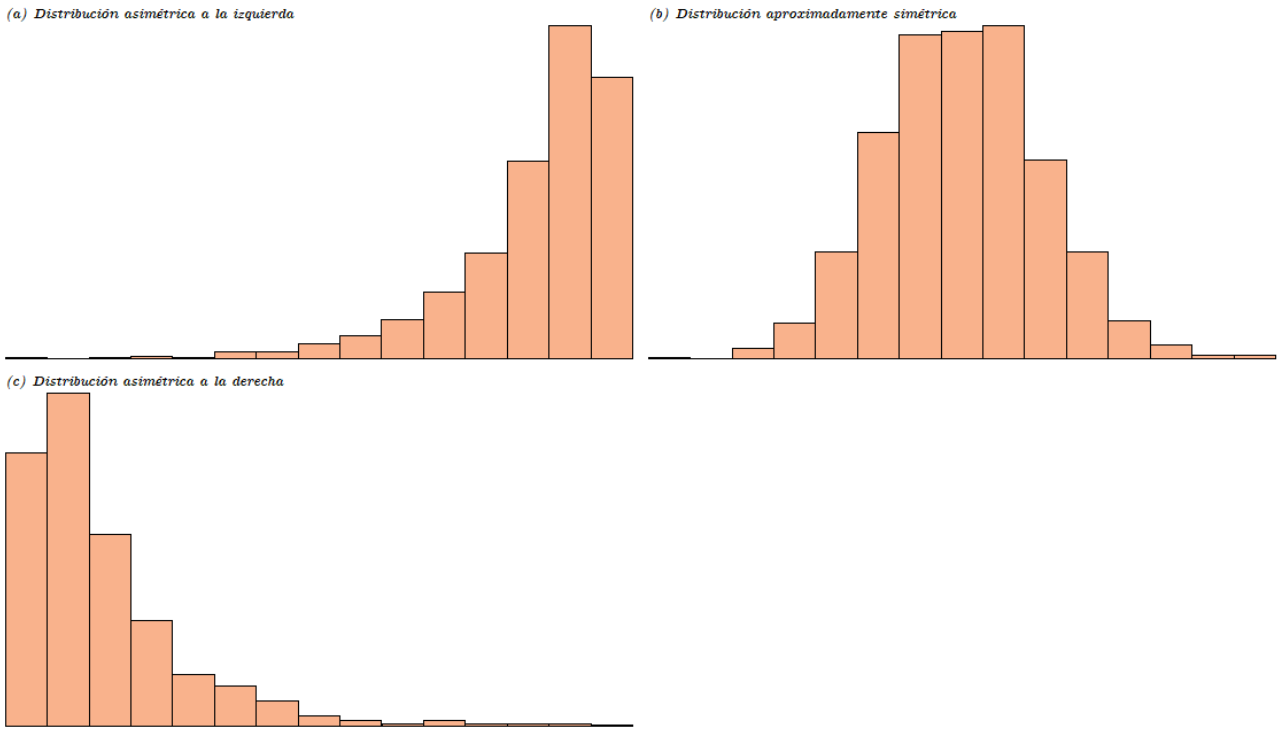


Figura 2.19 *Distribuciones de frecuencias simétricas y asimétricas*



Problema 1 (pág. 7): *De acuerdo con la Figura 2.7, la distribución de las longitudes de las barras es aproximadamente simétrica.*



Problema 5 (pág. 8): *De acuerdo con la Figura 2.9, la distribución de las cantidades de imperfecciones de los rollos es asimétrica a la derecha y unimodal.*



Con lo visto hasta ahora, se puede resolver la Actividad 15, Sección 2.5.

2.3.3 Medidas de dispersión o variabilidad

Las medidas de localización no son suficientes para describir a un conjunto de datos de manera apropiada y deben acompañarse con **medidas de dispersión o variabilidad**. Estas medidas ponen de manifiesto las diferencias entre los distintos valores de un conjunto de datos y son muy importantes para evaluar la calidad de un proceso de producción o de medición: puede decirse que la calidad del mismo está inversamente asociada a la variabilidad.

Dentro de este grupo de medidas se encuentran el rango y el rango intercuartílico, la variancia, el desvío estándar y el coeficiente de variación.

1. Rango

El **rango**, r, es la máxima diferencia que se puede observar en el conjunto de los datos.

Esta medida se puede obtener solo para variables cuantitativas medidas en escala de intervalo o razón. Se define como:

$$r = y_{\text{máx}} - y_{\text{mín}}$$

donde $y_{\text{máx}}$ es el máximo valor observado e $y_{\text{mín}}$ es el mínimo valor observado.

Para obtener el valor del rango, solo hace falta ordenar los datos, observar cuáles son el menor y el mayor valor de ese conjunto y hacer la diferencia entre ellos.

Esta medida proporciona una primera información sobre la dispersión de los datos; pero si los dos valores que intervienen en su cálculo estén anormalmente alejados de los restantes la dispersión del conjunto puede resultar sobrevalorada. Es decir, el rango es una medida que se ve afectada por la presencia de valores atípicos.

2. Rango intercuartílico

El **rango intercuartílico**, *ric*, es la máxima diferencia que se observa en el 50 % central de los datos ordenados.

Esta medida se puede obtener solo para variables cuantitativas medidas en escala de intervalo o razón. Se define como:

$$ric = q_3 - q_1$$

donde q_3 es el tercer cuartil y q_1 es el primero.

Tanto a la izquierda del cuartil 1 como a la derecha del cuartil 3 se encuentra un 25 % de las observaciones, de modo que entre ambos valores se encuentra el 50 % central. En ese conjunto central, ambos cuartiles constituyen los valores mínimo y máximo respectivamente y la diferencia entre ellos es un rango, de ahí el nombre de esta medida. A diferencia del rango definido en primer lugar, el rango intercuartílico no se ve afectado por valores atípicos, es decir, es una medida robusta a la presencia de estos valores.



Problema 1 (pág. 7): *En la sección anterior se obtuvieron los valores de $y_{\text{mín}}$, $y_{\text{máx}}$, q_1 y q_3 para las longitudes de las barras analizadas. Con estos valores se pueden obtener tanto el rango como el rango intercuartílico.*

Rango
El valor mínimo ($y_{mín}$) es 249,29 y el máximo ($y_{máx}$) es 250,71, por lo que el rango resulta r = 250,71 – 249,29 = 1,42. Es decir que en esa muestra de 120 barras, la máxima diferencia que se puede observar en sus longitudes es de 1,42 mm.
Rango intercuartílico
El primer cuartil (q_1) es 249,77 y el tercero (q_3) es 250,205, por lo que el rango intercuartílico resulta ric = 250,205 – 249,77 = 0,435. Es decir que en esa muestra de 120 barras, la máxima diferencia que se puede observar en el 50% central de las longitudes es de 0,435 mm.

3. Variancia y desviación estándar

Estas medidas se pueden obtener solo para variables cuantitativas medidas en escala de intervalo o razón.

Cuando la media aritmética ha sido elegida como la medida de localización del centro de un conjunto de observaciones, las diferencias (también llamadas **desvíos**) entre cada observación y ella ($y_i - \bar{y}$) dan información sobre la variabilidad del conjunto de los datos.

Una primera idea es promediar esos desvíos y utilizar ese valor como medida de variabilidad. El problema es que, dado que la media es el punto de equilibrio de la distribución, ese promedio vale siempre 0, como se muestra en el Ejemplo 2.3.

Ejemplo 2.3 Considerando los datos del Ejemplo 2.2, se calculan los desvíos de cada uno de ellos respecto a la media aritmética o promedio ($\bar{y}$ = 11,6).

Tabla 2.19 *Cálculo de los desvíos respecto a la media aritmética*

y_i	y_i - 11,6
10	-1,6
12	0,4
8	-3,6
15	3,4
13	1,4
Total	**0**

Como se puede observar la suma de los desvíos entre cada valor de la variable y el promedio o media aritmética es cero, y por lo tanto, el promedio de los mismos también.

Surge como alternativa para medir la variabililidad, la opción de elevar los desvíos al cuadrado y luego promediarlos. Esa medida se denomina **variancia** (s^2).

La **variancia**, s^2, es el *promedio* de los desvíos cuadrados entre cada observación y la media aritmética.

Al igual que el promedio, esta medida solo puede obtenerse para variables cuantitativas medidas en escala de intervalo o de razón.

La primera de las expresiones que se presentan a continuación corresponde al cálculo de la variancia cuando los datos no están organizados en una tabla de distribución de frecuencias y la segunda, cuando sí lo están, y en cada clase hay un único valor:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

donde $y_1, y_2, \ldots, y_n$ constituyen el conjunto de n observaciones, $\bar{y}$ es la media aritmética obtenida con dichas observaciones y n es el tamaño de la muestra.

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{k}(y_j - \bar{y})^2 n_j$$

donde $y_1, y_2, \ldots, y_k$ son los k valores distintos que asume la variable en estudio; $n_1, n_2, \ldots, n_k$ son las frecuencias absolutas correspondientes; $\bar{y}$ es la media aritmética obtenida con las n observaciones de la muestra y n es el tamaño del conjunto de los datos.

En el caso en que estén organizados en una tabla y en cada clase haya un intervalo de valores, se reemplaza el valor de y_j por el de y'_j (punto medio del intervalo) y se obtiene un valor aproximado de la variancia (que solo debe utilizarse si no se cuenta con los datos sin agrupar).

Observe que en el cálculo de la variancia de una muestra dada, se realiza una corrección en el denominador y se utiliza $(n-1)$ en vez de n. La cantidad $(n-1)$ se denomina **grados de libertad**.

En el cálculo de la variancia para una población finita, el denominador es N, es decir, el tamaño de dicha población. La expresión de la variancia de una población finita es la siguiente: $\sigma^2 = \frac{\sum_{i=1}^{N}(y_i-\mu)^2}{N}$, donde μ representa al promedio de la población.

En este capítulo, tanto la variancia de una muestra, como la de una población finita, se obtienen como valores numéricos a partir de sus respectivas expresiones de cálculo.

La utilización del denominador $(n-1)$ en el cálculo de la variancia muestral constituye una corrección que será analizada posteriormente en términos de propiedades de los estimadores, en el Capítulo 7.

Dado que en el cálculo de la variancia los valores de la variable están elevados al cuadrado, también lo está la unidad de medida. Para facilitar las interpretaciones y obtener una medida que tenga la misma unidad de medida que la variable en estudio, se utiliza el **desvío estándar** o **desviación estándar** (s).

La **desviación estándar**, s, es la raíz cuadrada positiva de la variancia.

Por lo tanto, resulta:

$$s = \sqrt{s^2}$$

donde s^2 es la variancia de la muestra.

Ejemplo 2.4 A partir del Ejemplo 2.3, se calculan la variancia y la desviación estándar:

Tabla 2.20 *Desvíos respecto a la media aritmética y cuadrados de los desvíos*

y_i	y_i - 11,6	$(y_i - 11,6)^2$
10	-1,6	2,56
12	0,4	0,16
8	-3,6	12,96
15	3,4	11,56
13	1,4	1,96
Total	**0**	**29,2**

Para obtener la variancia se aplica la fórmula de la página 85, para datos sin agrupar. El total de la tercera columna de la tabla anterior da el valor del numerador. El denominador es $(n-1) = 4$, ya que es razonable pensar que esos cinco valores analizados constituyen una muestra.

Una vez calculada la variancia, se obtiene la desviación estándar aplicándole la raíz cuadrada.

$$s^2 = \frac{1}{5-1}\sum_{i=1}^{5}(y_i - 11,6)^2 = \frac{1}{4}\cdot 29,2 = 7,3 \; unidades^2.$$

$$s = \sqrt{s^2} = \sqrt{7,3 \; unidades^2} = 2,7 \; unidades.$$

Si se interpreta la desviación estándar se puede decir que "en promedio", los valores se desvían aproximadamente 2,7 unidades respecto de la media del conjunto.



Problema 1 (pág. 7): *Para obtener la variancia de las longitudes de las 120 barras analizadas se va a utilizar la fórmula para datos sin agrupar (página 85). Luego se va a obtener la desviación estándar aplicando raíz cuadrada. Se tiene lo siguiente:*

$$s^2 = \frac{1}{120-1}\sum_{i=1}^{120}(y_i - 249,994)^2 = \frac{1}{119}\cdot 10,7814 = 0,0906;$$

$$s = \sqrt{s^2} = \sqrt{0,0906} = 0,3.$$

Observando el valor del desvío estándar se puede considerar que en promedio, las barras difieren (o se desvían) aproximadamente en 0,3 mm respecto del promedio del conjunto.

Para las 120 observaciones organizadas en una tabla de distribución de frecuencias, una aproximación de la variancia se obtiene aplicando la fórmula para datos agrupados (página 85), reemplazando el valor de y_j por el de y'_j (punto medio del intervalo). En la Tabla 2.21 se presenta el cálculo de $\sum_{j=1}^{8}(y'_j - 249,985)^2 n_j$, que resulta igual a 11,173.

Luego,

$$s^2 = \frac{1}{120-1}\sum_{j=1}^{8}(y'_j - 249,985)^2 n_j = \frac{1}{119}\cdot 11,173 = 0,094;$$

$$s = \sqrt{s^2} = \sqrt{0,094} = 0,307.$$

Tabla 2.21 *Cálculos requeridos para obtener el valor aproximado de variancia de las longitudes de las barras analizadas*

Longitud (mm) (y_j)	***Punto medio del intervalo*** (y'_j)***(mm)***	***Nro. de barras*** (n_j)	$(y'_j - 249,985)^2 n_j$ (mm^2)
(249,2 ; 249,4]	*249,3*	*3*	*1,408*
(249,4 ; 249,6]	*249,5*	*8*	*1,882*
(249,6 ; 249,8]	*249,7*	*26*	*2,112*
(249,8 ; 250,0]	*249,9*	*23*	*0,166*
(250,0 ; 250,2]	*250,1*	*30*	*0,397*
(250,2 ; 250,4]	*250,3*	*21*	*2,084*
(250,4 ; 250,6]	*250,5*	*6*	*1,591*
(250,6 ; 250,8]	*250,7*	*3*	*1,534*
Total		***120 (n)***	***11,173***

Observaciones: 1- En el cálculo del valor aproximado de la variancia, ee utiliza también el valor aproximado de la media muestral obtenido con la fórmula para datos agrupados.
2- Recuerde que este cálculo se realiza solo si no se cuenta con los datos originales y lo único que se tiene es la tabla de distribución de frecuencias.

4. Coeficiente de variación

El **coeficiente de variación**, *cv*, es el cociente entre la desviación estándar y el promedio, multiplicado por 100 %, y expresa a la desviación estándar como porcentaje del promedio.

Esta medida se puede calcular para variables cuantitativas medidas en escala de intervalo o de razón, de la siguiente manera:

$$cv = \frac{s}{\bar{y}}\,100\,\%$$

donde s es la desviación estándar e $\bar{y}$ es el promedio del conjunto de los datos.



Problema 1 (pág. 7): *Para las longitudes de las barras analizadas, la longitud promedio resultó 249,994 mm y el desvío estándar 0,3 mm. Dado que el coeficiente de variación expresa al desvío estándar como porcentaje del promedio, resulta:*

$$cv = \frac{0,3}{249,994} \cdot 100\,\% = 0,12\,\%$$

Es decir que el desvío estándar representa un 0,12 % del promedio de las longitudes analizadas.

Las medidas de dispersión mencionadas en los puntos anteriores brindan información sobre la variabilidad absoluta; en cambio, el coeficiente de variación es una medida de variabilidad relativa, que expresa al desvío estándar como un porcentaje del promedio y por lo tanto no viene expresado en unidades de la variable. Por ese motivo, es útil para comparar la variabilidad relativa entre distribuciones que no están medidas en la misma unidad o bien que poseen promedios dispares, como se presenta en el Ejemplo 2.5.

Ejemplo 2.5 Suponga que se cuenta con la información de tres muestras. La primera, referida al peso de piezas producidas con acero al carbón, presentó un peso promedio de 150 g y una desviación estándar de 15 g. La segunda, referida al peso de piezas producidas con acero al carbón y luego recubiertas con una capa de zinc, presentó un peso promedio de 130 g y una desviación estándar de 15 g. La tercera, referida al peso de piezas producidas con acero al carbón y luego recubiertas con una capa de estaño, presentó un peso promedio de 200 g y una desviación estándar de 20 g. La información se presenta en la Tabla 2.22.



¿En cuál de las tres muestras se presenta mayor variabilidad?

En términos absolutos, la muestra 3 presenta el mayor desvío estándar; pero dado que las tres muestras presentan diferentes valores promedio, resulta útil analizar la variabilidad relativa a través del coeficiente de variación.

Tabla 2.22 *Medidas de resumen para el peso de piezas producidas con acero al carbón correspondientes a tres muestras diferentes*

Muestra	Peso promedio (g)	Desvío estándar (g)	Coeficiente de variación (%)
1	150	15	10,0
2	130	15	11,5
3	200	20	10,0

Observe que las muestras 1 y 2 presentan la misma desviación estándar en términos absolutos. Sin embargo, la muestra 2 exhibe una mayor variabilidad relativa debido a que su promedio es menor. Por otra parte, las muestras 1 y 3 difieren en el valor absoluto de la desviación estándar, aunque ambas poseen la misma variabilidad relativa.

Considere ahora que, para la tercera muestra, además del peso, se registró la longitud (en mm). Como resultado, se obtuvo una longitud promedio de 350 mm y una desviación estándar de 27 mm

(Tabla 2.23). En términos absolutos, no es posible comparar la variabilidad del peso con la de la longitud, ya que ambas variables se expresan en distintas unidades. No obstante, dicha comparación sí puede realizarse en términos relativos, puesto que el coeficiente de variación es una medida adimensional.

Tabla 2.23 *Medidas de resumen para el peso y la longitud de piezas producidas con acero al carbón correspondientes a la Muestra 3*

Variable	Promedio	Desvío estándar	Coeficiente de variación
Peso	200 g	20 g	10,0 %
Longitud	350 mm	27 mm	7,71 %

Como se observa, las longitudes de la muestra de piezas recubiertas con estaño presentan menor variabilidad, en términos relativos, que los pesos.



Con lo visto hasta ahora, se pueden resolver las Actividades 16 a 18, Sección 2.5.

En el Apéndice, Sección 2.7.1, se presentan todas las medidas obtenidas a partir de los datos del Problema 1 y se presentan las medidas correspondientes a los datos de los Problemas 4 y 5.

2.3.4 Diagrama de caja y bigotes

En su versión más simple, este gráfico constituye una forma resumida de presentar información sobre 5 medidas: $y_{\text{mín}}$, q_1, mediana o q_2 , q_3, e $y_{\text{máx}}$.

Básicamente consta de una caja o rectángulo limitada por los cuartiles 1 y 3, en la que se señala la posición de la mediana con una línea. A ambos lados de la caja se ubican dos líneas o *bigotes* que van del q_1 al $y_{\text{mín}}$ y del q_3 al $y_{\text{máx}}$. La escala donde se localizan estos valores puede ser horizontal o vertical.

De la observación del gráfico se obtienen rápidamente las 5 medidas mencionadas y además se puede calcular aproximadamente el valor del rango (r) (ancho total del gráfico) y del rango intercuartílico (ric) (ancho total de la caja).

Existe otra versión del gráfico que permite además detectar **valores atípicos**. Para ello se requiere el cálculo de dos umbrales:

- Umbral inferior: $q_1 - 1,5\,ric$.
- Umbral superior: $q_3 + 1,5\,ric$.

Un valor puede considerarse atípico si resulta menor que el umbral inferior o mayor que el umbral superior. Del cálculo de los umbrales y del análisis de los datos obtenidos, se puede conocer si en el conjunto de datos se presentan o no valores atípicos y se grafican los bigotes en consecuencia [8]:

- Si no hay valores atípicos, es decir si $y_{\text{mín}}$ es mayor que el umbral inferior e $y_{\text{máx}}$ es menor que el umbral superior, se trazan los bigotes como se indica en la primera versión del diagrama.

- Si hay valores atípicos a la derecha, es decir, mayores que el umbral superior, el bigote se traza hasta el mayor valor entre los que no exceden el umbral. Análogamente, si hay valores atípicos a la izquierda, es decir, menores que el umbral inferior, el bigote se traza hasta el menor valor entre los no inferiores al umbral. En cualquiera de los casos, los valores atípicos se marcan con puntos, asteriscos o cruces (Figura 2.20).

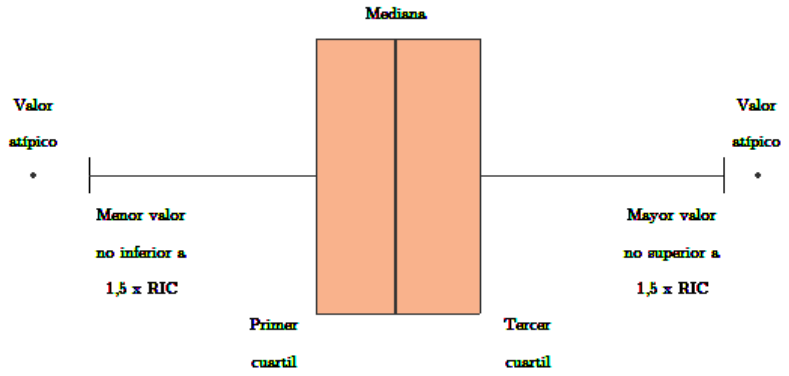


Figura 2.20 *Diagrama de caja y bigotes clásico utilizado para determinar si existen valores atípicos en un conjunto de datos*

Problema 1 (pág. 7): *En la Figura 2.21 se presenta el diagrama de caja y bigotes para las longitudes de las barras, construido a partir de los valores de las 5 medidas de resumen (todas expresadas en mm):*

- $y_{mín} = 249{,}29$;
- $q_1 = 249{,}77$;
- $q_2 = 250{,}005$;
- $q_3 = 250{,}205$ *e*
- $y_{máx} = 250{,}71$.

Los umbrales para analizar si se observan valores atípicos son los siguientes:

- *Umbral inferior:* $249{,}77 - 1{,}5 \cdot (250{,}205 - 249{,}77) = 249{,}1175$.
- *Umbral superior:* $250{,}205 + 1{,}5 \cdot (250{,}205 - 249{,}77) = 250{,}8575$.

En este caso no se observan valores atípicos ya que el valor mínimo es mayor que el umbral inferior y el valor máximo es menor que el umbral superior. Por lo tanto, los bigotes van desde cada uno de los cuartiles hasta los valores mínimo y máximo respectivamente.

[8] Cuando hay valores atípicos, algunos autores trazan el bigote hasta el valor del umbral; pero en este texto no se considera este criterio.

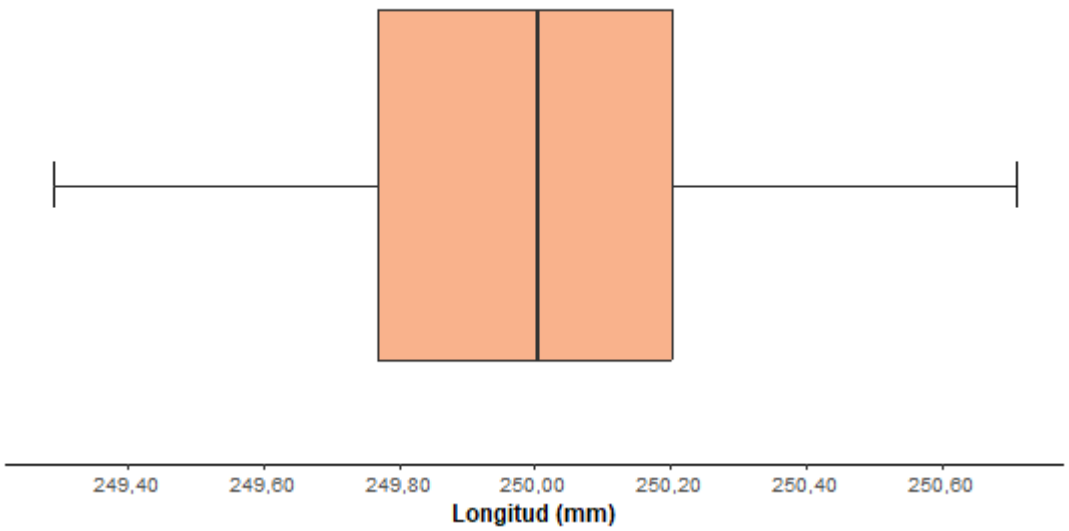


Figura 2.21 *Diagrama de caja y bigotes para las longitudes de las barras analizadas*

Problema 5 (pág. 8): *En la Figura 2.22 se presenta el diagrama de caja y bigotes para las cantidades de imperfecciones por rollo, construido a partir de los valores de las 5 medidas de resumen:*

- $y_{mín} = 0$;
- $q_1 = 0$;
- $q_2 = 1$;
- $q_3 = 1$ *e*
- $y_{máx} = 4$.

Los umbrales para analizar si se observan valores atípicos son los siguientes:

- *Umbral inferior:* $0 - 1{,}5 \cdot (1 - 0) = -1{,}5$.
- *Umbral superior:* $1 + 1{,}5 \cdot (1 - 0) = 2{,}5$.

En este caso, no se observan valores de la variable menores que el umbral inferior pero sí se observan valores mayores al umbral superior: se trata de los valores de la variable superiores a 2,5. Además, no se señala el bigote inferior ya que el valor mínimo coincide con el valor del primer cuartil ($y_{mín} = q_1 = 0$) *ni se distingue la línea que identifica a la mediana ya que el valor de la misma coincide con el valor del tercer cuartil* ($p_{0,50} = q_3 = 1$).

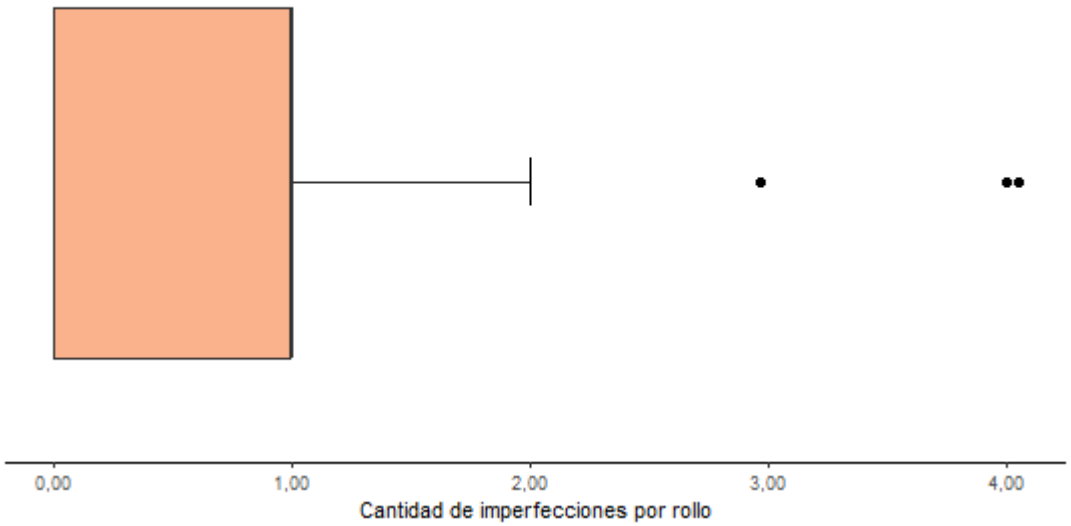


Figura 2.22 *Diagrama de caja y bigotes para las cantidades de imperfecciones de los rollos analizados*

Observación: En este caso, la mediana vale 1 imperfección y coincide con el tercer cuartil.

En la construcción de los diagramas se adoptó el siguiente criterio: ante la presencia de valores atípicos superiores (inferiores), el bigote superior (inferior) se extiende hasta el máximo (mínimo) valor de la variable que no es considerado atípico. En el caso particular del Problema 5, el bigote derecho se extiende hasta 2, que corresponde al mayor valor no atípico.

Relación entre los diagramas de caja y bigotes y los gráficos de distribuciones de frecuencias

Las distribuciones de frecuencias se representan gráficamente mediante histogramas, si se trata de variables continuas o mediante gráficos de bastones, si se trata de variables discretas. También pueden representarse a través de diagramas de puntos o de tallo y hoja. Estos gráficos permiten obtener una visión general y detallada de la distribución de un conjunto de datos; mientras que el diagrama de caja y bigotes ofrece una síntesis más global y simplificada.

En relación con la localización del conjunto de datos, los gráficos de distribuciones de frecuencias permiten obtener una aproximación de las medidas correspondientes, mientras que el diagrama de caja y bigotes muestra directamente las cinco medidas utilizadas para su construcción [9]. En cuanto a la variabilidad, los gráficos de frecuencias permiten visualizar el rango (diferencia entre valores extremos) y el patrón general de dispersión, mientras que el diagrama de caja y bigotes representa tanto el rango como el rango intercuartílico.

Si bien todos estos gráficos permiten identificar la presencia de valores atípicos, en el diagrama de caja y bigotes estos se señalan de forma explícita, lo que facilita su detección. En relación con la asimetría de la distribución, en el histograma puede observarse de manera directa, mientras que en el diagrama de caja y bigotes se infiere de forma indirecta a partir de la diferencia en la longitud de los bigotes y, en algunos casos, de la comparación entre las distancias mediana–primer cuartil y tercer cuartil–mediana.

[9] En algunos casos, puede incluirse la media aritmética en el diagrama de caja y bigotes mediante un símbolo (por ejemplo, un círculo), lo que aporta información adicional sobre esta medida.

En la Figura 2.23 se presentan dos histogramas con sus correspondientes diagramas de caja y bigotes.

En el primer histograma se observa una distribución claramente asimétrica a la derecha. El diagrama de caja y bigotes correspondiente muestra que:

- hay valores atípicos a la derecha;
- el bigote superior es mucho mayor que el inferior;
- la diferencia entre la mediana y el tercer cuartil es mayor que la diferencia entre la mediana y el primer cuartil.

En el segundo histograma, la distribución es aproximadamente simétrica. El diagrama de caja y bigotes correspondiente muestra que:

- hay valores atípicos a la derecha y a la izquierda;
- el bigote superior es similar en tamaño al inferior;
- la diferencia entre la mediana y el tercer cuartil es similar a la diferencia entre la mediana y el primer cuartil.

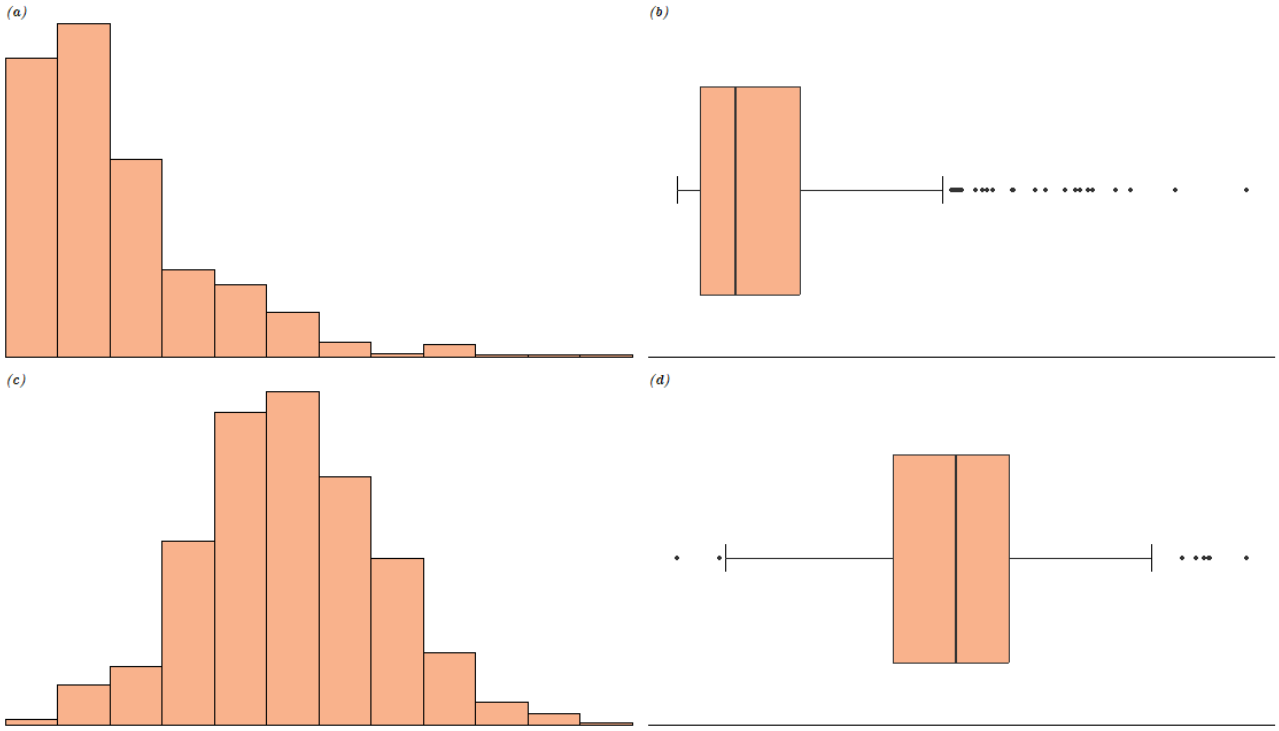


Figura 2.23 *Histogramas (a y c) y diagramas de caja y bigotes correspondientes (b y d), para dos conjuntos de datos*



Problema 1 (pág. 7): *En la Figura 2.24 se observa el histograma y el diagrama de caja y bigotes para las longitudes de las barras analizadas. Ambos gráficos revelan un comportamiento aproximadamente simétrico. En el diagrama de caja y bigotes, ambos bigotes presentan longitudes similares, así como las distancias desde la mediana hasta el primer y el tercer cuartil.*

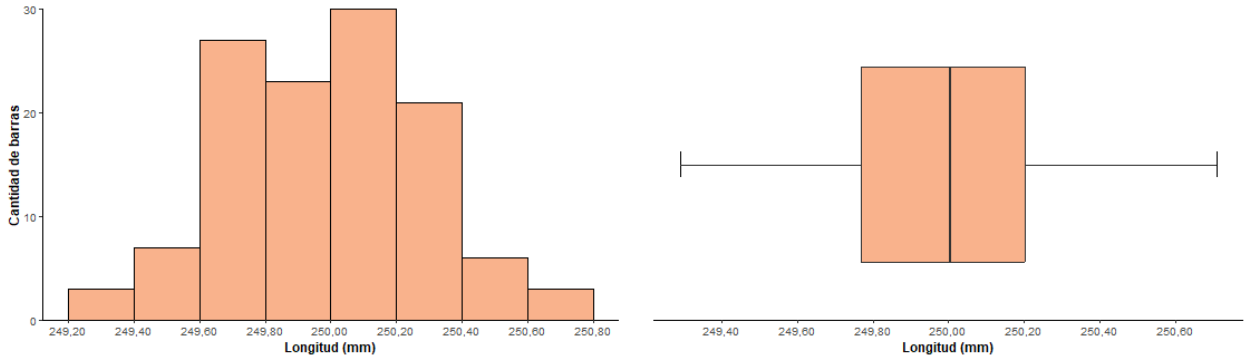


Figura 2.24 *Histograma y diagrama de caja y bigotes para las longitudes de las barras analizadas*

Problema 5 (pág. 8): *En la Figura 2.25 se observa el diagrama de bastones y el diagrama de caja y bigotes para las cantidades de imperfecciones de los rollos analizados. Ambos gráficos revelan que la distribución es asimétrica a la derecha. En el diagrama de caja y bigotes se observa que el bigote superior es mucho mayor que el inferior (que en este caso no existe ya que el valor del primer cuartil y del mínimo coinciden) e incluso se detectan valores atípicos.*

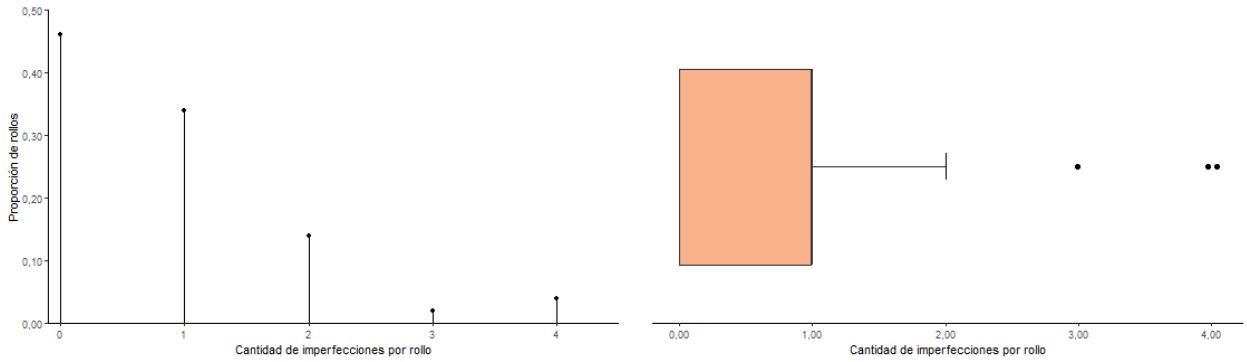


Figura 2.25 *Diagrama de bastones y diagrama de caja y bigotes para las cantidades de imperfecciones de los rollos analizados*

Con lo visto hasta ahora, se pueden resolver las Actividades 19 a 23, Sección 2.5.

2.3.5 Aspectos que involucran a las medidas de localización y de variabilidad

El patrón de variabilidad de una característica de interés en un conjunto de unidades se pone de manifiesto a través de su distribución de frecuencias, ya sea en forma tabular como gráfica. Las medidas de localización y de dispersión también aportan información sobre el comportamiento de esa característica de interés.

¿Qué información brindan conjuntamente las medidas de localización y las de dispersión sobre la distribución de los datos?

Si se obtienen los valores del promedio $(\bar{y})$ y de la desviación estándar (s), se pueden definir intervalos de la forma $\bar{y} \pm k\,s$, donde k es un número positivo que determina la amplitud del intervalo alrededor

del promedio, y analizar la proporción de unidades de la muestra con valores de la variable dentro de dichos intervalos.

Dependiendo de la simetría o asimetría de la distribución, estos intervalos pueden contener un menor o mayor porcentaje de unidades de la muestra.



Problema 1 (pág. 7): *Para las longitudes de las 120 barras analizadas, los valores del promedio y del desvío son 249,994 mm y 0,3 mm respectivamente. Si se definen intervalos de la forma* $(249,994 \pm k\,0,3)$ *mm para valores de* $k = 1,\ 2,\ 3$, *se tiene lo siguiente:*

Tabla 2.24 *Intervalos de la forma* $\bar{y} \pm k\,s$ *y proporción de barras con longitud dentro de cada uno*

k	$(\mathbf{249,994} \pm k\,\mathbf{0,3})$	***Proporción de barras con longitud en el intervalo***
$k = 1$	*(249,694 ; 250,294)*	*0,6583*
$k = 2$	*(249,394 ; 250,594)*	*0,9583*
$k = 3$	*(249,094 ; 250,894)*	*1*

Observación: Las proporciones de barras con longitudes en cada uno de los intervalos se obtuvieron a partir del conjunto de datos ordenados presentado en la Tabla 2.15. También se pudieron haber obtenido del diagrama de tallo y hoja presentado en la Figura 2.10.

En la Tabla 2.24 se muestra que aproximadamente el 66 % de las barras analizadas tienen longitud en el intervalo definido por $\bar{y} \pm s$*; que aproximadamente el 96 % de las barras analizadas tienen longitud en el intervalo definido por* $\bar{y} \pm 2s$*; mientras que todas las barras analizadas tienen longitud en el intervalo definido por* $\bar{y} \pm 3s$.

Análogamente se pueden construir otros intervalos, utilizando otros valores de k, y obtener la proporción de barras con longitudes en dichos intervalos.



Problema 5 (pág. 8): *Para las cantidades de imperfecciones de los 50 rollos de cable estudiados, los valores del promedio y del desvío son 0,84 y 1,017 respectivamente. Si se definen intervalos de la forma* $(0,84 \pm k\,1,017)$ *imperfecciones para valores de* $k = 1, 2, 3 y 4$, *se tiene lo siguiente:*

Tabla 2.25 *Intervalos de la forma* $\bar{y} \pm k\,s$ *y proporción de rollos con cantidad de imperfecciones dentro de cada uno*

k	$(\mathbf{0,84} \pm k\,\mathbf{1,017})$	***Proporción de rollos con cantidad de imperfecciones en el intervalo***
$k = 1$	*(-0,177 ; 1,857)* *[0 ; 2)*	*0,80*
$k = 2$	*(-1,194 ; 2,874)* *[0 ; 3)*	*0,94*
$k = 3$	*(-2,211 ; 3,891)* *[0 ; 4)*	*0,96*
$k = 4$	*(-3,228 ; 4,908)* *[0 ; 5)*	*1*

Observaciones: 1- Dado que el mínimo valor que puede tomar la variable es 0, se deben corregir los extremos del intervalo en caso de que resulten negativos. Además, por tratarse de una variable discreta, se toma como extremo superior al menor número natural que supera al extremo superior obtenido, sin incluirlo.
2- Las proporciones de rollos con cantidad de imperfecciones en cada uno de los intervalos se obtuvieron a partir de la distribución de frecuencias presentada en la Tabla 2.9.

En la Tabla 2.25 se muestra que el 80 % de los rollos analizados tienen una cantidad de imperfecciones en el intervalo definido por $\bar{y} \pm s$; que el 94 % de los rollos analizados tienen una cantidad de imperfecciones en el intervalo definido por $\bar{y} \pm 2s$; que el 96 % de los rollos analizados tienen una cantidad de imperfecciones en el intervalo definido por $\bar{y} \pm 3s$. La totalidad de los rollos analizados tienen una cantidad de imperfecciones en el intervalo definido por $\bar{y} \pm 4s$.

Análogamente se pueden construir otros intervalos, utilizando diferentes valores de k, y obtener la proporción de rollos con cantidad de imperfecciones en dichos intervalos.

Muchas veces la información relativa a alguna característica de interés de un producto o servicio se expresa en términos del valor promedio al que se suma y resta cierta cantidad. Por ejemplo, se informa que un cierto tipo de piezas plásticas tiene un espesor promedio de $(2 \pm 0{,}1)$ cm; o que el volumen de la solución contenida en envases es, en promedio de (100 ± 3) ml. **Generalmente el valor que acompaña al promedio corresponde al desvío estándar**.

Es importante considerar, como se vio en las Tablas 2.24 y 2.25, que esto no implica que el valor de la variable observado en todas las unidades se encuentre dentro del intervalo *(promedio $\pm$ desvío estándar)*. La proporción de valores que se encuentra en intervalos de la forma *(promedio $\pm$ k desvío estándar)* depende de la forma de la distribución.

A partir de lo visto en los ejemplos anteriores se destaca la utilidad de construir intervalos de la forma *promedio* $\pm$ *k desvío* para representar el rango de variación de la distribución de frecuencias completa o de una parte de ella e identificar valores atípicos. Se desprende de lo anterior que si un valor observado de la variable está a más de k desvíos del promedio, con k grande (por ejemplo $k \geq 3$), ese valor se puede considerar un valor atípico.

En los párrafos anteriores se consideró como medida de localización al promedio y como medida de variabilidad al desvío estándar; pero también se pueden relacionar otras medidas de localización y de dispersión.
Si se obtienen los valores de los cuartiles y del rango intercuartílico, estos también se pueden utilizar conjuntamente, por ejemplo, para detectar la presencia de valores atípicos, como se mostró en la Sección 2.3.4.

Si respecto de alguna variable de interés, existen límites de especificación que las unidades de la población deben cumplir, el histograma [10] y la información brindada por las medidas de localización y de dispersión resultan de mucha utilidad para analizar si el proceso es capaz de cumplir con las especificaciones fijadas.

En la Figura 2.26 se muestran diferentes situaciones posibles en relación a la capacidad del proceso para cumplir con lo pretendido.

[10] El histograma es el gráfico para el caso de variables continuas. Si se trata de una variable discreta, corresponde utilizar un gráfico de bastones.

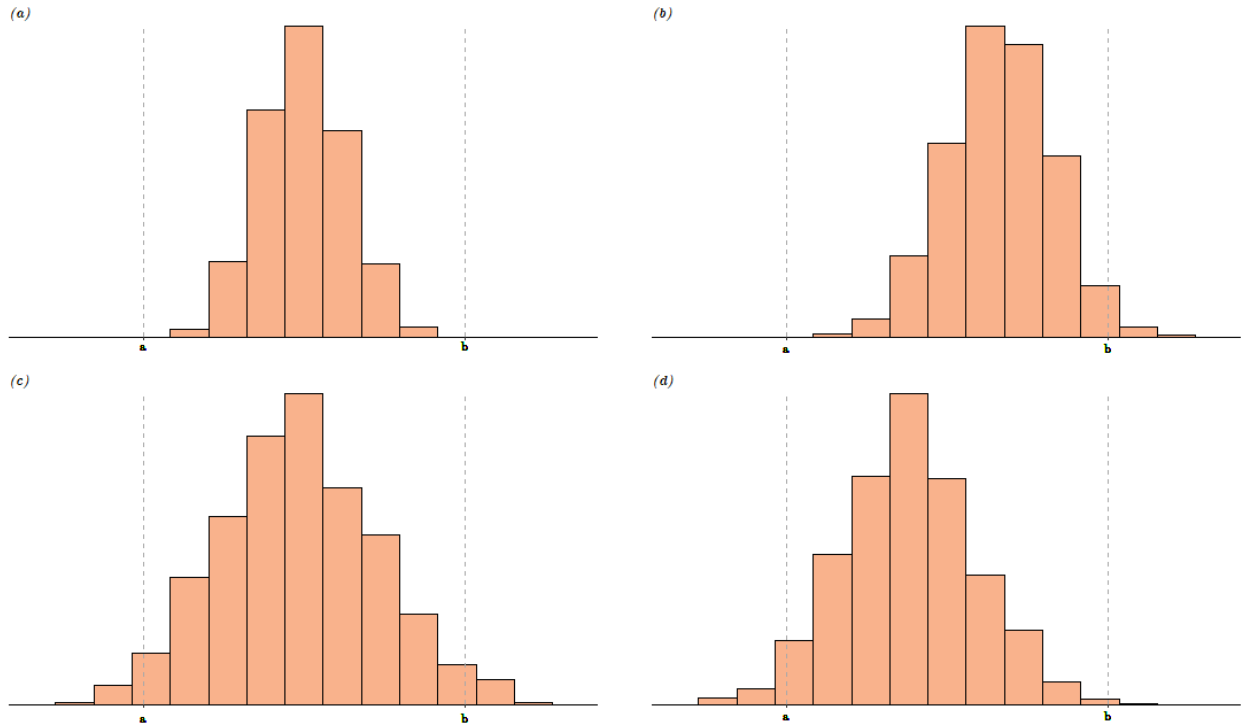


Figura 2.26 *Histogramas correspondientes a cuatro procesos, donde el requerimiento para la variable analizada es que su valor se encuentre entre a y b unidades*

En la Figura 2.26.a, todas las unidades analizadas presentan valores de la variable dentro de los límites de especificación. La distribución de frecuencias está centrada aproximadamente en el punto medio del intervalo definido por las especificaciones y la variabilidad parece adecuada ya que no hay unidades con valores fuera de las especificaciones.

En la Figura 2.26.b, la distribución de frecuencias no está centrada en el punto medio del intervalo definido por las especificaciones, sino en un valor mayor; pero la variabilidad parece adecuada[11]. Hay algunas unidades con valores fuera del límite superior de especificaciones.

En la Figura 2.26.c, la distribución de frecuencias está centrada aproximadamente en el punto medio del intervalo definido por las especificaciones; pero la variabilidad no parece adecuada ya que hay unidades con valores fuera de las especificaciones.

En la Figura 2.26.d, la distribución de frecuencias no está centrada en el punto medio del intervalo definido por las especificaciones y la variabilidad parece mayor que la adecuada[12].

Este análisis, que considera conjuntamente al centrado y a la variabilidad de un conjunto de datos, brinda conclusiones preliminares sobre la capacidad de un proceso para cumplir con las especificaciones, ya que se está considerando información de una muestra.

[11] La variabilidad se puede considerar adecuada ya que, si la distribución de frecuencias se desplazara a la izquierda para quedar centrada en el punto medio, la proporción de unidades fuera de especificaciones sería muy pequeña

[12] La variabilidad puede considerarse no adecuada ya que, aunque la distribución de frecuencias se desplazara a la derecha para quedar centrada en el punto medio, la proporción de unidades fuera de especificaciones no disminuiria

Problema 1 (pág. 7): *Respecto de las longitudes de las barras analizadas, se pretende que estas se encuentren en el intervalo* $(250 \pm 0,6)$ *mm.*
En la Figura 2.27 se muestra el histograma construido con las longitudes de las 120 barras analizadas, en relación con las especificaciones.

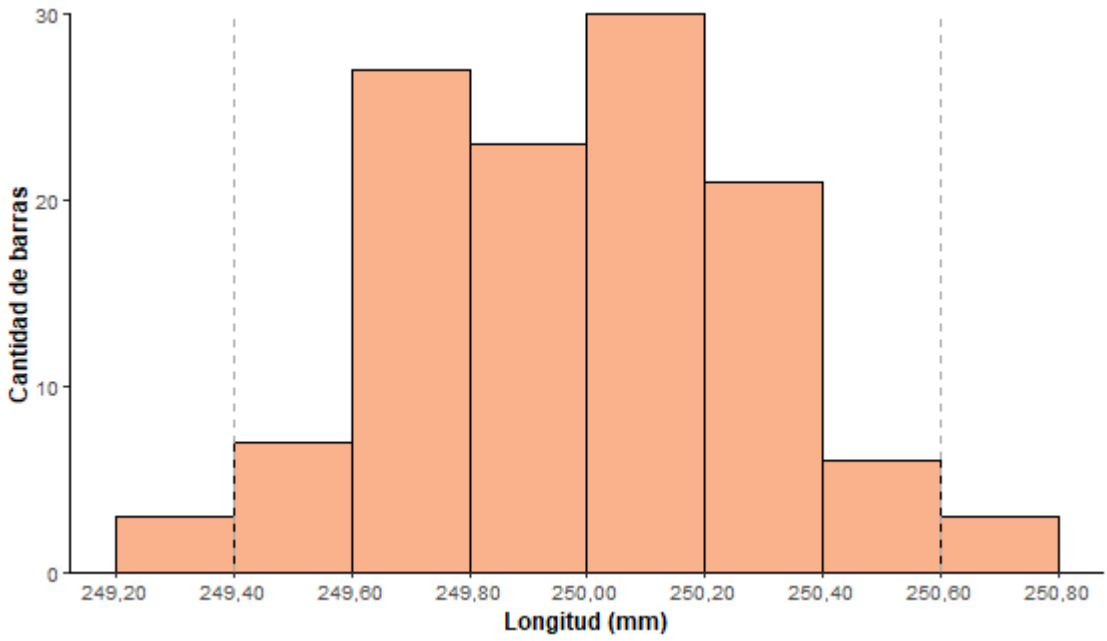


Figura 2.27 *Histograma para las longitudes de las barras analizadas. Las líneas punteadas indican los límites de las especificaciones*

En dicho gráfico se observa que la distribución está centrada aproximadamente en el punto medio del intervalo definido por las especificaciones. Para evaluar si la variabilidad es adecuada, es necesario establecer previamente la proporción máxima de unidades fuera de especificación que se admite. Si la proporción de unidades fuera de especificación es menor que dicho máximo, se concluye que la variabilidad es adecuada.
Como ya se manifestó en párrafos anteriores, este análisis brinda conclusiones preliminares sobre la capacidad del proceso de producción de barras de acero para cumplir con las especificaciones relativas a la longitud de las mismas, ya que se está considerando información de una muestra.

El análisis de la capacidad de un proceso para cumplir con ciertas especificaciones presupone que este se comporta de manera estable, lo cual se evalúa mediante gráficos de control, que no se desarrollan en el presente texto.

En muchas ocasiones, las especificaciones que debe cumplir alguna variable de interés en la población, se definen de la forma $T \pm h$, donde T es el valor nominal o valor objetivo y h es la tolerancia que se admite.

Sería ideal que la media del proceso, μ, coincida con T y que el desvío estándar del proceso, σ sea claramente menor que h.

Estas cuestiones se van a profundizar en el Capítulo 5.

El histograma es otra de las **"7 herramientas básicas de la calidad"** mencionadas al finalizar el Capítulo 1 (página 30).

Este gráfico permite visualizar de manera clara la distribución de los datos y resulta especialmente útil cuando existen especificaciones o valores objetivo. En estos casos, facilita el análisis del centrado del proceso respecto del valor objetivo, así como la evaluación de su variabilidad en relación con los límites establecidos, brindando además una idea preliminar de la capacidad del proceso para cumplirlos.

De forma equivalente, otros gráficos también permiten analizar el comportamiento de la variable en estudio. Entre ellos se destacan el diagrama de bastones, cuando la variable es discreta, y el diagrama de puntos o el diagrama de tallo y hoja, para variables cuantitativas en general.



Con lo visto hasta ahora, se puede resolver las Actividades 24 a 26, Sección 2.5.

2.3.6 Medidas de resumen cuando se aplica una transformación lineal

Dados un conjunto de observaciones $y_1, y_2, \ldots, y_n$, asociadas a una variable cuantitativa y una transformación lineal que se aplica a cada una de ellas, se obtiene un nuevo conjunto de observaciones $x_1, x_2, \ldots, x_n$, asociadas a una nueva variable cuantitativa, donde $x_i = a + b\,y_i$ con a y b pertenecientes a los números reales.

Si la transformación aplicada a cada una de las observaciones originales consiste en la suma de una constante ($x_i = a + y_i$, es decir, $b = 1$), gráficamente se puede observar un desplazamiento de la distribución de frecuencias de las nuevas observaciones, a la derecha o a la izquierda de la distribución original, dependiendo si el valor de a es positivo o negativo. Sin embargo, las diferencias que existen entre las distintas observaciones en la distribución original, se mantienen iguales en la nueva distribución (Figura 2.28.a). En este caso, todas las medidas de localización como, por ejemplo, el promedio, la mediana, o la moda del nuevo conjunto de observaciones son a unidades mayores o menores que las medidas del conjunto original; es decir que el nuevo valor de cada una de estas medidas se obtiene aplicando la misma transformación a las medidas correspondientes al conjunto original. En cambio, las medidas de variabilidad como, por ejemplo, la variancia, la desviación estándar y el rango, no ven modificados sus valores en ambos conjuntos.

Si la transformación aplicada a cada una de las observaciones originales consiste en el producto por un factor ($x_i = b\,y_i$, es decir, $a = 0$), gráficamente se puede observar un desplazamiento de la distribución de frecuencias de las nuevas observaciones, a la derecha o a la izquierda de la distribución original, dependiendo si el valor de b es mayor o menor que 1; pero además, las diferencias que existen entre las distintas observaciones en la distribución original, aumentan o disminuyen en la nueva distribución, también según el valor de b (Figuras 2.28.b y 2.28.c).

En este caso, no solo se modifican las medidas de localización como el promedio, la mediana, o la moda del nuevo conjunto, sino también las medidas de variabilidad como el desvío estándar o el rango. Los valores de las nuevas medidas de localización se obtienen aplicando la misma transformación a la medida correspondiente al conjunto original. Para el caso de las medidas de variabilidad, los valores

de rango, rango intercuartílico y desviación estándar se obtienen multiplicando a las mismas medidas, correspondientes al conjunto original por el valor absoluto de b; mientras que en el caso de la variancia, se multiplica por b^2.

Dado que al sumar una constante a las observaciones originales, el promedio se modifica pero el desvío estándar no, el coeficiente de variación sí se modifica. En cambio, si se multiplica a las observaciones originales por un factor, el valor del nuevo promedio y el del nuevo desvío estándar, quedan también multiplicados por el mismo factor y entonces el coeficiente de variación no se modifica.

En síntesis, conociendo el valor de algunas medidas de resumen para el conjunto original de observaciones, pueden obtenerse los valores de dichas medidas para el conjunto de observaciones transformadas, como se describe en la Tabla 2.26.

Observaciones:

- Lo que se expresa en la Tabla 2.26 respecto del promedio o media aritmética, vale también para cualquiera de las medidas de localización, como por ejemplo para la mediana.
- Lo que se expresa en la Tabla 2.26 respecto del desvío estándar, vale también para cualquiera de las otras medidas de variabilidad presentadas en este texto, con excepción de la variancia.

Tabla 2.26 *Promedio, variancia y desvío estándar de las observaciones transformadas expresados en función de las mismas medidas en las observaciones originales*

Para las observaciones originales (y_i)	**Para las observaciones transformadas** ($x_i = a + b\,y_i$)
$\bar{y}$	$\bar{x} = a + b\,\bar{y}$
s_y^2	$s_x^2 = b^2\,s_y^2$
s_y	$s_x = \lvert b\rvert\,s_y$

Ejemplo 2.6 Considere la muestra de n = 5 datos analizada en los Ejemplos 2.2 y 2.3:
$y_1 = 10$; $y_2 = 12$; $y_3 = 8$; $y_4 = 15$; $y_5 = 13$.
Se sabe que $\bar{y}$ = 11,6 unidades y s_y = 2,7 unidades.

1. Suponga que a cada uno de los valores se le suman 5 unidades ($x_i = y_i + 5$), luego:
   $x_1 = 15$; $x_2 = 17$; $x_3 = 13$; $x_4 = 20$; $x_5 = 18$.
   Los valores originales y transformados se presentan en la Figura 2.28.a.
   Resulta: $\bar{x}$ = 11,6 + 5 unidades = 16,6 unidades y $s_x = s_y$ = 2,7 unidades.
2. Suponga que cada uno de los valores se multiplica por 0,5 ($x_i = 0,5\,y_i$), luego:
   $x_1 = 5$; $x_2 = 6$; $x_3 = 4$; $x_4 = 7,5$; $x_5 = 6,5$.
   Los valores originales y transformados se presentan en la figura 2.28.b.
   Resulta: $\bar{x} = 0,5 \cdot 11,6$ unidades = 5,8 unidades y $s_x = 0,5 \cdot 2,7$ = 1,35 unidades.
3. Suponga que cada uno de los valores se multiplica por 3 ($x_i = 3\,y_i$), luego:
   $x_1 = 30$; $x_2 = 36$; $x_3 = 24$; $x_4 = 22,5$; $x_5 = 19,5$.
   Los valores originales y transformados se presentan en la figura 2.28.c.
   Resulta: $\bar{x} = 3 \cdot 11,6$ unidades = 34,8 unidades y $s_x = 3 \cdot 2,7$ = 8,1 unidades.

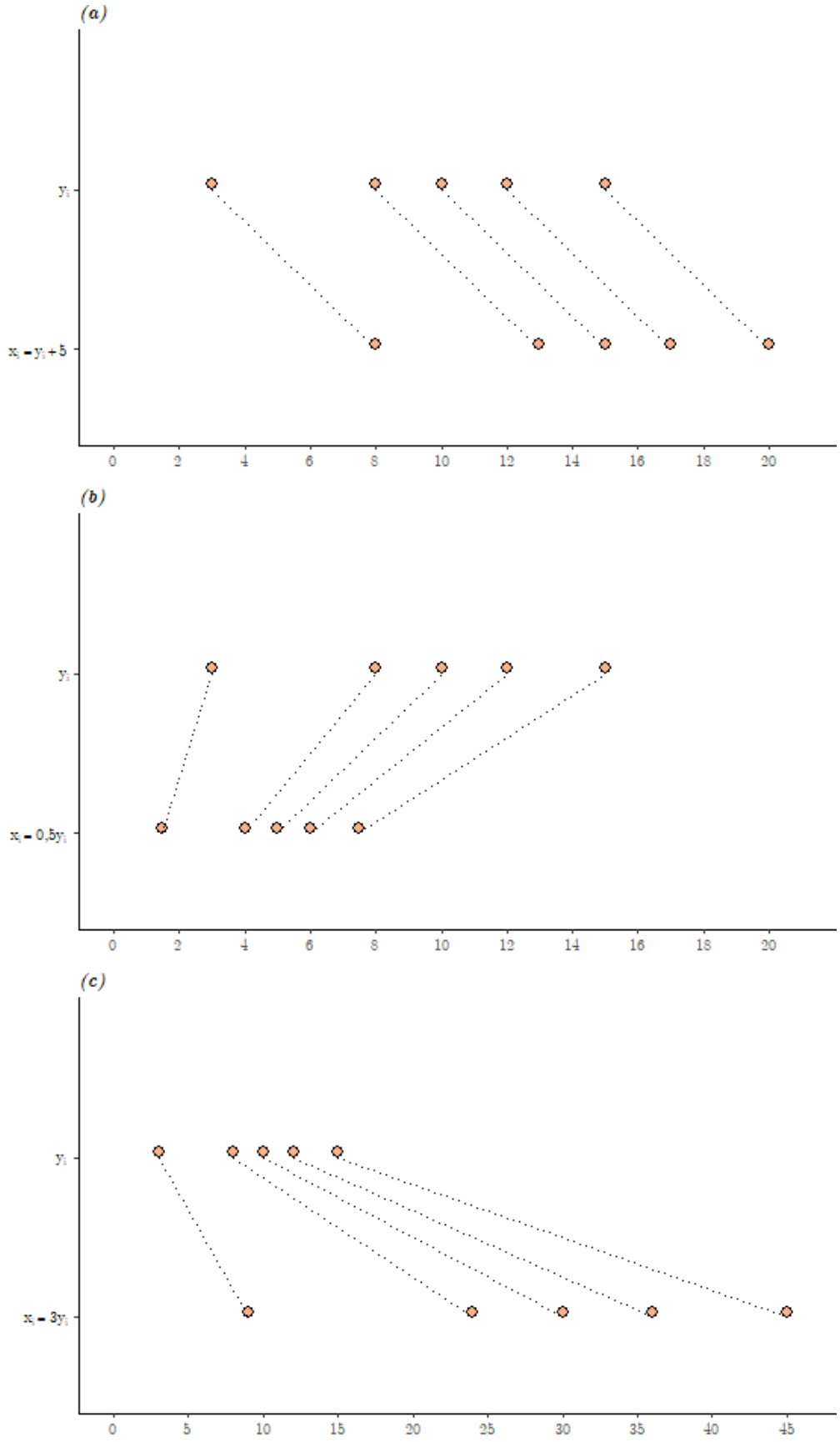


Figura 2.28 *Observaciones originales (y_i) y transformadas (x_i) para las situaciones 1, 2 y 3 respectivamente*

Con lo visto hasta ahora, se pueden resolver las Actividades 27 a 30, Sección 2.5.

2.4 Síntesis

Cuando se cuenta con un conjunto de datos correspondiente a una única variable y se va a llevar a cabo un análisis descriptivo de los mismos, es importante tener en cuenta lo siguiente:

- **¿Cuál es el objetivo del estudio?**

  Considere las siguientes situaciones:

  1. Se quiere determinar, en un estudio a escala piloto, si un tanque construido con cierto material resistirá la presión de un fluido bajo ciertas condiciones. Para ello se lleva adelante un estudio y se analiza la resistencia de las paredes del tanque.
  2. En un laboratorio están evaluando la precisión de un nuevo equipo de medición para cierta magnitud. Realizan 15 mediciones de una pieza patrón, de la cual se conoce el valor de la magnitud.
  3. Las mascarillas utilizadas por los bomberos deben resistir altas temperaturas. En la estación de bomberos están considerando comprar mascarillas a un nuevo fabricante. En un ensayo (en laboratorio) una muestra de 40 mascarillas es sometida a temperaturas superiores a los 90°C y se observa en cada una si los lentes se desprenden o no.

  ¿Considera que el valor promedio es una medida de interés en alguna de las situaciones?

  En la primera no sería razonable ya que si se trata de evaluar la resistencia de las paredes del tanque, conviene tener información sobre la máxima presión que estas resistirán y no sobre la presión promedio (que será superada posiblemente la mitad de las veces).
  En la segunda tampoco es razonable, ya que la precisión de un método de medición no se ve reflejada en un valor promedio sino en alguna medida de variabilidad (entre una medición y el valor de la magnitud).
  En la tercera, una medida de interés es la proporción de mascarillas que resisten las altas temperaturas sin dañarse, o bien, la menor temperatura a partir de la cual las mascarillas se dañan.

  Estas situaciones muestran que es muy importante tener claro el objetivo del estudio y ver cuáles son las medidas de interés en relación a él. El promedio, una medida ampliamente conocida y aplicada, no siempre es la que debe obtenerse.

- **¿De qué tipo es la variable?, ¿cuál es su nivel de medición?**

  En la Tabla 2.27 se resumen las tablas y gráficos que pueden construirse, así como las medidas que pueden obtenerse según el tipo de variable que se trate. En la Tabla 2.28 se listan las medidas junto con la simbología que se usa en este libro.

- **¿Cómo están tomados los datos en relación al tiempo?**

  En relación al tiempo, pueden considerarse las siguientes situaciones:

  1. Los datos están tomados en un mismo momento en el tiempo.
  2. Los datos están tomados a través del tiempo, pero manifiestan un comportamiento estable.
  3. Los datos están tomados a través del tiempo, pero no manifiestan un comportamiento estable. Se presentan ciclos, tendencias, cambios de nivel, etc.

  Si se presenta la Situación 2, los datos pueden analizarse como si fueran tomados en un mismo momento en el tiempo (es decir, como para la Situación 1) y pueden construirse tablas y gráficos

Tabla 2.27 *Representación tabular y gráfica y medidas de resumen según el tipo de variable y el nivel de medición*

<table>
<tr><th>Tipo de variable (nivel de medición)</th><th>Representación Tabular</th><th>Representación Gráfica</th><th>Medidas de resumen</th></tr>
<tr><td>Cualitativa (nominal)</td><td>Tabla de distribución de frecuencias absolutas y relativas (n_j, f_j)</td><td>Gráfico de sectores
Gráfico de barras
Diagrama de Pareto</td><td>Proporciones
Moda</td></tr>
<tr><td>Cualitativa (Ordinal)</td><td rowspan="3">Tabla de distribución de frecuencias absolutas, relativas y acumuladas (n_j, f_j, N_j, F_j)</td><td>Gráfico de sectores
Gráfico de barras</td><td>Proporciones
Moda
Mediana
Otros percentiles</td></tr>
<tr><td>Cuantitativa discreta (de razón o de intervalo)</td><td>Gráfico de bastones
Gráfico escalonado
Diagrama de tallo y hoja
Diagrama de puntos
Diagrama de caja y bigotes</td><td rowspan="2">Proporciones
Medidas de posición o localización:
Moda
Mediana
Otros percentiles
Promedio (media aritmética)
Media geométrica
Media truncada
Medidas de variabilidad o dispersión:
Rango
Rango intercuartílico
Variancia
Desviación estándar
Coef. de variación</td></tr>
<tr><td>Cuantitativa continua (de razón o de intervalo)</td><td>Histograma
Polígono de frecuencias
Polígono acumulativo
Diagrama de tallo y hoja
Diagrama de puntos
Diagrama de caja y bigotes</td></tr>
</table>

Observación: en esta tabla solo se consideran estudios transversales.

Tabla 2.28 *Simbología utilizada en este texto para las medidas de resumen, según correspondan a una muestra dada o a una población*

Clasificación	Medida	Muestra	Población
Medidas de posición o localización	Percentil de orden α	p_α	P_α
	Promedio	$\bar{y}$	μ
	Moda	mo	Mo
	Mediana	$\tilde{y} = q_2 = p_{0,50}$	$\tilde{\mu} = Q_2 = P_{0,50}$
	Cuartiles	$q_1 = p_{0,25}$, $q_3 = p_{0,75}$	$Q_1 = P_{0,25}$, $Q_3 = P_{0,75}$
	Media geométrica	$\bar{y}_G$	μ_G
	Media truncada, p%	$\bar{y}_{p\%}$	$\mu_{p\%}$
Medidas de variabilidad o dispersión	Rango	r	R
	Rango intercuartílico	ric	RIC
	Variancia	s^2	σ^2
	Desviación estándar	s	σ
	Coeficiente de Variación	cv	CV
Otra	Proporción	f_0	π

Observación: No existe una notación ampliamente aceptada para la moda. En este texto se propone usar *mo y Mo*.

y calcularse medidas como las que se presentan en el diagrama anterior.
Si se presenta la Situación 3, carece de sentido obtener la mayoría de las medidas presentadas en este libro.

- **¿Cómo es la forma de la distribución? ¿Se presenta asimetría? ¿Se observan valores atípicos?**

  Apenas se obtienen los datos y se depuran los posibles errores de carga, de medición y otros que se puedan detectar, es importante construir gráficos para ver cómo es la forma de la distribución y si hay o no valores atípicos.
  Además, algunas de las medidas presentadas en este capítulo son sensibles a datos atípicos y sus valores se ven modificados por ellos; mientras que otras no lo son. Por ese motivo el cálculo de diferentes medidas y la comparación de sus valores también da información sobre la presencia de valores atípicos y sobre la asimetría de la distribución.

- **¿Qué ocurre cuando se analizan los datos de una muestra aleatoria de una determinada población?**

  Si se lleva a cabo un estudio observacional por muestreo o un estudio experimental, se cuenta con un conjunto de datos y no con toda la población. En ese caso, del análisis descriptivo se puede tener una idea del comportamiento de la variable en estudio y obtener los valores de los estadísticos de interés. Pero solo con eso no alcanza, ya que deben generalizarse las conclusiones a toda la población de interés aplicando herramientas inferenciales (las cuales se presentan en los Capítulos 8 y 9).
  Estas herramientas se basan en diversos requerimientos relativos al conjunto de los datos (por ejemplo, que estos constituyan una muestra aleatoria, que la distribución de la variable en la población presente ciertas características, etc.). El análisis descriptivo permite, de manera aproximada, evaluar el cumplimiento de dichos requerimientos y elegir la herramienta inferencial apropiada[13].

- **En muchas ocasiones no se toma una muestra, sino que se analizan datos ya disponibles sobre el proceso de interés. ¿Son de utilidad estos datos?**

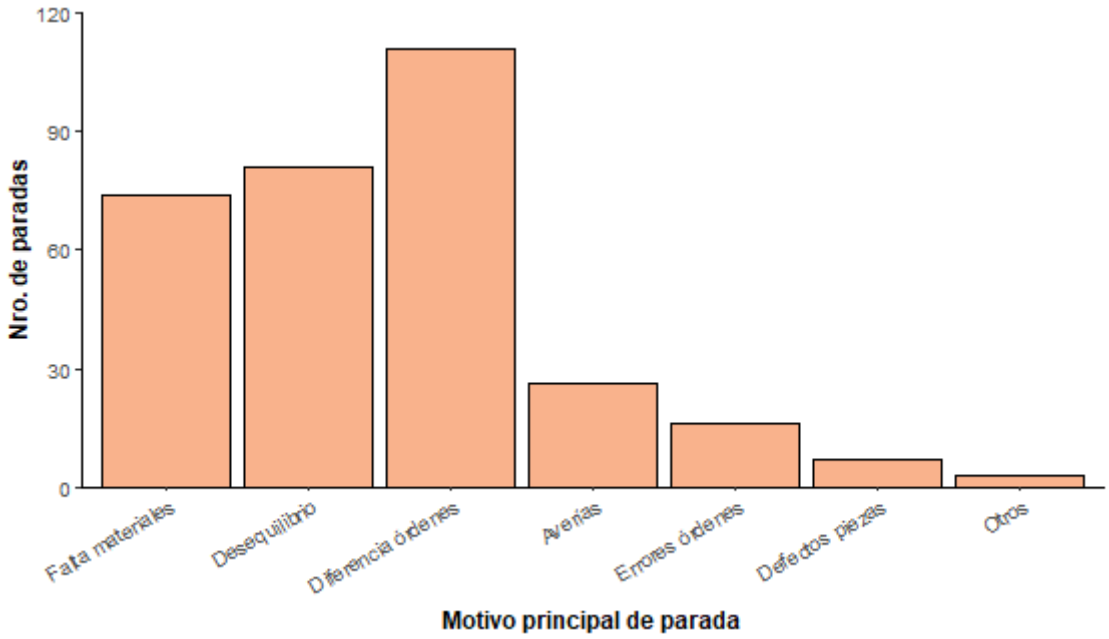
  Aunque no se haya tomado una muestra de una población de interés, analizar estos datos puede ser de mucha utilidad ya que brinda información sobre el problema a resolver e incluso, puede dar una idea de cuál es el problema a resolver. Se suele hablar de "análisis exploratorio de datos", el cual se lleva a cabo sin ideas previas, para conocer qué puede estar pasando.
  Luego de definir adecuadamente el problema, se recomienda tomar una muestra, relevar las variables de interés y generalizar las conclusiones a la población.

[13] Existen algunas herramientas inferenciales que permiten evaluar este cumplimiento, algunas de las cuales se presentan en el Capítulo 9.

2.5 Actividades propuestas

1. Una empresa que fabrica insumos para la industria del mueble (tiradores de cajones, patas, refuerzos, soportes, etc.) está recibiendo últimamente quejas por demoras en las entregas. El ingeniero responsable del proceso sospecha que estas demoras se deben a un incremento de las paradas debido a diferentes causas. Con el objetivo de reducir dichas paradas decide, en primer lugar, registrar el motivo principal de cada una de las ocurridas en los últimos seis meses. Con los datos recabados construye la tabla de distribución de frecuencias y el gráfico de barras que a continuación se presentan:

Motivo principal de la parada	Nro. de paradas	Proporción de paradas
Falta de materiales por problemas logísticos	74	0,2327
Desequilibrio entre la carga de trabajo y la capacidad disponible	81	0,2547
Diferencia entre órdenes de trabajo sucesivas	111	0,3491
Averías de las máquinas	26	0,0818
Errores en las órdenes de fabricación	16	0,0503
Defectos en las piezas a procesar	7	0,0094
Otros	3	0,0220
Total	**318**	**1**



a) Mencione algún objetivo para el cual las paradas analizadas constituyan una muestra. Especifique unidad elemental, población y variable en estudio.

b) ¿Qué puede decir de los motivos de las paradas? Interprete los resultados obtenidos.

c) Construya un diagrama de Pareto con la misma información.

d) Además del motivo, el tiempo en que el proceso está parado en cada caso y el costo que demanda solucionar el problema puede ser diferente. Si contara con esa información, ¿consideraría útil utilizarla en la construcción del diagrama de Pareto? Comente.

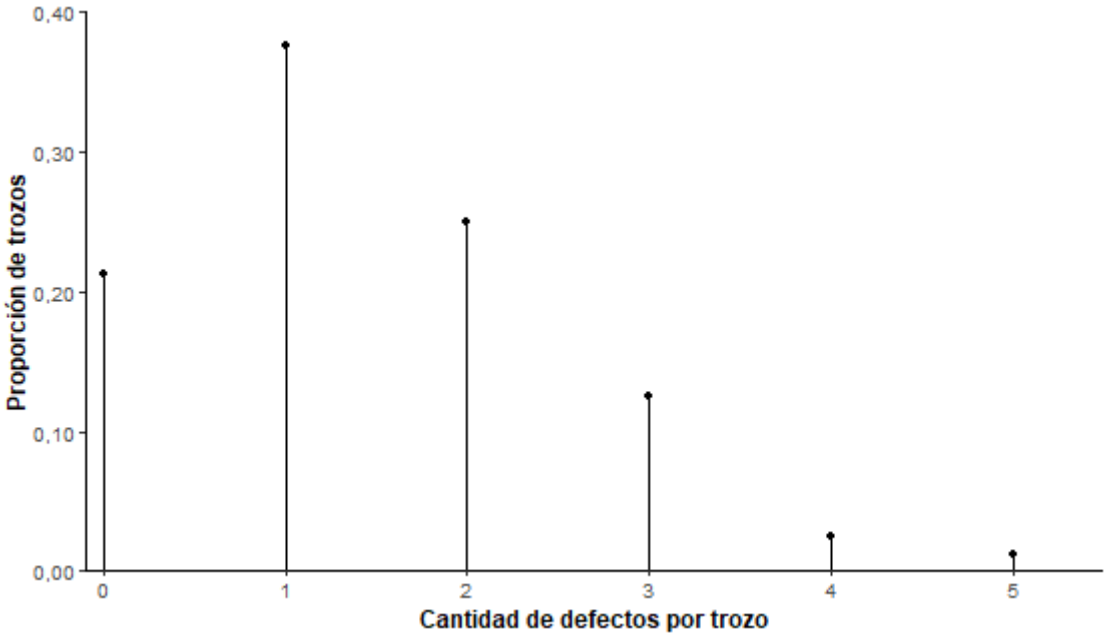
2. La gerenta de operaciones de una planta desea estudiar las fallas que se observan en el proceso de envasado. Los datos sin procesar que se muestran a continuación, corresponden a la falla principal registrada en 50 envases con fallas, los cuales se tomaron de la producción de una semana. Se codificó con 1 al etiquetado incorrecto, con 2 al envase roto, con 3 al envase manchado, con 4 al envase arrugado, con 5 al etiquetado ilegible y con 6 al envase agrietado.

Falla principal observada en cada envase				
3	**2**	**3**	**1**	**5**
6	**1**	**2**	**3**	**3**
4	**3**	**2**	**3**	**2**
2	**5**	**3**	**2**	**4**
2	**1**	**3**	**2**	**2**
3	**2**	**4**	**6**	**2**
4	**2**	**5**	**6**	**3**
4	**6**	**3**	**2**	**5**
2	**5**	**2**	**3**	**2**
3	**4**	**2**	**2**	**3**

a) Plantee un problema de interés para el estudio. Defina la población y la variable correspondiente.
b) Una persona observa los datos y afirma que la variable de interés es cuantitativa discreta. ¿Qué opina Ud.?
c) Realice el análisis descriptivo completo de los datos obtenidos. Interprete los resultados.
d) ¿Sobre qué tipo de falla aconsejaría trabajar con prioridad para corregirla y a futuro evitarla? Redacte un breve informe (un párrafo) respondiendo al problema planteado.

3. En una industria textil interesa evaluar la calidad del estampado de cierto tipo de tela de algodón. Con ese objetivo, de la producción de un día en particular se toma una muestra de 80 trozos de tela (todos de 1 m^2 de superficie) y en cada uno se observa la cantidad de defectos. Con los datos obtenidos se construyen la tabla de distribución de frecuencias y el gráfico de bastones que a continuación se presentan:

Cant. de defectos por trozo	Nro. de trozos	Proporción de trozos	Nro. de trozos acumulados	Proporción de trozos acumulada
0	17	0,2125	17	0,2125
1	30	0,3750	47	0,5875
2	20	0,2500	67	0,8375
3	10	0,1250	77	0,9625
4	2	0,0250	79	0,9875
5	1	0,0125	80	1,0000
Total	**80**	**1**		



a) Especifique unidad elemental, población y variable en estudio.
b) Interprete una fila de la tabla de distribución de frecuencias.
c) Complete las siguientes oraciones y señale las frecuencias mencionadas en el gráfico de bastones cuando corresponda:
   1) El ..... % de los trozos analizados tiene como máximo 1 defecto.
   2) El ..... % de los trozos analizados tiene exactamente 2 defectos.
   3) ..... trozos analizados tienen al menos 4 defectos.
   4) El 98,7 % de los trozos analizados tiene como máximo ..... defectos.

4. De una máquina que fabrica piezas especiales, se registra la cantidad de piezas defectuosas producidas por día. Para un mes en particular (30 días), se conoce lo siguiente:
   - Solo 1 día de los que estuvo en funcionamiento produjo 4 piezas defectuosas y ningún otro día superó ese valor.
   - En 19 días no produjo piezas defectuosas.
   - El 10 % de los días produjo 2 piezas defectuosas.
   - El 80 % de los días produjo a lo sumo 1 pieza defectuosa.

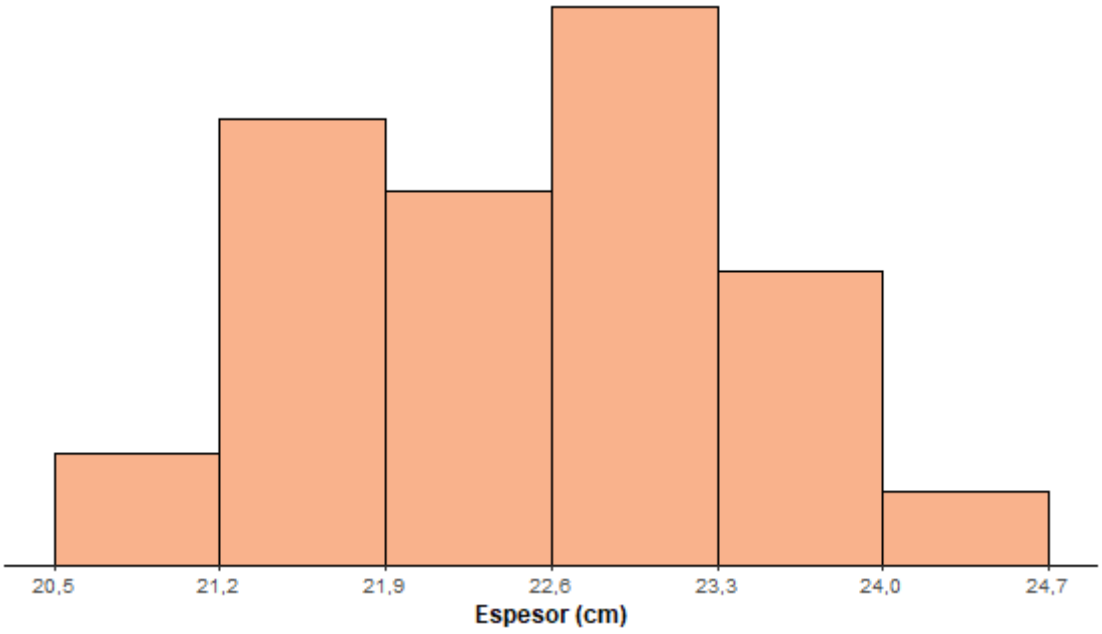
   a) Plantee un objetivo de interés para el estudio, de tal manera que los datos obtenidos constituyan una muestra.
   b) Determine unidad elemental, población y variable en estudio.
   c) Construya una tabla de frecuencias de la distribución del número diario de piezas defectuosas que produjo la máquina en el período considerado.
   d) Construya un gráfico en el que se representen los datos de la tabla.

5. Reconsidere el Problema 6 (Sección 1.4, página 7). En dicho problema se requiere conocer si un tramo de ruta ya construido cumple con las exigencias de calidad. Respecto de la resistencia (medida en megapascales) se pretende que como máximo el 5 % de los puntos del tramo tengan resistencia menor o igual que 30 MPa. A partir de una muestra aleatoria de 50 puntos, se obtiene la siguiente distribución de frecuencias:

Resistencia (MPa)	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
(29,0 - 30,0]	2	0,04	2	0,04
(30,0 - 31,0]	9		11	0,22
(31,0 - 32,0]	16	0,32		0,54
(32,0 - 33,0]		0,34	44	0,88
(33,0 - 34,0]	5	0,10	49	0,98
(34,0 - 35,0]	1	0,02	50	1
Total	50	1		

a) Especifique unidad elemental, población y variable en estudio.
b) Complete la tabla de distribución de frecuencias con los valores faltantes e interprete una fila en términos del problema.
c) ¿Qué puede decir de la resistencia del 54 % de los puntos de menor resistencia?
d) ¿Cuánto vale la resistencia del 12 % de los puntos de mayor resistencia?
e) Realice algún gráfico que muestre la distribución de la resistencia de los puntos.
f) ¿Ud. cree que en ese tramo de ruta se cumple con las exigencias de calidad en cuanto a la resistencia? Indique si estas conclusiones se pueden considerar preliminares o definitivas. Justifique.

6. Reconsidere nuevamente el Problema 6 (Sección 1.4). En el estudio estadístico llevado a cabo para evaluar el tramo de ruta, también se midió el espesor (en centímetros), en cada uno de los 50 puntos seleccionados al azar. Con la información obtenida, se construyó el gráfico que se presenta a continuación:



a) Especifique unidad elemental, población y variable en estudio.
b) Indique cuál es la amplitud de los intervalos y trace el polígono de frecuencias. Complete el gráfico con lo que corresponda.
c) ¿En qué intervalos se observan las mayores frecuencias?
d) ¿Podría indicar aproximadamente la proporción de valores en cada clase? ¿Cómo lo haría?

7. En un taller que se especializa en reparaciones de cierto tipo de equipos electrónicos, están interesados en estudiar el comportamiento de los tiempos de reparación (en minutos). En particular, están considerando aumentar la cantidad de horas de trabajo de los técnicos, si la proporción de reparaciones que demandan más de 180 min supera 0,10. Para ello, seleccionan al azar una muestra de 50 reparaciones y registran el tiempo que demanda cada una. Los datos obtenidos se grafican en el siguiente diagrama de tallo y hoja:

```
14  9
15
15  58
16  000001133
16  5557778888899999
17  0223334
17  55588889999
18  23
18  5
19  1
```

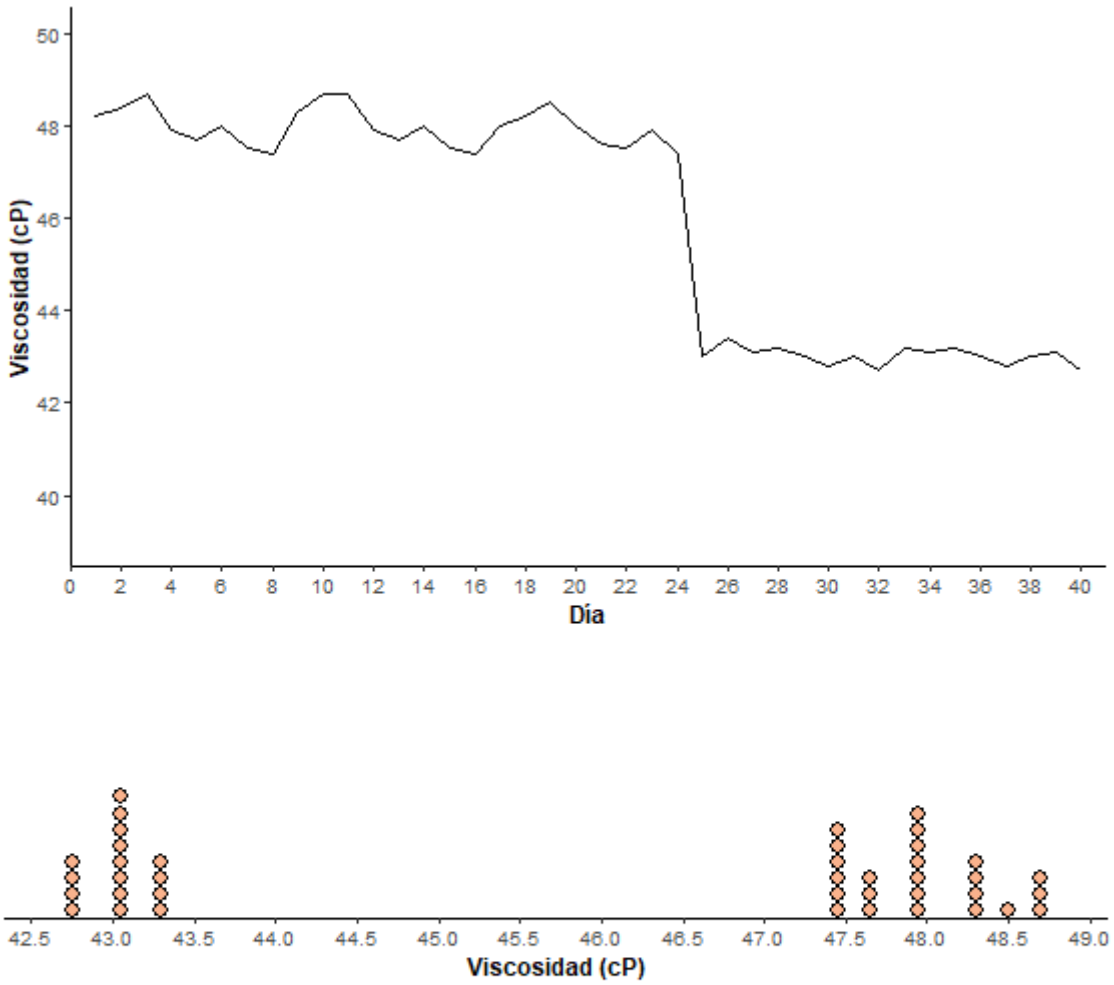
   a) Respecto del diagrama, indique qué se utilizó como "tallo" y cómo se organizaron las "hojas".
   b) ¿Qué puede decir del comportamiento de los tiempos de reparación analizados?
   c) ¿Considera que se debe aumentar la cantidad de horas de trabajo de los técnicos? Explique su conclusión e indique si esta es preliminar o definitiva.
   d) Construya un diagrama de puntos con la misma información.

8. Un fabricante de componentes electrónicos está interesado en conocer el comportamiento del tiempo de vida de cierto tipo de baterías para computadoras que produce. Para ello selecciona una muestra aleatoria de 20 baterías y registra la duración (en horas) de cada una de ellas. Los valores obtenidos son los siguientes:

123 - 121 - 116 - 122 - 131 - 118 - 109 - 180 - 126 - 111
118 - 115 - 125 - 117 - 110 - 112 - 124 - 132 - 124 - 126

   a) Defina variable (clasifíquela), población y población estadística.
   b) Construya un diagrama de tallo y hoja. Comente el criterio para considerar en el "tallo" y en las "hojas".
   c) Comente lo que se observa a partir del gráfico.

9. En una industria química se realiza una medición diaria de la viscosidad de una solución (en centipoises). Para dicha solución la viscosidad debe fluctuar en el intervalo (48 +/- 2) cP. Los valores obtenidos, correspondientes a los últimos 40 días, se presentan a continuación, acompañados por un diagrama de puntos.



a) Identifique: unidad de observación, población, población estadística y variable en estudio.
b) ¿Qué puede observar en ambos gráficos respecto a la especificación requerida?
c) ¿Qué sugieren los datos respecto al comportamiento de esta variable en el tiempo? ¿A qué puede deberse?
d) En este caso, ¿recomendaría presentar la información de la totalidad de los datos recabados? ¿O los analizaría por separado? Comente.

10. Reconsidere la Actividad 4, en la cual se registró el número de piezas defectuosas producidas por día, durante 30 días. ¿Qué utilidad le brindaría construir un diagrama de series cronológicas con dicha información? Explique.

11. Reconsidere la Actividad 8, en la cual se registró el tiempo de vida (en horas), de una muestra de 20 baterías para computadoras. A continuación se presentan algunas medidas de localización obtenidas con dichos datos:

Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
109,0	115,8	121,5	123,0	125,2	180,0

a) Interprete las diferentes medidas de localización obtenidas.
b) ¿Alguna de las medidas se ve afectada por valores atípicos? ¿Cuál/es?
c) Obtenga la media truncada al 5 %.
d) Indique si las medidas obtenidas son estadísticos o parámetros. Justifique.

12. Reconsidere la Actividad 7, en la cual se registró el tiempo de reparación de cierto tipo de equipo electrónico (en minutos), para una muestra aleatoria de 50 reparaciones.
    a) Obtenga el valor de los tres cuartiles e interprételos en contexto.
    b) El tiempo promedio en dicha muestra resultó igual a 169,84 min. Interprete dicho valor en contexto.
    c) Se eliminaron los 5 tiempos más bajos y los 5 tiempos más altos y se obtuvo un tiempo promedio igual a 169,75 min.
        1) ¿Qué medida se calculó? Justifique.
        2) ¿Tenía sentido, en este caso, obtenerla? ¿En qué casos es recomendable?

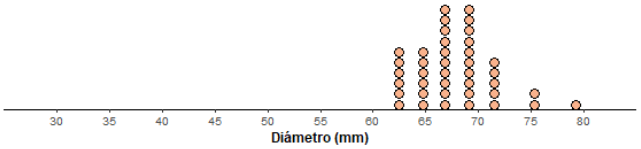
13. Una empresa fabrica piezas con tres máquinas diferentes. De la máquina A se toma una muestra de 30 piezas; de la máquina B se toma una muestra de 28 piezas y de la máquina C se toma una muestra de 20 piezas. Con la información de cada muestra, se obtiene que el peso promedio de las piezas (en gramos) es 58 para A; 61 para B y 68 para C.
    a) Identifique la variable en estudio.
    b) Calcule el peso promedio para la muestra total de 78 piezas.
    c) Si tuviera la mediana y la moda de los pesos para cada una de las tres muestras, ¿podría calcular a partir de ellas las correspondientes a la muestra total?

14. Reconsidere la Actividad 1, relativa a las paradas de un proceso de producción de insumos para fábricas de muebles y la Actividad 2, relativa a las fallas de un proceso de envasado. Para cada una, obtenga e interprete dos medidas de resumen, en contexto.

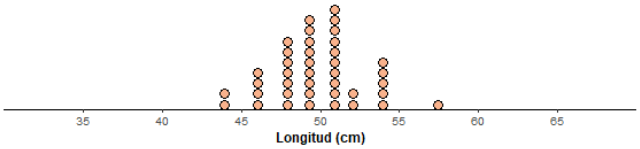
15. Revise las actividades resueltas hasta el momento.
    a) Identifique una de ellas en la cual la distribución de frecuencias resulte asimétrica. Indique si dicha asimetría es a la derecha o a la izquierda.
    b) Para la distribución mencionada en el ítem anterior, observe/obtenga los valores de la mediana, la moda y el promedio y comente sobre las diferencias entre ellos.
    c) Identifique alguna actividad en la que la distribución sea aproximadamente simétrica y realice lo mismo que se pide en el ítem anterior.

16. Una empresa utiliza dos máquinas para producir piezas metálicas. Los diagramas de puntos que se presentan en las siguientes figuras corresponden al diámetro (en mm) y a la longitud (en mm) de piezas metálicas de dos muestras, una formada por piezas tomadas de la máquina 1 y otra formada por piezas tomadas de la máquina 2. Determine si las siguientes afirmaciones son verdaderas o falsas.

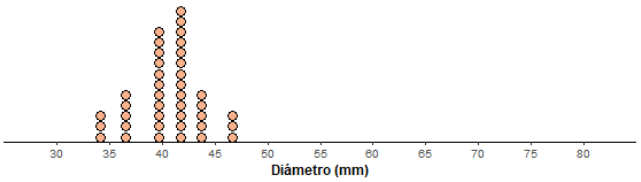
(a) Diámetro de las piezas de la Máquina 1



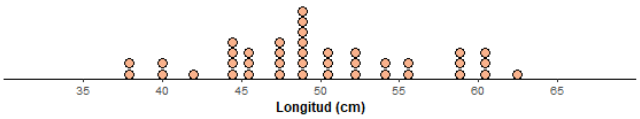
(c) Longitud de las piezas de la Máquina 1



(b) Diámetro de las piezas de la Máquina 2



(d) Longitud de las piezas de la Máquina 2



a) Los diámetros de las piezas provenientes de la máquina 1 y de la máquina 2 difieren principalmente en cuanto a su variabilidad.

b) La longitud promedio de las piezas de ambas máquinas son similares.

c) La longitud de las piezas producidas por la máquina 2 presenta menor variabilidad que la de las piezas producidas por la máquina 1.

d) La medida apropiada para comparar la variabilidad del diámetro con la de la longitud es el desvío estándar de ambas variables.

17. Reconsidere nuevamente la Actividad 8, en la cual se registró el tiempo de vida (en horas), de una muestra de 20 baterías para computadoras y la Actividad 11 en la cual se obtuvieron algunas medidas de localización. Con la información registrada se calculó también la variancia de las duraciones, la cual resultó 223,77 (h^2).

    a) Obtenga rango, rango intercuartílico y desvío estándar. Interprételos en contexto.

    b) En términos relativos, ¿cómo expresaría a la variabilidad?

    c) ¿Alguna de las medidas se ve afectada por valores atípicos? ¿Cuál/cuáles?

18. Reconsidere la Actividad 3, relativa a la evaluación de la calidad del estampado de un cierto tipo de tela de algodón.

    a) Obtenga e interprete dos medidas de localización y dos medidas de dispersión.

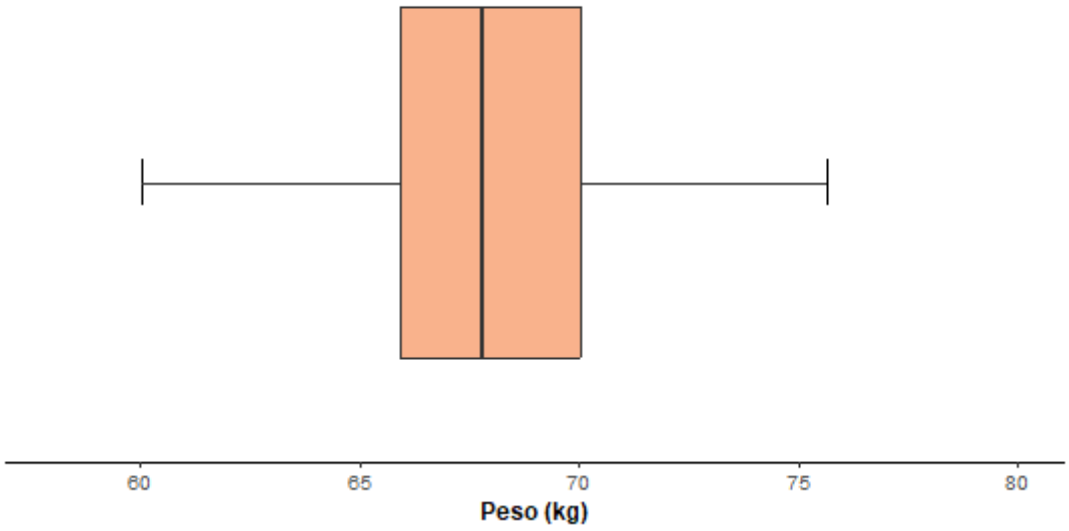
    b) Indique si estas medidas son estadísticos o parámetros.

19. Una empresa produce dos tipos de artículos. Para los artículos tipo A, la característica más importante es el peso (en kilogramos); mientras que para los artículos tipo B,la característica más importante es su longitud (en centímetros). A continuación se presentan salidas para una muestra de 40 artículos tipo A a los cuales se les midió su peso. Para los artículos tipo B también se tomó una muestra de 40 unidades a las que se midió la longitud. Los datos obtenidos se tabularon.

    Medidas de resumen para el peso (en kg) de los artículos tipo A analizados (muestra de n = 40 artículos):

    - Peso promedio $(\bar{y})$: 67,77
    - Desvío estándar (s): 3,64

Diagrama de caja y bigotes para el peso de los artículos tipo A analizados (muestra de n = 40 artículos):



Medidas de resumen para la longitud (en cm) de los artículos tipo B analizados (muestra de n = 40 artículos):

- $\bar{x}$: 98
- s: 7
- x_{min}: 75
- x_{max}: 114
- q_1: 93,23
- q_2: 100,95
- q_3: 104

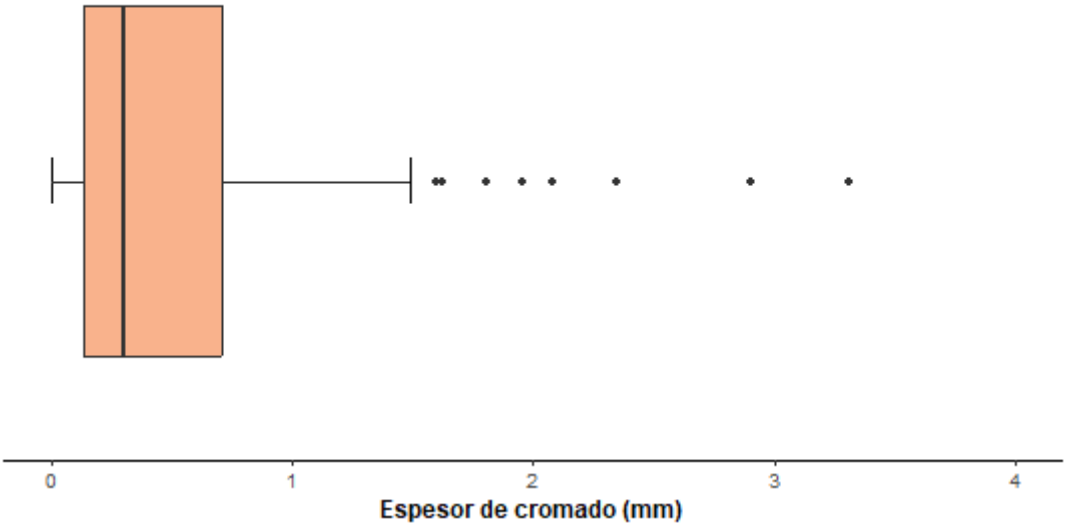
a) Indique los valores aproximados de las medidas que se representan en el diagrama de caja y bigotes construido para el peso de los artículos tipo A. Interprételas.

b) Interprete las medidas de resumen para las longitudes de los artículos tipo B y construya el diagrama de caja y bigotes. Suponga que las 5 menores longitudes observadas fueron: 75 - 76,3 - 77,4 - 78 - 80.

c) ¿Cómo evaluaría la variabilidad para cada conjunto de datos? Obtenga e interprete todas las medidas de variabilidad correspondientes a cada uno.

d) ¿Cuál de los dos conjuntos presenta mayor variabilidad? Justifique.

20. En una empresa que fabrica partes de motores, se seleccionó una muestra aleatoria de 30 ejes de levas y se midió en cada uno de ellos el espesor de la capa de cromado (en milímetros). El diagrama de caja que se presenta a continuación se construyó con la información obtenida. Observe detenidamente el diagrama y señale cuál/es de las opciones corresponden. Justifique.



a) La distribución de los datos de los espesores de cromado presenta:
- asimetría a la derecha.
- asimetría a la izquierda.
- simetría.

b) El espesor promedio es:
- mayor que la mediana.
- menor que la mediana.
- similar a la mediana.

c) Entre las medidas de variabilidad, la que no se modifica ante la presencia de valores atípicos es:
- el rango.
- el rango intercuartílico.
- la desviación estándar.

d) La longitud del bigote derecho es:
- menor o igual a $1,5\,ric$.
- igual a $(y_{\text{máx}} - q_3)$.

e) La longitud del bigote izquierdo es:
- menor o igual a $1,5\,ric$.
- igual que $(q_1 - y_{\text{mín}})$.

f) Indique el valor aproximado de todas las medidas de centrado y de variabilidad que se pueden obtener a partir de la lectura del diagrama de caja y bigotes. Interprete dos de ellas en contexto.

21. Reconsidere la Actividad 5, relativa a la resistencia (en megapascales) medida en una muestra de puntos de un tramo de ruta (Problema 6, Sección 1.4). Obtenga e interprete las medidas necesarias para construir diagrama de caja y bigotes y, luego, constrúyalo.
Utilice los datos disponibles en el archivo https://sites.google.com/view/libro-estadistica.

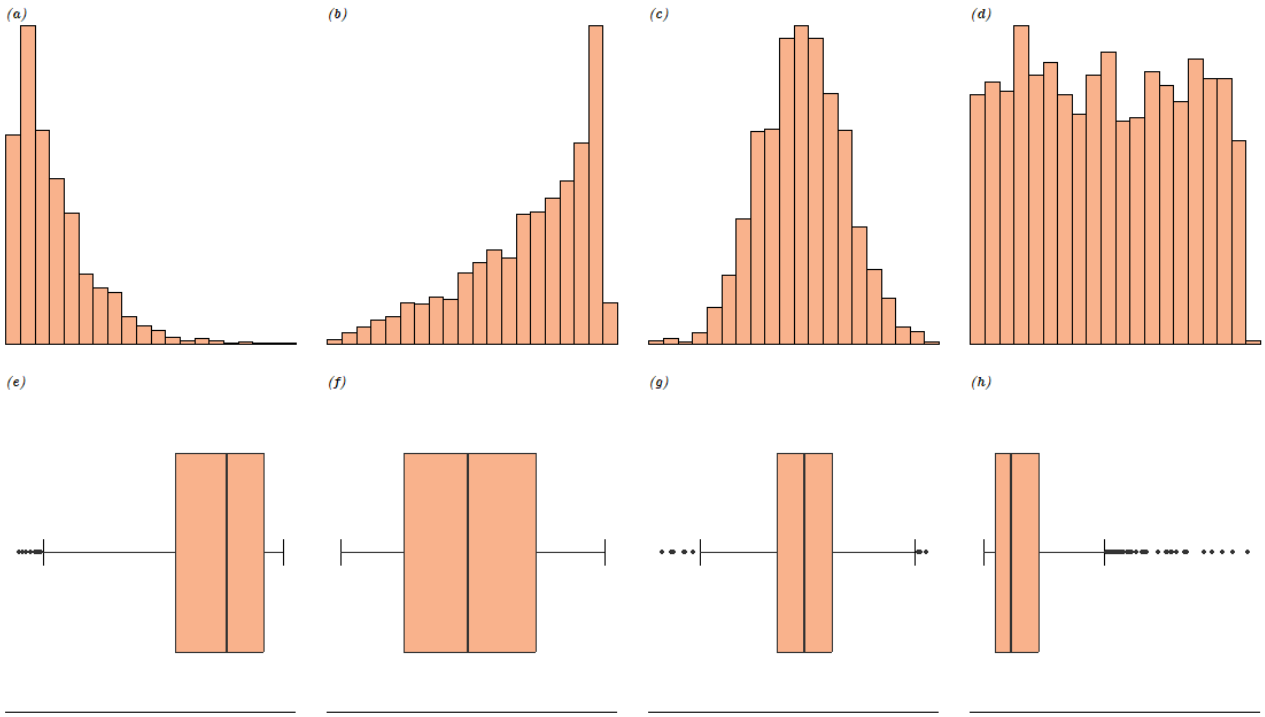
22. Reconsidere la Actividad 4, relativa al número de piezas defectuosas producidas por día durante un mes, en una fábrica.

a) Obtenga e interprete las medidas que se necesitan para construir el diagrama de caja y bigotes y constrúyalo.

b) ¿Qué particularidad presenta este diagrama de caja y bigotes? Comente.

c) ¿Puede obtener alguna/s medida/s de dispersión a partir del diagrama de caja y bigotes? Si es así, indique cuáles son e informe el valor de las mismas.

23. Para los siguientes conjuntos de datos se construyeron el histograma y el diagrama de caja y bigotes correspondiente. A continuación se presentan los gráficos obtenidos:



a) Relacione cada histograma con el diagrama de caja y bigotes que representa el mismo conjunto de datos.

b) Las distribuciones (c) y (d) tienen el mismo rango. Indique cuál de las dos tiene mayor variancia. Justifique.

24. En un proceso de producción de cierto tipo de válvulas, el espesor (en pulgadas) de las paredes del cuerpo de las mismas (en adelante, espesor de la válvula) debe cumplir con las especificaciones 2,1” +/- 0,1”. Si las válvulas presentan espesor superior a 2,2” se deben reprocesar, mientras que aquellas cuyo espesor esté por debajo de 2” se deben desechar.
Con el objeto de conocer el comportamiento del espesor de las válvulas producidas, en la empresa deciden tomar una muestra de 100 unidades. Con los datos se construyó la tabla de distribución de frecuencias que se presenta a continuación.

Espesor (")	Frecuencia Absoluta	Frecuencia Relativa
(1,85 ; 1,90]	1	0,01
(1,90 ; 1,95]	3	0,03
(1,95 ; 2,00]	6	0,06
(2,00 ; 2,05]	16	0,16
(2,05 ; 2,10]	16	0,16
(2,10 ; 2,15]	30	0,3
(2,15 ; 2,20]	20	0,2
(2,20 ; 2,25]	6	0,06
(2,25 ; 2,30]	2	0,02
Total	**100**	**1**

a) Plantee un objetivo de interés para el estudio llevado a cabo.

b) Realice un análisis exhaustivo de los resultados obtenidos.

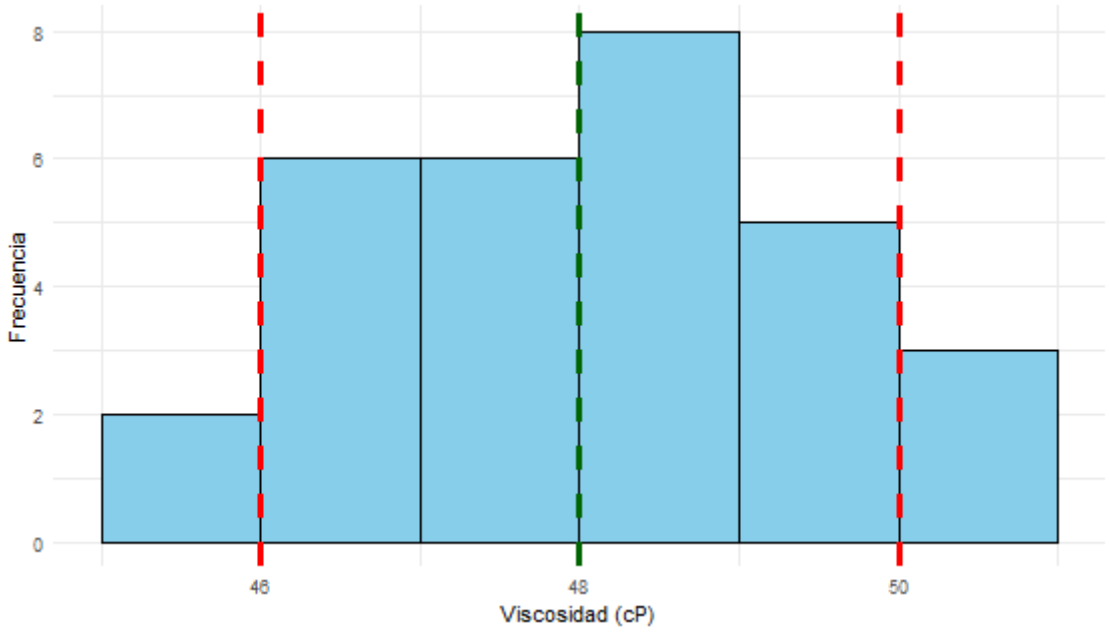
c) Considera que el proceso de producción de válvulas es capaz de cumplir con las especificaciones? Si no es así, ¿habría problemas con el centrado? ¿con la variabilidad? Comente, e indique si necesita información adicional para dar su respuesta.

d) Indique si sus conclusiones son preliminares o definitivas. Justifique.

25. En la Actividad 9 se analizó el comportamiento de la viscosidad de una solución (en centipoises), la cual debe fluctuar en el intervalo (48 ± 2) cP. Suponga ahora que **el proceso finalmente se estabilizó** y que se cuenta con los registros diarios de viscosidad realizados en el último mes. A partir de dichas lecturas se obtuvo el siguiente conjunto de datos, ordenados cronológicamente:

```
visco <- c(
  48.7, 46.4, 49.3, 47.8, 48.1, 50.2, 47.2, 48.5, 46.8, 49.7,
  45.6, 48.9, 47.4, 50.8, 46.1, 49.1, 48.3, 47.6, 46.7, 48.6,
  50.5, 49.8, 47.1, 48.2, 45.9, 49.4, 46.5, 47.9, 48.8, 46.9
)
```

Con la información obtenida se construyó el siguiente gráfico:



a) Defina población, variable y parámetros de interés.

b) Comente, a partir de la observación del gráfico, sobre el cumplimiento de las especificaciones. Indique si necesita información adicional y, en caso afirmativo, mencione qué información sería necesaria para hacer un juicio más preciso.

c) En caso que, según el análisis del gráfico, las especificaciones no se cumplan, explique si esto se debe a un problema de centrado, de variabilidad, o de ambos. Justifique su respuesta a partir de la distribución de los datos y su relación con los límites de especificación.

d) Comente si sus conclusiones se pueden considerar preliminares o definitivas. Justifique.

e) En el enunciado se menciona que el proceso se encuentra estabilizado. ¿Dispone de suficiente información para evaluar esta afirmación, es decir, para evaluar la estabilidad del proceso durante el mes considerado? Justifique su respuesta y si lo cree conveniente realice el análisis que considere más apropiado.

26. Una empresa metalmecánica produce anillos de pistón para motores de combustión interna. Una de las características críticas de calidad es el diámetro interior (en milímetros) de cada pieza. La especificación de diseño establece un valor nominal de 74 mm con una tolerancia de ± 0,08 mm; es decir, se consideran conformes las piezas cuyo diámetro interior se encuentra en el intervalo $(73,92; 74,08)$ mm. Las piezas que no cumplen con este requisito son rechazadas por los clientes.
Se relevaron muestras de 80 piezas en cuatro situaciones distintas del proceso de producción. Los datos obtenidos, ordenados cronológicamente, son los siguientes:

```
datosI <- c(
  73.976, 74.022, 74.006, 73.967, 73.987, 74.036, 73.947, 73.991, 74.028,
  73.981, 73.985, 73.998, 74.033, 73.986, 73.990, 73.990, 74.049, 74.048,
  74.022, 74.008, 74.016, 74.033, 73.979, 74.026, 73.972, 73.986, 74.020,
  73.969, 73.997, 73.981, 73.994, 73.938, 73.961, 73.985, 74.020, 73.996,
  74.000, 74.015, 73.981, 74.006, 73.982, 73.962, 73.991, 74.013, 74.007,
  74.000, 74.053, 74.009, 74.022, 74.049, 73.972, 73.977, 74.038, 73.982,
  74.001, 74.024, 74.020, 74.039, 74.033, 74.024, 73.983, 74.017, 74.007,
  73.971, 74.031, 74.018, 74.001, 73.995, 73.974, 74.004, 74.010, 73.982,
  74.026, 73.976, 73.953, 74.023, 73.991, 73.997, 73.982, 73.965
)

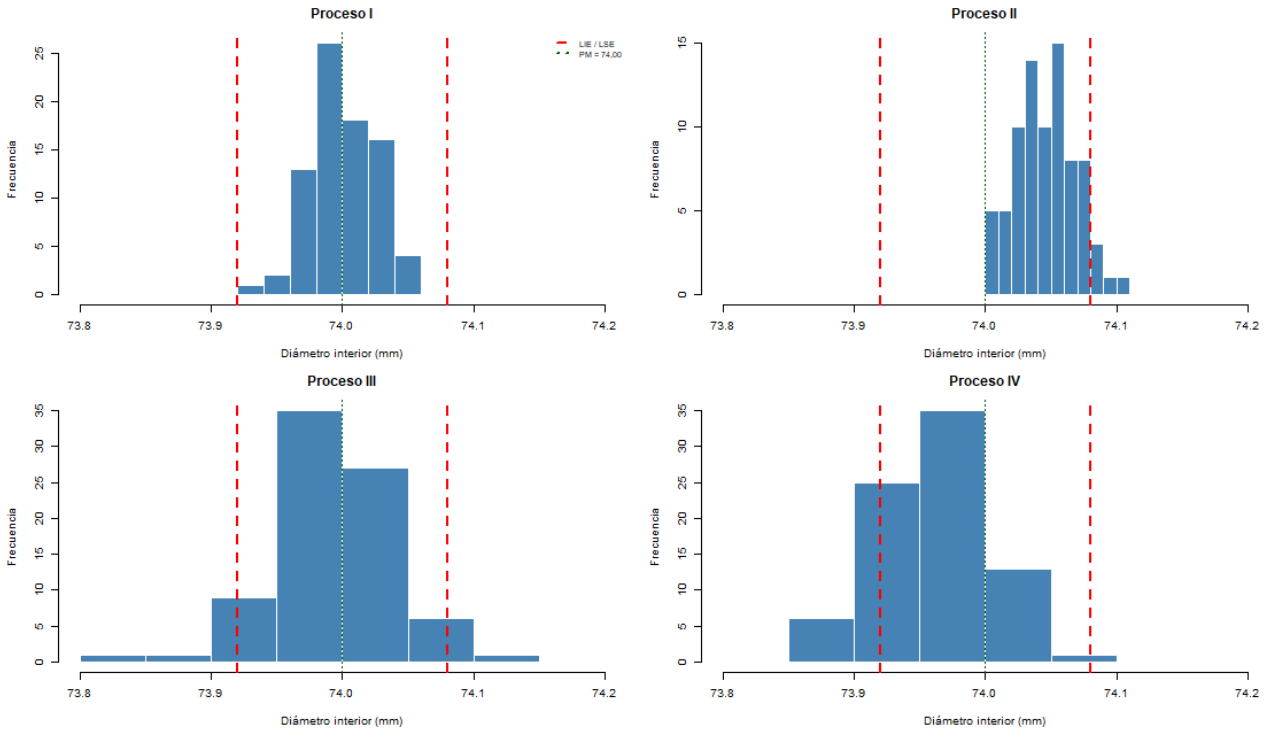
datosII <- c(
  74.073, 74.030, 74.082, 74.063, 74.038, 74.021, 74.029, 74.018, 74.091,
  74.049, 74.070, 74.017, 74.049, 74.071, 74.038, 74.068, 74.021, 74.015,
  74.053, 74.037, 74.059, 74.001, 74.061, 74.102, 74.044, 74.046, 74.049,
  74.004, 74.054, 74.010, 74.036, 74.072, 74.029, 74.056, 74.067, 74.051,
  74.015, 74.038, 74.088, 74.000, 74.039, 74.033, 74.048, 74.061, 74.080,
  74.039, 74.063, 74.056, 74.055, 74.033, 74.023, 74.021, 74.028, 74.052,
  74.062, 74.052, 74.033, 74.085, 74.078, 74.037, 74.027, 74.048, 74.073,
  74.052, 74.057, 74.040, 74.055, 74.079, 74.040, 74.048, 74.051, 74.051,
  74.014, 74.004, 74.023, 74.049, 74.057, 74.033, 74.075, 74.042
)
```

```
datosIII <- c(
  74.001, 73.990, 74.007, 74.037, 74.035, 73.953, 74.079, 73.943,
  74.004, 73.986, 73.945, 73.996, 73.961, 74.004, 74.021, 74.077,
  74.016, 73.968, 73.980, 74.007, 74.005, 74.076, 74.073, 73.981,
  73.971, 73.867, 73.971, 73.949, 73.982, 74.020, 74.009, 73.998,
  74.010, 73.993, 74.010, 73.832, 73.986, 73.994, 73.982, 73.989,
  74.037, 73.969, 74.114, 74.036, 74.000, 73.989, 73.996, 73.952,
  73.995, 74.014, 74.030, 74.030, 73.986, 73.926, 73.965, 74.084,
  74.047, 74.019, 73.960, 74.000, 73.935, 73.971, 73.987, 73.981,
  74.050, 73.926, 73.955, 73.929, 73.936, 74.006, 73.917, 74.039,
  73.987, 74.004, 74.017, 73.977, 74.054, 73.990, 74.031, 73.990
)

datosIV <- c(
  73.978, 73.978, 73.976, 73.921, 73.985, 73.951, 73.993, 73.868,
  73.899, 73.981, 74.049, 74.003, 73.934, 73.889, 73.998, 73.961,
  73.992, 73.970, 73.966, 74.098, 73.965, 73.958, 73.960, 73.939,
  74.010, 73.949, 73.994, 73.978, 74.020, 73.984, 73.910, 73.905,
  74.028, 73.938, 73.906, 73.935, 74.022, 73.901, 73.926, 73.935,
  73.909, 73.912, 73.897, 73.944, 73.949, 73.934, 73.891, 74.021,
  73.905, 73.978, 73.917, 73.995, 74.008, 73.926, 73.878, 73.996,
  74.049, 73.984, 73.972, 74.002, 73.999, 73.928, 73.935, 73.993,
  73.959, 74.024, 73.933, 73.999, 73.980, 73.959, 74.018, 73.996,
  73.920, 73.994, 74.027, 73.959, 73.954, 73.935, 73.964, 73.996
)
```

A continuación se presentan los histogramas correspondientes a cada situación, con los límites de especificación indicados mediante líneas punteadas y un conjunto de medidas de resumen.

Situacion	Media	Desvio	Prop_fuera_LIE	Prop_fuera_LSE	Prop_fuera_total
I	73.998	0.026	0.000	0.000	0.000
II	74.048	0.024	0.000	0.050	0.050
III	73.987	0.048	0.038	0.025	0.063
IV	73.958	0.041	0.162	0.013	0.175



Para cada una de las situaciones:

a) Evalúe si la distribución está centrada respecto del valor nominal. Justifique.

b) Evalúe si la variabilidad del proceso es adecuada en relación al ancho del intervalo de especificaciones. Justifique.

c) Comente sobre la proporción de piezas que no cumplen con las especificaciones. ¿Los rechazos se producen principalmente por el límite inferior, el superior o por ambos?

d) Concluya si el proceso es capaz de cumplir con las especificaciones de diseño. En caso de no serlo, identifique si la principal causa es la falta de centrado, el exceso de variabilidad o ambos.

e) Desde el punto de vista de un cliente que admite como máximo un 1 % de anillos fuera de especificación, ¿alguna de las situaciones analizadas puede considerarse aceptable? Justifique su respuesta.

f) ¿Cambiaría su conclusión si el cliente admitiera como máximo un 10 % de piezas fuera de especificación? Explique.

g) Para todas las conclusiones elaboradas en esta actividad, mencione si se pueden considerar preliminares o definitivas. Justifique.

27. Una empresa cuenta con una máquina destinada a producir componentes metálicos. La producción diaria (en unidades) presenta cierta variabilidad debido a condiciones operativas. Actualmente, el promedio de producción diaria es de 150.000 unidades. Se analizan dos propuestas para aumentar la producción:

- Incorporar un módulo adicional que incrementa la producción diaria en 15.000 unidades, cualquiera sea el nivel de producción alcanzado ese día.
- Implementar un sistema de optimización que incrementa la producción diaria multiplicándola por una constante a.

Suponga que el valor de la producción de un día determinado bajo las condiciones actuales se denota por x_i; mientras que bajo las nuevas propuestas se denota por y_i y z_i respectivamente.

a) Exprese a y_i y a z_i en función de x_i.

b) Encuentre el valor de la constante a, tal que, para ambas propuestas, la producción diaria promedio resulte igual.

c) ¿Resultarán semejantes las distribuciones de frecuencias de las producciones diarias con ambas propuestas? Justifique la respuesta.

d) Compare el coeficiente de variación de la producción diaria actual con el de la producción diaria bajo las nuevas propuestas. Comente sus conclusiones.

e) Todas las medidas mencionadas en esta actividad, ¿se pueden considerar estadísticos o parámetros? Justifique.

28. Una empresa se dedica a la reparación de equipos electrónicos. Al finalizar cada trabajo, factura por separado el costo de los materiales utilizados y el de la mano de obra. Para calcular el costo de la mano de obra, se multiplica la cantidad de horas trabajadas por \$5000 y al resultado se le suma un cargo fijo de \$10000.

La gerencia está interesada en caracterizar el costo de la mano de obra de sus servicios y solicita que se lleve adelante un estudio estadístico. Los responsables de realizar dicho estudio seleccionaron una muestra aleatoria de 100 reparaciones realizadas en los últimos 6 meses y encontraron lo siguiente:

- el tiempo promedio de reparación fue de 10 h.
- el desvío estándar fue de 0,5 h.
- el tiempo de reparación más frecuente fue de 7 h.
- los valores de los tres cuartiles resultaron, respectivamente, 5; 8,5 y 15 h.

a) Defina la población de interés en este caso.

b) Obtenga el monto promedio facturado en concepto de mano de obra y el desvío estándar para la muestra dada. Interprételos en contexto.

c) Calcule la moda y la mediana del monto facturado por mano de obra e interprételas en contexto.

d) Además del desvío estándar, ¿puede obtener alguna otra medida de dispersión? En caso afirmativo, calcúlela e interprétela.

e) Un posible cliente con gran cantidad de equipos desea contratar a la empresa siempre y cuando en el 75 % de las reparaciones el monto de mano de obra sea inferior a \$90000. Con la información muestral disponible, ¿considera que debería contratar a dicha empresa? ¿Su conclusión es preliminar o definitiva? Justifique.

29. Una empresa de servicios informáticos cuenta con un sistema de atención telefónica de los clientes, habilitado de lunes a viernes, de 9 a 17 horas. Dos operadores capacitados atienden las llamadas y buscan solucionar los problemas y reclamos ellos mismos y derivan, si es necesario, la consulta a un especialista, que actúa como supervisor.

Cuando un cliente no puede comunicarse telefónicamente con la empresa porque los operadores están ocupados, el sistema le permite dejar un mensaje breve. Este mensaje es luego analizado directamente por el supervisor, que devuelve la llamada al cliente y trata de solucionar su problema.

En la empresa consideran que, si el número promedio diario de mensajes breves es mayor que 4, debe incorporarse un nuevo operador al servicio, para no restar tiempo al supervisor, que tiene asignadas además otras tareas. En los últimos 45 días hábiles se registró el número de mensajes breves.

La planilla de volcado correspondiente se presenta a continuación:

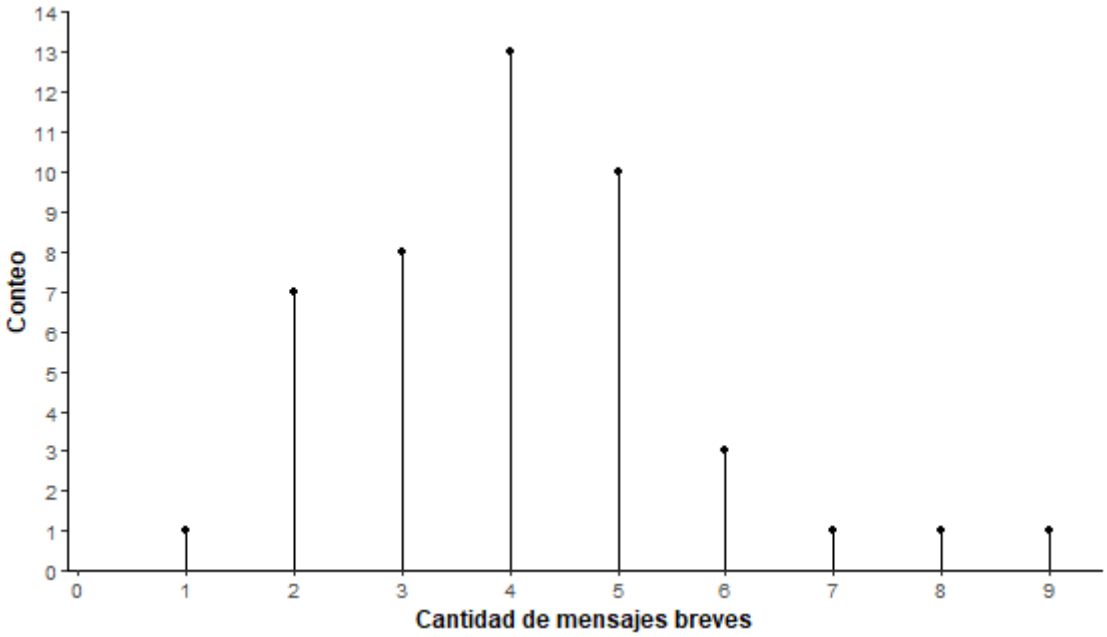
Fecha	Cantidad de mensajes breves	Fecha	Cantidad de mensajes breves	Fecha	Cantidad de mensajes breves
16-ago	4	6-sept	4	27-sept	5
19-ago	3	9-sept	2	30-sept	2
20-ago	5	10-sept	5	1-oct	3
21-ago	1	11-sept	5	2-oct	3
22-ago	4	12-sept	3	3-oct	2
23-ago	4	13-sept	4	4-oct	5
26-ago	5	16-sept	5	7-oct	8
27-ago	2	17-sept	9	8-oct	6
28-ago	3	18-sept	7	9-oct	2
29-ago	6	19-sept	4	10-oct	4
30-ago	4	20-sept	6	11-oct	4
2-sept	4	23-sept	2	14-oct	3
3-sept	2	24-sept	4	15-oct	5
4-sept	4	25-sept	3	16-oct	4
5-sept	5	26-sept	3	17-oct	5

a) Realice el planteo completo del problema.

b) Los analistas construyeron en primer lugar un diagrama de bastones. Usted, ¿está de acuerdo con ellos? Si no está de acuerdo, ¿qué otro gráfico considera que debieron construir en primer lugar? Comente.

c) Suponiendo que no hay inconvenientes para analizar la información brindada por la distribución de frecuencias, ¿qué decisión deben tomar en la empresa en relación a la contratación de un nuevo operador?

d) Indique si sus conclusiones son preliminares o definitivas.



30. Para cada una de las siguientes afirmaciones, indique si es verdadera o falsa. Justifique.

a) El primer cuartil es el punto medio entre el mínimo y la mediana.

b) En una distribución de frecuencias de forma simétrica, el rango es aproximadamente igual a 6 desviaciones estándar.

c) La moda de una distribución de frecuencias es igual a 4, por lo tanto 4 es el mayor valor encontrado en la muestra.

d) La media de una distribución de frecuencias es el valor que siempre ocurre con más frecuencia.

e) Dado un conjunto de datos, se puede decir que todos se encuentran comprendidos en el intervalo: media muestral +/- un desvío estándar

31. Reconsidere los problemas presentados en la Sección 1.4 (página 7).



Problema 1 (pág. 7):

a) Haga un resumen de todo lo actuado hasta ahora en relación a este problema, desde el planteo propiamente dicho.

b) ¿Qué puede decir de las barras producidas en la industria metalúrgica, en relación a las pretensiones de la empresa automotriz?

c) Las conclusiones que menciona, ¿se pueden considerar definitivas?



Problema 2 (pág. 7): *El conjunto de datos correspondiente al lote de placas de madera se encuentra en el archivo https://sites.google.com/view/libro-estadistica.*

a) ¿Considera que puede enviar el lote al cliente? Obtenga sus conclusiones siguiendo las etapas del ciclo PPDAC.

b) Pensando un poco más sobre este problema:

1) ¿Qué otra información hubiera registrado Ud. sobre las fallas de cada placa de madera, además de la cantidad total?, ¿con qué objetivo registraría esa información? ¿Cómo sería una planilla de recolección de datos acorde?

2) ¿Qué análisis podría hacer con esta información adicional?



Problema 4 (pág. 8):

a) Haga un resumen de todo lo actuado hasta ahora en relación a este problema, desde el planteo propiamente dicho.

b) ¿Qué puede decir de los motivos principales de las devoluciones?

c) Las conclusiones que menciona, ¿se pueden considerar definitivas? Justifique.



Problema 5 (pág. 8):

a) Haga un resumen de todo lo actuado hasta ahora en relación a este problema, desde el planteo mismo.

b) ¿Qué puede decir de las cantidades de imperfecciones que se presentan en los rollos de 100 m de cable unipolar de 2,5 mm?

c) Las conclusiones que menciona, ¿se pueden considerar definitivas?

2.6 Estadística con R

2.6.1 Análisis descriptivo de variables cualitativas

Para calcular las frecuencias absolutas y relativas correspondientes a una variable cualitativa, es posible utlilizar funciones que componen la versión *base* de R.

Si el conjunto de datos que se está utilizando se denomina *Base* y contiene una variable a analizar de tipo cualitativa llamada *Variable*, las frecuencias absolutas se pueden obtener sencillamente ejecutando la función `table(Base$Variable)`. Los resultados se mostrarán en el panel de *Consola*.

Para obtener las frecuencias relativas, se deben dividir las frecuencias absolutas sobre la cantidad de datos, el cual coincide con el número de filas del conjunto de datos, siempre que todos los datos se encuentren completos. Entonces, basta con ejecutar la sentencia `table(Base$Variable)/nrow(Base)`. Al igual que antes, los resultados se mostrarán en el panel de *Consola*. A continuación, se ejemplifica esta tarea para el Problema 4.



Problema 4 (pág. 8): *Para evaluar las causas de las devoluciones de productos adquiridos en línea, la empresa que los comercializa decide calcular las frecuencias absolutas y relativas correspondientes a cada clase. Para ello, se utiliza el siguiente script. Recuerde que las líneas que comienzan con # no son ejecutadas y sirven para comentar los realizado:*

```
#Paquetes a utilizar
library(readxl)
library(ggplot2)
library(dplyr)
library(forcats)

#Importa los datos
#Toma solo las dos primeras columnas, filas 1 a 351
Datos4 <- read_excel("C:/Users/Usuario/datos_problemas.xlsx",
  sheet="problema_4", range="A1:B351")

#Indica que la segunda columna de Datos (Datos[[2]]) debe ser considerada como
#variable cualitativa (factor)
Datos4[[2]] <- as.factor(Datos4[[2]])

#Recodifica la variable, indicando a qué clase corresponde
#cada número de la segunda columna de Datos (Datos[[2]])
Datos4[[2]] <- fct_recode(Datos4[[2]],
           "Errores facturación" = "1",
           "Ped. equivocado" = "2",
           "Precio no coincidente"= "3",
           "Prod. dañado" = "4",
           "Prod. mal embalado" = "5",
           "Prod. vencido/mal estado" = "6",
```

```
          "Retraso entrega" = "7")

#Calcula frecuencias absolutas y relativas
FrecAbsolutas <- table(Datos4$motivo_principal)
FrecRelativas <- round((table(
  Datos4$motivo_principal)/nrow(Datos4)),2)

#Une las frecuencias calculadas en una tabla y mostrarlas
Tabla <- cbind(FrecAbsolutas, FrecRelativas)
print(Tabla)
```

La tabla de frecuencias, "Tabla", se mostrará en el panel de Consola (Figura 2.29). También aparecerá disponible en el panel Environment. Para poder visualizarla, se debe hacer click sobre el ícono  *que se encuentra en la fila donde aparece el archivo, a la derecha.*

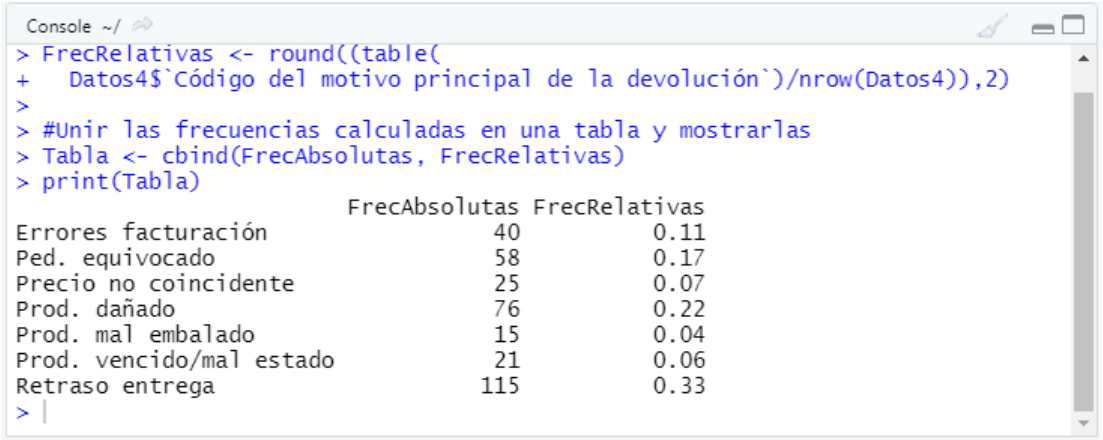


Figura 2.29 *Frecuencias absolutas y relativas correspondientes al motivo principal de devolución*

Para realizar los gráficos de barras, de sectores y de Pareto, es aconsejable utilizar el paquete `ggplot2` [10], dado que permite obtener gráficos elegantes y personalizados. En este libro no se explica en detalle cada una de las opciones, pero se presentan ejemplos con información suficiente para que puedan ser adaptados a otras situaciones donde se requiera un análisis similar.



Problema 4 (pág. 8): *Para representar las frecuencias absolutas correspondientes a los motivos principales de la devolución, se decide realizar un gráfico de barras. Una versión básica de este gráfico se puede obtener con el código que se presenta a continuación. El resultado obtenido puede visualizarse en la pestaña Plots del panel inferior derecho. Sin embargo, la figura resultante presenta detalles confusos que dificultan su lectura y utilidad:*

```
#Versión básica
ggplot(Datos4)+ #Se indica que se trabaja en base a Datos4
  #Se indica la variable a graficar
```

```
  aes(x=motivo_principal) +
  geom_bar() #Se pide un gráfico de barras
```

El gráfico anterior puede ser mejorado con algunas sentencias adicionales, que permitirán personalizar los nombres de los ejes y acomodar las etiquetas de cada categoría:

```
#Versión un poco mejorada
ggplot(Datos4)+ #Se indica que se en base a Datos4
  #Se indica la variable a graficar
  aes(x=motivo_principal) +
  geom_bar()+ #Se pide un gráfico de barras
  #Se especifican los nombres de los ejes
  labs(x = "Motivo principal de una devolución", y = "Nro. de devoluciones") +
  #Se inclinan las etiquetas de las categorías en el eje
  theme(axis.text.x = element_text (size=10, angle=30, hjust=1))
```

El gráfico de barras puede personalizarse aún más a partir de la tabla de frecuencias construida anteriormente. El siguiente código permite obtener el gráfico presentado en la Figura 2.1.b de la página 49:

```
#Guarda la tabla de frecuencias como data.frame
Tabla <- data.frame(Tabla)
#Se agrega a Tabla una columna con los motivos de devolución
Tabla$Motivo <- rownames(Tabla)

#Se indica que los datos se encuentran en el archivo Tabla
#En el eje horizontal se representa el motivo de devolución
#En el eje vertical se representan las frec aboslutas
ggplot(Tabla, aes(x=Motivo, y=FrecAbsolutas)) +
  #Configuración de color de borde y relleno de las barras
  geom_bar(stat="identity",color="black", fill="#f9b28c") +
  #Nombre de los ejes
  labs(x = "Motivo principal de una devolución", y = "Nro. de devoluciones") +
  #Configuraciones de formato
  scale_x_discrete(breaks = Tabla$Motivo) +
  guides(fill = FALSE, color = FALSE) +
  scale_y_continuous(expand=c(0,0), limits=c(0,120)) +
  #Estilo
  theme_classic()+
  #Fuente para los ejes
  theme(axis.title.x = element_text(face="bold", colour="black", size = 12),
        axis.title.y = element_text(face="bold", colour="black", size = 12),
        axis.text.x = element_text (size=10, angle=30, hjust=1))
```

Para obtener el gráfico con las frecuencias relativas, es suficiente con cambiar `y=FrecAbsolutas` *por* `y=FrecRelativas` *en la primera línea del código del gráfico, y adaptar la escala del eje vertical*

`scale_y_continuous(expand=c(0,0),limits=c(0,120))` *a, por ejemplo,* `scale_y_continuous(expand=c(0,0),limits=c(0,1))`.

Otra opción para representar las frecuencias relativas es el gráfico de sectores. Las siguientes líneas de código permiten obtener el gráfico presentado en la Figura 2.1.a de la página 49:

```
#Se indica que los datos se encuentran en el archivo Tabla
#Se asignan colores para las distintas clases de Motivo
ggplot(data = Tabla, aes(x = "", y = FrecRelativas,
                         fill = Motivo)) +
  #Color de borde de los sectores
  geom_bar(stat = "identity", color = "black") +
  #Para que realice un gráfico circular
  coord_polar("y") +
  #Configuraciones de formato
  theme_void()+
  #Agrega los porcentajes sobre cada sector
  geom_text(aes(label=paste(round(100*FrecRelativas, 2), "%")),
            position = position_stack(vjust=0.5), size=3)+
  #Tipo de fuente general
  theme(text=element_text(family="CMU Serif", size=12),
        plot.title = element_text(face="bold.italic", colour="black",
                                  size = 12))+
  #Paleta de colores para rellenar sectores
  scale_fill_brewer(palette="Accent")
```

Dado que la situación en estudio en este problema lo amerita, también se realiza un diagrama de Pareto. Para ello, es necesario ordenar la tabla de frecuencias en forma descendente, de acuerdo a las frecuencias absolutas y, luego, calcular las frecuencias relativas acumuladas, que pueden expresarse como porcentajes. Este gráfico presenta cierta complejidad, dado que se requiere la utilización de un eje vertical secundario para las frecuencias relativas acumuladas, que tiene una escala diferente a la del eje vertical principal, donde se representan las frecuencias absolutas. El siguiente script permite su realización, y el gráfico obtenido se muestra en la Figura 2.2 de la página 50:

```
#Ordena la tabla de frecuencias en forma decreciente s/frec. abs.
Tabla <- Tabla[order(Tabla$FrecAbsolutas, decreasing = TRUE),]
#Modifica el formato de la variable Motivo
Tabla$Motivo <- factor(Tabla$Motivo, levels=Tabla$Motivo)
#Se generan las frecuencias relativas acumuladas (en porcentaje)
Tabla$Acumulada <- cumsum(Tabla$FrecRelativas)*100

#Dado que se requieren dos ejes verticales que tienen escalas
#diferentes, es necesario corregirlas por cierto factor (sf)
sf<-max(Tabla$Acumulada)/max(Tabla$FrecAbsolutas)

#Se indica que los datos se encuentran en el archivo Tabla
```

```
#En el eje horizontal se representa el motivo de devolucion
#En el eje vertical se representan las frec aboslutas
ggplot(Tabla, aes(x=Motivo, y=FrecAbsolutas)) +
  #Configuración de color de borde y relleno de las barras
  geom_bar(stat="identity",color="black", fill="#f9b28c") +
  #Se traza un diagrama de lineas para representar
  #las frecuencias relativas acumuladas a traves de los motivos
  geom_line(aes(x=as.integer(Motivo), y = Acumulada/sf)) +
  #Se agregan puntos en los pares ordenados (Motivo, frec. acum.)
  geom_point(aes(x=as.integer(Motivo), y = Acumulada/sf),
             pch = 19) +
  #Configuraciones generales
  scale_x_discrete(breaks = Tabla$Motivo) +
  guides(fill = FALSE, color = FALSE) +
  scale_y_continuous(expand=c(0,0), limits=c(0,120),
                     sec.axis = sec_axis(~.*sf, name = "Frecuencia acumulada (%)"))+
  #Nombres de los ejes
  labs(x ="Motivo principal", y = "Nro. de devoluciones") +
  #Configuraciones de formato
  #Estilo
  theme_classic()+
  #Fuente para los ejes
  theme(axis.text.x = element_text (size=10, angle=30, hjust=1))
```

2.6.2 Análisis descriptivo de variables cuantitativas discretas

Para calcular las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas correspondientes a una variable cuantitativa discreta, es posible utlilizar funciones que componen la versión *base* de R, de manera similar a lo realizado para el caso de las variables cualitativas.

Si el conjunto de datos que se está utilizando se denomina *Base* y contiene una variable a analizar de tipo cuantitativa discreta llamada *Variable*, las frecuencias absolutas se pueden obtener sencillamente ejecutando la función `table(Base$Variable)`. Los resultados se mostrarán en el panel de *Consola*.

Para obtener las frecuencias relativas, se deben dividir las frecuencias absolutas sobre la cantidad de datos, el cual coincide con el número de filas del conjunto de datos, siempre que todos los datos se encuentren completos. Entonces, basta con ejecutar la sentencia `table(Base$Variable)/nrow(Base)`. Al igual que antes, los resultados se mostrarán en el panel de *Consola*.

Si se pretende obtener las frecuencias absolutas acumuladas, será necesario guardar, previamente, las frecuencias absolutas calculadas, para luego usar la función `cumsum`. Por ejemplo:

```
Absolutas <- table(Base$Variable)
print(Absolutas)
AbsolutasAcumuladas <- cumsum(Absolutas)
print(AbsolutasAcumuladas)
```

Para la realización de los gráficos de bastones, se recomienda el uso del paquete `ggplot2`. Partiendo de la tabla de frecuencias realizada anteriormente, es posible utilizar segmentos y puntos para lograr la representación deseada, ya sea considerando las frecuencias absolutas o las frecuencias relativas. Similarmente, se puede obtener el gráfico escalonado, para representar las frecuencias relativas acumuladas. A continuación, se presenta el *script* correspondiente a la obtención de la tabla de frecuencias y los gráficos de bastones y escalonado para los datos del Problema 2.



Problema 2 (pág. 7): *Se desea realizar la tabla de frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas para el número de defectos por placa. Además, se quiere obtener la representación de la información mediante un gráfico de bastones y un gráfico escalonado.*

```
#Paquetes a utilizar
library(readxl)
library(ggplot2)

#Levanta los datos
Datos2 <- read_excel("C:/Users/Usuario/datos_problemas.xlsx",
sheet="problema_2")

#Calcula frecuencias absolutas, relativas y acumuladas
FrecAbsolutas <- table(Datos2$nro_defectos)
FrecRelativas <- table(Datos2$nro_defectos)/nrow(Datos2)
AbsAcumuladas <- cumsum(FrecAbsolutas)
RelAcumuladas <- cumsum(FrecRelativas)

#Une las frecuencias calculadas en una tabla
Tabla <- cbind(FrecAbsolutas, FrecRelativas, AbsAcumuladas, RelAcumuladas)

#Guarda la tabla en un conjunto de datos
Tabla <- data.frame(Tabla)
```

La tabla de frecuencias, "Tabla", aparecerá disponible en el panel Environment. Para poder visualizarla, se debe hacer click sobre el ícono  *que se encuentra en la fila donde aparece el archivo, a la derecha. Una versión básica del gráfico de bastones puede obtenerse mediante el código que se presenta a continuación. Sin embargo, el mismo carece de algunos detalles dificultando así su correcta lectura e interpretación:*

```
#Agrega una columna con los valores de la variable
Tabla$NroDefectos<-as.numeric(rownames(Tabla))

#Gráfico de bastones con frecuencias absolutas
#Versión básica
ggplot(data=Tabla) +
  #Se trazan bastones sobre c/valor de la variable (NroDefectos)
  #Los bastones van desde 0 hasta c/frec. absoluta
```

```
  geom_segment(aes(x=NroDefectos,y=0,xend=NroDefectos,
                   yend=FrecAbsolutas)) +
  #Se agregan puntos en los pares ordenados (valor, frec. abs.)
  geom_point(aes(NroDefectos,FrecAbsolutas))
```

A continuación, se muestra el código considerado para la obtención del gráfico de bastones en su versión mejorada (Figura 2.30) y del gráfico escalonado correspondiente (Figura 2.31). Los gráficos se pueden visualizar en la pestaña Plots, en el panel inferior derecho.

```
#Gráfico de bastones con frecuencias absolutas
#Se indica el nombre de la tabla de frecuencias
ggplot(data=Tabla) +
  #Se trazan bastones sobre c/valor de la variable (NroDefectos)
  #Los bastones van desde 0 hasta c/frec. absoluta
  geom_segment(aes(x=NroDefectos,y=0,xend=NroDefectos, yend=FrecAbsolutas)) +
  #Se agregan puntos en los pares ordenados (valor, frec. abs.)
  geom_point(aes(NroDefectos,FrecAbsolutas),size=1.5) +
  #Se indican los nombres de los ejes
  labs(x = "Cantidad de defectos por placa", y = "Cantidad de placas") +
  #Configuraciones de formato
  #Estilo
  theme_classic()+
  #Fuente para los ejes
  theme(axis.title.x = element_text(face="bold", colour="black", size = 12),
        axis.title.y = element_text(face="bold", colour="black", size = 12))+
  #Límites de los ejes
  scale_x_continuous(expand=c(0,0), limits = c(-0.1,5.5), breaks=c(0:5)) +
  scale_y_continuous(expand=c(0,0), limits = c(0,30))

#Gráfico escalonado
#Se indica el nombre de la tabla de frecuencias
ggplot(data=Tabla) +
  #Se trazan segmentos horizontales a la altura de la frecuencia relativa acumulada
  #La posición respecto al eje horizontal comienza en los distintos
  #valores de la variable (NroDefectos), hasta el siguiente (NroDefectos+1)
  geom_segment(aes(x=NroDefectos,y=RelAcumuladas,
  xend=NroDefectos+1, yend=RelAcumuladas)) +
  #Se agrega un último segmento para la frec. relativa acum.
  geom_segment(aes(5,1,xend=5.5,yend=1)) +
  #Se dibujan círculos abiertos en los pares ordenados
  #(valor variable, frec. rel. acum.)
  geom_point(aes(NroDefectos,RelAcumuladas),size=1.5, shape=16) +
  #Nombre de los ejes
  labs(x = "Cantidad de defectos por placa", y = "Proporción acumulada de placas") +
  #Configuraciones de formato
  #Estilo
```

```
theme_classic()+
#Configuración fuente de ejes
theme(axis.title.x = element_text(face="bold", colour="black", size = 12),
      axis.title.y = element_text(face="bold", colour="black", size = 12))+
#Límites de los ejes
scale_x_continuous(expand=c(0,0), limits = c(-0.1,5.5), breaks=c(0:5)) +
scale_y_continuous(expand=c(0,0), limits = c(0,1.05),
                   labels = scales::label_number(accuracy = 0.01, decimal.mark = ','))
```

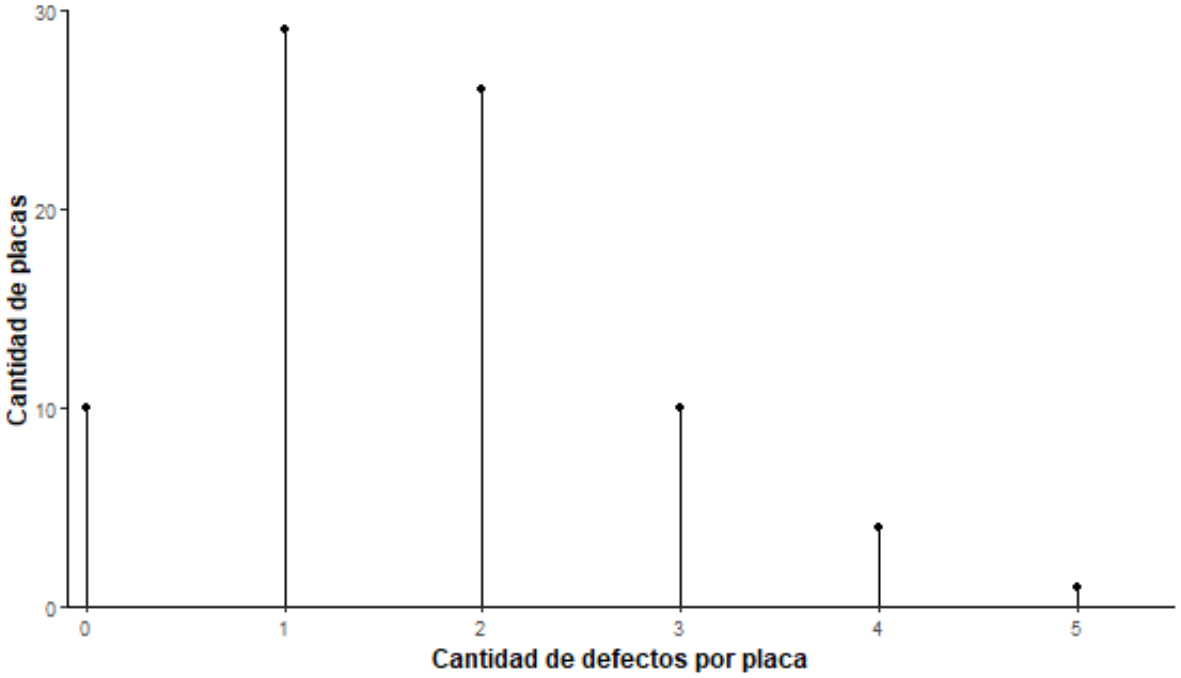


Figura 2.30 *Frecuencias absolutas de la cantidad de defectos por placa*

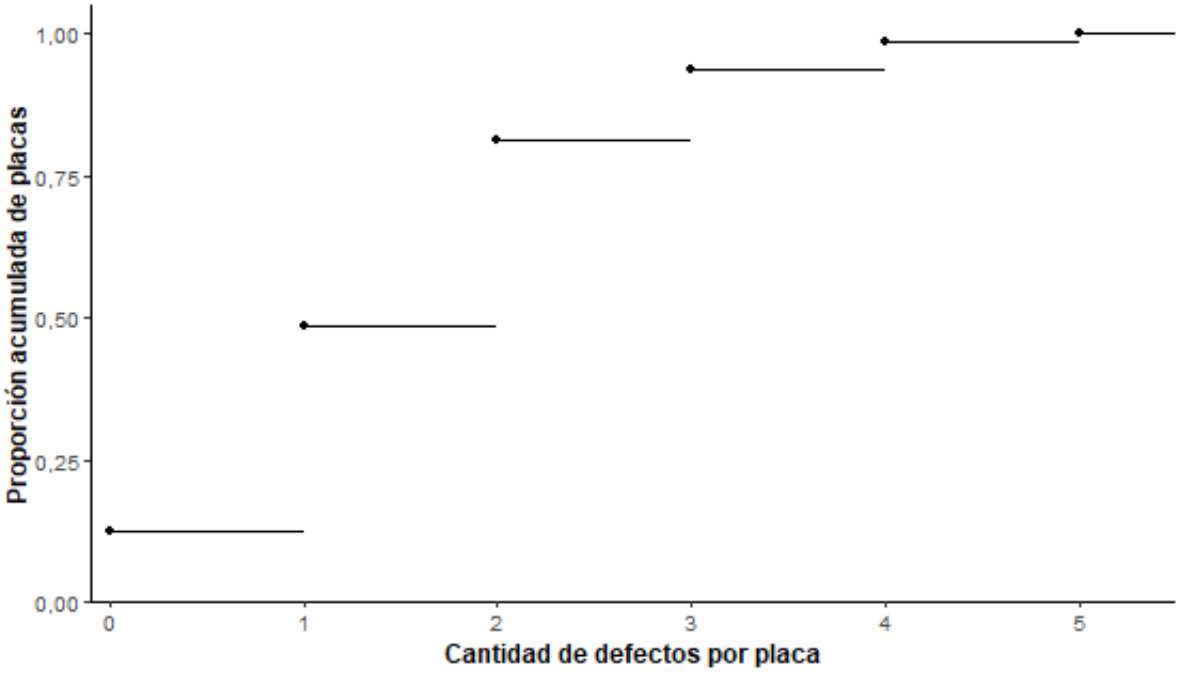


Figura 2.31 *Frecuencias relativas acumuladas de la cantidad de defectos por placa*

2.6.3 Análisis descriptivo de variables cuantitativas continuas

Como ya se ha estudiado, para un correcto análisis de datos correspondientes a una variable cuantitativa continua, es útil agruparlos en intervalos de clase. Una vez definida la cantidad de intervalos de clase a considerar, se puede determinar su amplitud y definir, entonces, los límites de cada uno de ellos.

Para realizar la tabla de frecuencias correspondiente a la variable cuantitativa en estudio, se consideran los intervalos de clase definidos y se realiza, en principio, el conteo de datos que se encuentran dentro de cada uno de ellos, determinando así las frecuencias absolutas y, a partir de ellas, se pueden calcular las frecuencias relativas, absolutas acumuladas y relativas acumuladas. A continuación, se presenta un ejemplo del armado de la tabla de frecuencias para los datos del Problema 1.



Problema 1 (pág. 7): *Para organizar los datos correspondientes a las longitudes de las 120 barras (en mm), se decide agruparlos en 8 intervalos de clase, cada uno de amplitud 0,2 mm. Para que todos los valores sean incluidos, el primer intervalo tiene como límite inferior 249,2 mm, mientras que el último intervalo tiene como límite superior 250,8 mm. El siguiente script permite agrupar los datos según estos intervalos y realizar la tabla de frecuencias.*

```
#Levanta los datos
Datos1 <- read_excel("C:/Users/Usuario/datos_problemas.xlsx",
sheet="problema_1")

#Una vez establecido el número de intervalos a utilizar
#se fija el límite inferior del primer intervalo (min)
#y el límite superior del último intervalo (max)
#IMPORTANTE tener en cuenta que se van a generar intervalos abiertos por
#izquierda y cerrados por derecha
min <- 249.20
max <- 250.80
intervalos <- 8

#Se genera una secuencia con los valores límites de los intervalos
#partiendo desde min, hasta max
cortes <- seq(min, max, length.out=intervalos+1)

#Se dividen los datos de las longitudes de las barras en intervalos
#abiertos por izquierda y cerrados por derecha
división <- cut(as.double(Datos1$longitud), cortes)

#En función a la división de los datos en intervalos, se calculan
#las frecuencias absolutas, relativas y acumuladas
#en forma similar a las variables cuantitativas discretas
FrecAbsolutas <- table(división)
FrecRelativas <- round((FrecAbsolutas/nrow(Datos1)),2)
```

```
AbsAcumuladas <- cumsum(FrecAbsolutas)
RelAcumuladas <- cumsum(FrecRelativas)

#Une las frecuencias calculadas en una tabla
Tabla <- cbind(FrecAbsolutas, FrecRelativas, AbsAcumuladas, RelAcumuladas)

print(Tabla)
```

La tabla de frecuencias, "Tabla", se mostrará en el panel de Consola (Figura 2.32). También aparecerá disponible en el panel Environment. Para poder visualizarla, se debe hacer click sobre el ícono  *que se encuentra en la fila donde aparece el archivo, a la derecha.*

```
Console ~/
> RelAcumuladas <- cumsum(FrecRelativas)
> #Unir las frecuencias calculadas en una tabla
> Tabla <- cbind(FrecAbsolutas, FrecRelativas, AbsAcumuladas,
+                RelAcumuladas)
> print(Tabla)
             FrecAbsolutas FrecRelativas AbsAcumuladas RelAcumuladas
(249.2,249.4]             3          0.02             3          0.02
(249.4,249.6]             8          0.07            11          0.09
(249.6,249.8]            26          0.22            37          0.31
(249.8,250]              23          0.19            60          0.50
(250,250.2]              30          0.25            90          0.75
(250.2,250.4]            21          0.18           111          0.93
(250.4,250.6]             6          0.05           117          0.98
(250.6,250.8]             3          0.02           120          1.00
>
```

Figura 2.32 *Frecuencias absolutas, relativas y acumuladas de las longitudes (en mm) de las barras*

Una vez establecidos los puntos de cortes o límites de los intervalos, es posible realizar el histograma de frecuencias absolutas utilizando el paquete `ggplot`, el cual permite adicional, si se requiere, el polígono de frecuencias.



Problema 1 (pág. 7): *Se realiza el histograma para representar las frecuencias absolutas en función de los intervalos especificados anteriormente. Además, se traza el polígono de frecuencias. El gráfico obtenido corresponde al presentado en la Figura 2.7.a de la página 62.*

```
ggplot(Datos1, aes(longitud)) +
  #Se solicita realizar un histograma y se especifican los colores
  #y los límites de los intervalos con la opción (breaks=cortes)
  geom_histogram(aes(y=..count..),
                 color="black", fill="#f9b28c", breaks=cortes) +
  #Optativo: se agrega el polígono de frecuencias
  geom_freqpoly(aes(longitud), breaks=cortes)+
  scale_x_continuous(breaks=cortes)+
  #Configuraciones generales
```

```
  #Nombres de los ejes
  labs(x = "Longitud (mm)", y = "Cantidad de barras") +
  #Configuraciones de formato
  #Estilo
  theme_classic()+
  scale_y_continuous(expand=c(0,0))+
  #Fuente para los ejes
  theme(axis.title.x = element_text(face="bold", colour="black", size = 12),
        axis.title.y = element_text(face="bold", colour="black", size = 12))
```

El resultado puede visualizarse en la pestaña Plots del panel inferior derecho.

Además, se puede realizar el polígono para representar las frecuencias relativas acumuladas. Para ello, se cuenta con estas frecuencias acumuladas en la tabla correspondiente, aunque es necesario indicar que antes del primer intervalo, no se acumulan datos. Esto se logra adicionando el valor 0 a las frecuencias relativas acumuladas ya generadas.



Problema 1 (pág. 7): *A continuación, se muestra el código utilizado para generar el polígono de frecuencias acumuladas, que se presenta en la Figura 2.7.b de la página 62.*

```
#Polígono frecuencias acumuladas
#Se agrega el valor 0 al inicio de las frec. relativas acumuladas
RelAcumuladas <- c(0, RelAcumuladas)

#Se une en un archivo de datos la información sobre los límites
#de los intervalos y las frecuencias relativas acumuladas
Base <- data.frame(cbind(cortes, RelAcumuladas))

#Polígono de frecuencias relativas acumuladas
ggplot(Base, aes(x = cortes, y = RelAcumuladas)) +
  geom_line() + geom_point()+
  scale_x_continuous(breaks=cortes)+
  #Configuraciones generales
  #Nombres de los ejes
  labs(x = "Longitud de una barra (en mm)", y = "Proporción acumulada de barras") +
  #Configuraciones de formato
  #Estilo
  theme_classic()+
  scale_y_continuous(expand=c(0,0), limits = c(0,1.05),
                     labels = scales::label_number(accuracy = 0.01,
                                                   decimal.mark = ','))+
  #Fuente para los ejes
  theme(axis.title.x = element_text(face="bold", colour="black", size = 12),
        axis.title.y = element_text(face="bold", colour="black", size = 12))
```

2.6.4 Otros gráficos

1. Diagrama de puntos

Mediante el paquete `ggplot2`, es posible realizar un diagrama de puntos de manera sencilla mediante la función específica `geom_dotplot`. Se presenta a continuación un ejemplo de su aplicación sobre los datos correspondientes al Problema 2.



Problema 2 (pág. 7): *El siguiente código permite realizar y personalizar un diagrama de puntos para representar gráficamente los datos correspondientes al número de defectos por placa.*

```
#Paquetes a utilizar
library(readxl)
library(ggplot2)

#Levantar los datos
Datos2 <- read_excel("C:/Users/Usuario/datos_problemas.xlsx",
  sheet="problema_2")

ggplot(Datos2, aes(x = nro_defectos)) +
  geom_dotplot(dotsize=0.5, color="black", fill="#f9b28c")+
  #Nombres de los ejes
  labs(x = "Número de defectos de una placa") +
  #Configuraciones de formato
  #Estilo
  theme_classic()+
  scale_y_continuous(expand=c(0,0), NULL, breaks = NULL)+
   #Configuración para los ejes
  theme(axis.title.x = element_text(face="bold", colour="black", size = 12),
        axis.line.y = element_blank())
```

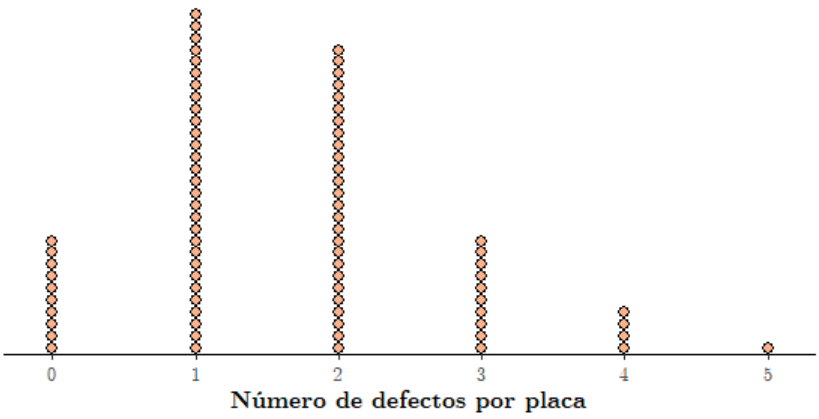


Figura 2.33 *Diagrama de puntos para el número de defectos por placa*

2. Diagrama de tallo y hojas

La función `stem` (*tallo*), disponible en la *base* de R, permite generar en forma muy sencilla un diagrama de tallo y hojas. Si se dispone de un conjunto de datos que se denomina *Base* y contiene una variable cuantitativa llamada *Variable*, entonces el diagrama de tallo y hojas para analizar el comportamiento de los datos de esa variable se obtiene ejecutando la sentencia `stem(Base$Variable)`. El resultado se mostrará en el panel de *Consola*.



Problema 1 (pág. 7): *Se quiere realizar un diagrama de tallo y hojas para representar los datos de las longitudes de las 120 barras, medidas en mm. Se ejecuta el siguiente script:*

```
#Paquetes a utilizar
library(readxl)

Datos1 <- read_excel("C:/Users/Usuario/datos_problemas.xlsx", sheet="problema_1")

stem(Datos1$longitud)
```

```
Console C:/Users/Usuario/Desktop/FCEIA/Libro/

  The decimal point is 1 digit(s) to the left of the |

  2492 | 9
  2493 | 34
  2494 | 59
  2495 | 3688
  2496 | 001222345588
  2497 | 0133557777899
  2498 | 0001125788
  2499 | 002233334567789
  2500 | 01334444555677888
  2501 | 002444445688
  2502 | 001234666699
  2503 | 011334567
  2504 | 0012
  2505 | 068
  2506 | 037
  2507 | 1

>
```

Figura 2.34 *Diagrama de tallo y hojas para las longitudes de las barras (en mm)*

3. Serie de tiempo

Para graficar datos que fueron obtenidos a lo largo del tiempo, se puede utilizar la función `geom_line` del paquete `ggplot2`, poniendo atención al formato en el que se encuentran los valores que indican el momento de obtención de cada dato (fechas, horas, combinación de ambas, otros formatos). Esto permitirá que el software trate de manera adecuada a la escala horizontal, en la que se refleja el tiempo.

Es habitual que, si los datos están guardados en un archivo de Microsoft Excel o de texto plano, R los identifique, en principio, como una cadena de caracteres. Una vez levantados los datos en R, es posible verificar el formato de lectura de cada variable dentro de la pestaña *Environment* del panel superior derecho, haciendo click en el ícono  a la izquierda del conjunto de datos en cuestión. Así aparecerá un listado de los nombres de las variables del conjunto de datos con un indicador del formato de lectura.

Si una columna contiene información sobre horarios y/o fechas y presenta un formato distinto a *Date*, será necesario modificarlo para su adecuado tratamiento. Para ello, serán de utilidad las funciones `as.Date` y el paquete `zoo`.



Problema 9 (pág. 9): *En la empresa fabricante de productos de limpieza, están interesados en conocer el comportamiento de la demanda trimestral de desengrasante en los últimos 5 años para poder predecir valores futuros y planificar la producción para el próximo año. Para ello, se grafican los datos obtenidos en cada trimestre. El resultado se muestra en la Figura 2.12 de la página 66.*

```
#Paquetes a utilizar
library(readxl)
library(ggplot2)
library(zoo)

Datos9 <- read_excel("C:/Users/Usuario/datos_problemas.xlsx",
sheet="problema_9")

#Se indica que los datos de la variable Período son fechas
#con formato de 4 cifras para el año (Y) y luego el número del trimestre (m)

Datos9$fecha <- as.yearqtr(Datos9$fecha, format = "%Y-%q")
Datos9$fecha <- as.numeric(Datos9$fecha)

ggplot(Datos9, aes(x = fecha, y = unidades_vendidas, group=1)) +
  geom_line()+
  labs(x = "Año - Trimestre", y = "Cantidad de unidades vendidas de un trimestre") +
  scale_x_continuous(breaks = unique(Datos9$fecha),
                     labels = format(as.yearqtr(unique(Datos9$fecha)), "%Y-%q")) +
  scale_y_continuous(expand=c(0,0), labels = scales::label_number(),
                     limits = c(0, 800000), breaks=seq(0,800000, 100000)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, size = 8))
```

4. Diagrama de caja y bigotes

Para la realización de diagramas de caja y bigotes (*box-plots*), el paquete `ggplot2` ofrece una función específica: `geom_boxplot`. Este gráfico se puede personalizar de forma similar a los anteriormente vistos.

Para determinar el alcance de los *bigotes*, el software calcula el menor de los valores del conjunto de datos que no es inferior al cuartil 1 menos 1,5 veces el rango intercuartil y extiende hasta él el bigote inferior, y el mayor de los valores del conjunto de datos que no supera al cuartil 3 más 1,5 veces el rango intercuartil y extiende hasta él el bigote superior. En caso de existir valores menores al extremo del bigote inferior o mayores al extremo del bigote superior, son considerados valores atípicos (o *outliers*) y se grafican mediante puntos.

Problema 1 (pág. 7): *Se presenta a continuación el script necesario para realizar un diagrama de caja y bigotes que resuma la información correspondiente a las 120 mediciones de las longitudes de las barras, en mm.*

```
#Paquetes a utilizar
library(readxl)
library(ggplot2)

Datos1 <- read_excel("C:/Users/Usuario/datos_problemas.xlsx",
sheet="problema_1")

ggplot(Datos1, aes(x = longitud, y="")) +
  #se trazan líneas en los extremos de los bigotes
  stat_boxplot(geom ='errorbar', width = 0.1) +
  #Se establece el color de relleno de la caja
  geom_boxplot(fill = "#f9b28c") +
  #Se simboliza con un punto el valor del promedio (mean)
  stat_summary(fun = mean, geom = "point")+
  #Configuraciones generales
  #Nombres de los ejes
  labs(x = "Longitud (mm)", y = "") +
  #Configuraciones de formato
  #Estilo
  theme_classic()+
  scale_x_continuous(breaks = seq(249.2,250.8,by=0.2))+
  #Fuente para los ejes
  theme(axis.title.x = element_text(face="bold", colour="black", size = 12),
        axis.line.y = element_blank(),
        axis.ticks.y = element_blank())
```

El gráfico resultante se encuentra en la pestaña Plots del panel inferior derecho. En el mismo se observa que no hay datos atípicos, por lo que los bigotes se extienden hasta el valor mínimo y el máximo (Figura 2.21 de la página 91).
Si, por alguna razón, se modificara el valor mínimo del conjunto de datos (249,29 mm) por el valor 249 mm, ese nuevo valor resultaría menor al cuartil 1 menos 1,5 veces el rango intercuartil, por lo que se consideraría atípico. El bigote inferior llegaría hasta el valor siguiente en orden de magnitud y el gráfico resultaría como en la Figura 2.35.

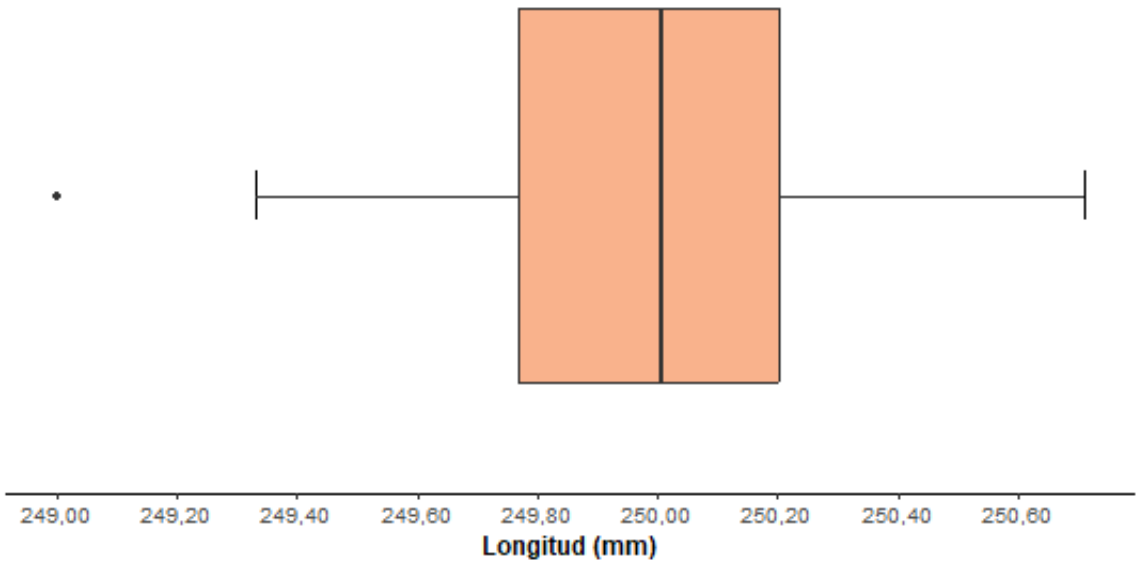


Figura 2.35 *Diagrama de caja y bigotes correspondiente a las longitudes de las barras analizadas, con valor mínimo modificado*

2.6.5 Medidas de resumen

Una función muy útil para calcular las principales medidas de resumen de variables cuantitativas es `summarise` (o, equivalentemente, `summarize`), que es parte del paquete `dplyr` [11]. Indicando el nombre del conjunto de datos a utilizar, es posible definir sencillamente las medidas a calcular, especificando entre paréntesis el nombre de la variable.

A continuación, se presenta como ejemplo el cálculo de las medidas de resumen de la variable de interés en el Problema 1, el cual puede adaptarse adecuadamente para ser empleado en otras situaciones.



Problema 1 (pág. 7): *Se desea obtener las medidas de resumen para la variable Longitud de la barra (en mm). El siguiente script permite cargar los paquetes necesarios, levantar los datos, guardarlos con el nombre de "Datos1" y obtener el valor mínimo, el máximo, el rango, la media, los cuartiles 1, 2 (mediana) y 3, el percentil de orden 0,90, la varianza, el desvío estándar y el coeficiente de variación para la variable en estudio. Estas medidas son guardadas en un nuevo conjunto de datos llamado "Medidas".*

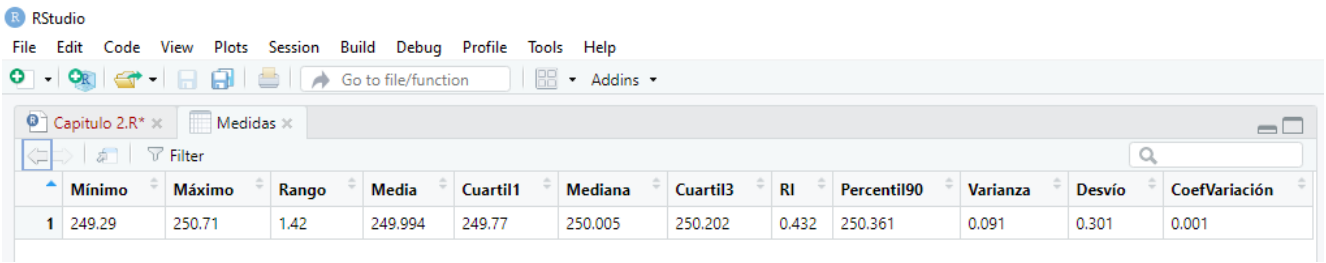
```
#Paquetes a utilizar
library(readxl)
library(dplyr)

#Levanta los datos
Datos1 <- read_excel("C:/Users/Usuario/datos_problemas.xlsx",
sheet="problema_1")

#Calcula medidas descriptivas
```

```
Medidas <- summarise(Datos1,
  Mínimo=min(longitud),
  Máximo=max(longitud),
  Rango= Máximo - Mínimo,
  Media=mean(longitud),
  Cuartil1=quantile(longitud,probs=0.25),
  Mediana=median(longitud),
  Cuartil3=quantile(longitud,probs=0.75),
  RI=Cuartil3 - Cuartil1,
  Percentil90=quantile(longitud,probs=0.90),
  Varianza=var(longitud),
  Desvío=sd(longitud),
  CoefVariación=Desvío/Media)
  Medidas <- round(Medidas,3)
```

El conjunto de datos generado, "Medidas", aparecerá disponible en el panel Environment. Para poder visualizarlo, se debe hacer click sobre el ícono  *que se encuentra en la fila donde aparece el conjunto de datos, a la derecha.*



	Mínimo	Máximo	Rango	Media	Cuartil1	Mediana	Cuartil3	RI	Percentil90	Varianza	Desvío	CoefVariación
1	249.29	250.71	1.42	249.994	249.77	250.005	250.202	0.432	250.361	0.091	0.301	0.001

Figura 2.36 *Medidas descriptivas para las longitudes (en mm) de las barras*

2.7 Apéndice

2.7.1 Medidas de resumen obtenidas en la resolución de los problemas 1, 4 y 5

A continuación se presentan los valores de algunas medidas de resumen, obtenidos con información recolectada para la resolución de los Problemas 1, 4 y 5, y sus interpretaciones.



Problema 1 (pág. 7):

- *Población: todas las barras de cierto tipo producidas por la metalúrgica santafesina.*
- *Muestra: n = 120 barras.*
- *Variable: longitud de una barra de acero (mm), cuantitativa, medida en escala de razón.*

Tabla 2.29 *Medidas de resumen para las longitudes de las 120 barras de acero analizadas*

Medida	*Valor*	*Interpretación*
Valores mínimo y máximo	*249,29 mm y 250,71 mm respectivamente*	*La barra con menor longitud entre las analizadas medía 249,29 mm; mientras que la de mayor longitud medía 250,71 mm.*
Cuartil 1	*249,77 mm*	*El 25 % de las barras analizadas tiene longitud menor o igual que 249,77 mm y el 75 % restante, 249,77 mm o más.*
Cuartil 3	*250,205 mm*	*El 75 % de las barras analizadas tiene longitud menor o igual que 250,205 mm y el 25 % restante, 250,205 mm o más.*
Promedio o media aritmética	*249,994 mm (249,985 mm)*	*En promedio, las barras analizadas miden 249,994 mm de longitud.*
Media truncada 10 %	*249,99 mm*	*En promedio, las barras analizadas miden 249,99 mm de longitud (excluyendo el 10 % de las barras de menor longitud y el 10 % de las barras con de mayor longitud)*
Media geométrica	*249,994 mm*	*En promedio, las barras analizadas miden 249,994 mm de longitud.*
Moda	*No tiene sentido obtenerla con datos sin agrupar. Intervalo modal: ((250 ; 250,2] mm)*	*Más frecuentemente las barras analizadas miden entre 250 mm y 250,2 mm de longitud.*
Mediana	*250,005 mm*	*El 50 % de las barras analizadas tiene longitud menor o igual que 250,005 mm y el 50 % restante, 250,005 mm o más.*
Rango	*1,42 mm*	*La máxima diferencia que se observa entre dos de las barras analizadas es de 1,42 mm.*
Rango intercuartílico	*0,435 mm*	*La máxima diferencia que se observa en el 50 % central de las barras es de 0,435 mm.*
Variancia	*0,091 mm^2 (0,094 mm^2)*	*En promedio, los desvíos cuadrados entre la longitud de cada barra y el valor promedio de las barras analizadas vale 0,091 mm^2.*
Desvío estándar	*0,300 mm (0,307 mm)*	*"En promedio", las longitudes de las barras analizadas se desvían en aprox. 0,3 mm del promedio.*
Coeficiente de variación	*0,12 %*	*El desvío estándar de las barras analizadas representa un 0,12 % del valor de la media aritmética.*
Proporción de barras con longitud menor que 249,4 mm	*0,025*	*La proporción de barras con longitud por debajo del límite inferior de especificaciones entre las analizadas es 0,025 (2,5 %).*
Proporción de barras con longitud fuera de especificaciones	*0,05*	*La proporción de barras con longitud fuera de especificaciones es 0,05. Es decir que el 5 % de las barras analizadas no cumple con las especificaciones.*

Observación: Entre paréntesis se indican los valores aproximados de las medidas, obtenidos a partir de los datos agrupados en una tabla de distribución de frecuencias.



Problema 5 (pág. 8):

- *Población: rollos de 100 m de cable unipolar de 2,5 mm.*
- *Muestra: n = 50 rollos.*
- *Variable: cantidad de imperfecciones de un rollo, cuantitativa, medida en escala de razón.*

Tabla 2.30 *Medidas de resumen para las cantidades de imperfecciones de los 50 rollos analizados*

Medida	*Valor*	*Interpretación*
Valores mínimo y máximo	*0 y 4 imperfecciones respectivamente*	*En los rollos analizados, como mínimo no se observaron imperfecciones y como máximo se observaron 4.*
Cuartil 1	*0 imperfecciones*	*En el 25% de los rollos analizados se observaron 0 imperfecciones (o menos)(*) y el 75% restante se observaron 0 o más.*
Cuartil 3	*1 imperfección*	*En el 75% de los rollos analizados se observó 1 imperfección o menos y el 25% restante se observó 1 o más.*
Promedio o media aritmética	*0,84 imperfecciones*	*En promedio, se observaron 0,84 imperfecciones por rollo, en los rollos analizados.*
Media truncada 10%	*0,675 imperfecciones*	*En promedio, se observaron 0,675 imperfecciones por rollo, en los rollos analizados (sin considerar el 10% de rollos con valores mayores y el 10% de rollos con valores menores)*
Media geométrica	*No se puede obtener porque algunos valores observados de la variable son nulos.*	
Moda	*0 imperfecciones*	*La mayoría de los rollos analizados no presentó imperfecciones*
Mediana	*1 imperfección*	*En el 50% de los rollos se observó 1 imperfección o menos y en el 50% restante se observó 1 o más.*
Rango	*4 imperfecciones*	*La máxima diferencia entre dos rollos cualesquiera analizados es de 4 imperfecciones.*
Rango intercuartílico	*1 imperfección*	*La máxima diferencia en el del 50% central de los rollos es de 1 imperfección.*
Variancia	*1,035 (imperfecciones)^2*	*Los desvíos cuadrados entre la cantidad de imperfecciones por rollo y el valor promedio son, en promedio, de 1,035 (imperfecciones)^2 para los rollos analizados.*
Desvío estándar	*1,0174 imperfecciones*	*"En promedio", las cantidad de imperfecciones por rollo desvían en aproximadamente 1,02 imperfecciones de la media aritmética.*
Coeficiente de variación	*121,1%*	*El desvío estándar representa un 121,1% del valor de la media aritmética*
Proporción de rollos sin imperfecciones(**)	*0,46*	*En el 46% de los rollos analizados no se observaron imperfecciones*
Proporción de meses con 3 o más fallas(**)	*0,06*	*El 6% de los meses analizados se produjeron 3 o más fallas*

(*) *Carece de sentido decir 0 imperfecciones o menos pero se incluye como modelo de interpretación general.*
(**) *Se incluyeron estas proporciones a modo de ejemplo; podrían haberse incluido otras.*



Problema 4 (pág. 8):

- *Población: todas las devoluciones.*
- *Muestra: n = 350 devoluciones.*
- *Variable: motivo principal de una devolución, cualitativa, medida en escala nominal.*

Tabla 2.31 *Medidas de resumen para los motivos principales de las 350 devoluciones analizadas*

Medida	***Valor***	***Interpretación***
Moda	*"retraso en la entrega por problemas del transporte"*	*El motivo de devolución más frecuente, entre las analizadas, fue el retraso en la entrega por problemas de transporte.*
Proporción de devoluciones debidas al retraso en la entrega por problemas del transporte(*)	*0,329*	*El retraso por transporte representa el 0,329 (32,9 %) de las devoluciones analizadas.*
Proporción de devoluciones debidas al producto dañado(*)	*0,217*	*El producto dañado representa el 0,217 (21,7 %) de las devoluciones analizadas.*

(*) *Se incluyeron estas proporciones a modo de ejemplo; podrían haberse incluido otras.*



3. Distribuciones de probabilidades para variables aleatorias

3.1 Introducción

En el Capítulo 2, se realiza un análisis descriptivo de un conjunto de datos correspondientes a una variable de interés. Cuando estos datos se refieren a una población finita relevada de manera exhaustiva, la distribución de frecuencias relativas constituye un modelo que representa el comportamiento de la variabilidad de la variable en dicha población. La misma se puede visualizar con un diagrama de bastones o un histograma, según se trate de variables cuantitativas discretas o continuas respectivamente.

Cuando la población es infinita o finita pero de tamaño muy grande, para conocer el modelo de comportamiento de la variable de interés en dicha población, se requeriría seleccionar una muestra de unidades pertenecientes a la misma (ya que no se podría acceder a todas) y estudiar la distribución de las frecuencias relativas de la variable en esa muestra. Cuando el tamaño de la muestra tiende a infinito, por la propiedad de estabilidad o regularidad de las frecuencias relativas, cada frecuencia relativa se estabiliza alrededor de un valor que puede considerarse como la frecuencia relativa poblacional o probabilidad. Surge entonces una distribución de probabilidades que constituye el modelo buscado.

En este capítulo se estudian las distribuciones de probabilidades y se muestra cómo estos modelos permiten obtener probabilidades y otros parámetros de interés y describir, de manera aproximada, el comportamiento de una variable aleatoria en la población.

Los objetivos de este capítulo son:

- Presentar el concepto de variable aleatoria.
- Definir a la probabilidad como frecuencia relativa poblacional.
- Presentar el concepto de distribución de probabilidades para variables aleatorias continuas y discretas.
- Utilizar a las distribuciones de probabilidades (y a las medidas que se puedan obtener a partir de ellas) para describir de manera aproximada el comportamiento de una variable aleatoria en la población.

3.2 Variable aleatoria y su distribución de probabilidades

El proceso de seleccionar al azar una unidad de cierta población y observar en ella alguna característica de interés constituye un experimento aleatorio, es decir, una experiencia que puede repetirse una gran cantidad de veces en igualdad de condiciones y en la cual se verifica que el resultado puede variar de repetición a repetición pero a priori es siempre desconocido [1].



Problema 1 (pág. 7): *En este problema la población está compuesta por todas las barras de cierto tipo que se producen (en condiciones similares) en la empresa metalúrgica. Un experimento aleatorio puede ser seleccionar al azar una barra de la producción y registrar su longitud, en mm.*



Problema 4 (pág. 8): *En este problema la población está compuesta por todas las devoluciones de productos vendidos en línea por la empresa. Un experimento aleatorio consiste en seleccionar al azar una devolución y observar el motivo principal de la misma.*

Dado un experimento aleatorio, se define una variable aleatoria Y que brinda un criterio para asignar un valor numérico al resultado de cada repetición del mismo y, consecuentemente, a cada unidad de la población. Formalmente,

Una **variable aleatoria,** Y, brinda un criterio que permite asociar a cada unidad de la población un valor numérico.



Problema 1 (pág. 7): *Retomando el experimento aleatorio que consiste en seleccionar al azar una barra de la producción y medir su longitud, en mm, se pueden definir, por ejemplo, las siguientes variables aleatorias:*

- *Y_1: cumplimiento de las especificaciones de una barra, en relación a la longitud.*

Y_1 toma el valor 1 si la longitud de la barra está entre 249,4 mm y 250,6 mm (es decir, si la barra cumple con las especificaciones definidas por la empresa automotriz en relación a la longitud) y 0 en caso contrario.

- *Y_2: longitud de una barra.*

Y_2 toma directamente el valor medido de la longitud, sin hacer ninguna modificación.



Problema 4 (pág. 8): *Retomando el experimento aleatorio que consiste en seleccionar al azar una devolución y observar el motivo principal de la misma, se pueden definir, por ejemplo, las siguientes variables aleatorias:*

- *Y_3: devolución por retraso.*

Y_3 toma el valor 1 si el motivo principal de la devolución es el retraso y 0 si el motivo principal es cualquier otro.

[1] La definición y las características de un experimento aleatorio se amplían en el Capítulo 4

- *Y_4: devolución por problemas administrativos.*

Y_4 toma el valor 1 si el motivo principal de la devolución es un error de facturación o un precio no coincidente y 0 si el motivo principal es cualquier otro.

Definida una variable aleatoria Y, su recorrido, R_Y, es el conjunto de valores que puede asumir dicha variable. Si R_Y es un intervalo o una unión de intervalos de números reales, Y es una **variable aleatoria continua**; en cambio, si R_Y es un conjunto finito o infinito numerable, Y es una **variable aleatoria discreta**.



Problema 1 (pág. 7):

- *Y_1: cumplimiento de las especificaciones en relación a la longitud.*

Recorrido de la variable Y_1: $R_{Y1} = \{0,1\}$

- *Y_2: longitud.*

Recorrido de la variable Y_2: $R_{Y2} = \Re^+$

Y_1 es una variable aleatoria discreta e Y_2 es una variable aleatoria continua.



Problema 4 (pág. 8):

- *Y_3: devolución por retraso.*
- *Y_4: devolución por problemas administrativos.*

Para Y_3 e Y_4, sus recorridos coinciden con el de Y_1 . Ambas son variables aleatorias discretas.

Si la variable aleatoria Y es continua, su distribución de probabilidades se representa de manera aproximada por la función de densidad de probabilidad mientras que, si es discreta, se representa por la función de probabilidad puntual.

3.2.1 Variable aleatoria continua. Función de densidad de probabilidad

Considere a modo de ejemplo la situación del Problema 1 relativa a un proceso de fabricación de barras de acero, respecto del cual interesa estudiar el comportamiento de la característica *Y: longitud de una barra de acero (en mm).* La variable aleatoria Y asocia a cada barra el valor numérico de su longitud y por lo tanto, es una variable aleatoria continua. Al tratarse de una producción de barras, se piensa en las barras que fueron, son y serán fabricadas bajo las mismas condiciones y por ese motivo, esta población se puede considerar infinita. Si se tuviesen los valores de Y sobre todas las barras del proceso, se contaría con la población estadística.

Suponga que se toma una muestra de $n = 20$ barras de la población y se representa la variabilidad de esas 20 longitudes mediante un histograma con el correspondiente polígono de frecuencias. El reducido tamaño de la muestra obligará a definir intervalos de amplitud grande teniendo algunos de ellos escasa frecuencia y, como consecuencia, el polígono va a estar descripto por picos muy pronunciados. Si la muestra fuese de $n = 200$ barras, el histograma resultante podría construirse con intervalos de menor amplitud y como resultado el polígono de frecuencias tendría picos menos pronunciados. Si se considerara una muestra de 2000 barras, podrían utilizarse intervalos de amplitud todavía menor y el polígono de frecuencias tendría picos menos pronunciados, más suaves. Las formas de los histogramas para las muestras de tamaño 20, 200 y 2000 podrían ser las que se muestran en la Figura 3.1, respectivamente.

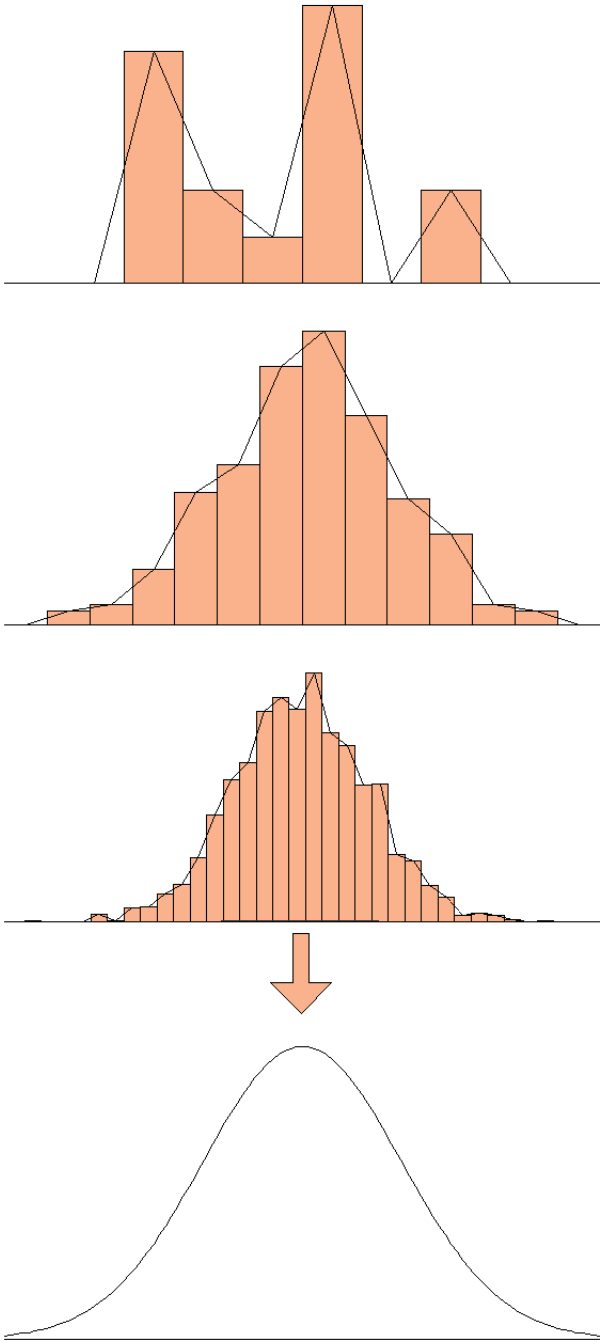


Figura 3.1 *Histogramas para determinadas muestras de tamaño 20, 200 y 2000 unidades y función f (de arriba a abajo respectivamente)*

Si se pudiera seguir aumentando el tamaño de muestra indefinidamente, el polígono de frecuencias que se obtendría sería, seguramente, una curva suave como la gráfica de la función f representada en último lugar en la figura mencionada. La propiedad que se acaba de describir se conoce como **propiedad de estabilidad o regularidad de las frecuencias relativas en los grandes números**; se trata de una propiedad que solo emerge después de muchas observaciones. La función f es el límite del polígono de frecuencias cuando el tamaño de muestra tiende a infinito y brinda una descripción aproximada del comportamiento de la variable en la población.

Por conveniencia para el desarrollo que sigue, se consideran histogramas donde el área de la barra asociada a cada intervalo de clase es igual a la frecuencia relativa de la misma [2] y en consecuencia el área total del histograma vale 1.

La función f (suele simbolizarse también f_Y) recibe el nombre de **función de densidad de probabilidad** de Y. Es una representación matemática (o modelo matemático[3]) del comportamiento de una variable aleatoria continua Y en la población.

[3] Un modelo es una representación de un fenómeno donde eventualmente se realizan simplificaciones, pero se respetan las características sobresalientes del mismo. Usando las palabras de George Box, "todos los modelos son falsos, pero algunos son útiles"

Este modelo matemático da una idea aproximada de la variabilidad de los valores de la variable aleatoria continua Y en la población, ayuda a comprender qué valores de la variable son posibles y con qué frecuencia relativa aparecen dichos valores en la población, facilitando así la toma de decisiones.

Conocer la función de densidad de probabilidad f es útil para hacer previsiones sobre todas las unidades de la población. No se puede afirmar que la variabilidad de la variable Y en la población sea exactamente la implicada por f sino que dicha función describe en forma “razonablemente ajustada” el comportamiento de dicha variabilidad.

Una función de densidad de probabilidad cumple con las siguientes condiciones que se derivan de su relación con los polígonos de frecuencias:

- $f_Y(y) \geq 0, \forall\, y$
- $\int_{-\infty}^{\infty} f_Y(y)dy = 1$ (propiedad conocida como **condición de cierre**: establece que el área del recinto limitado por la gráfica de f y el eje de las abscisas vale 1).

La proporción de unidades de la población con valores de la variable continua en un intervalo $[y_1, y_2]$ está dada por el área del recinto limitado por la curva de densidad sobre dicho intervalo y el eje de abscisas, es decir, por el valor de la integral de la función de densidad sobre dicho intervalo. Esta frecuencia relativa límite o proporción poblacional recibe el nombre de **probabilidad**.

Se puede decir, entonces, que la probabilidad (P) de que la variable aleatoria Y tome valores en el intervalo $[y_1, y_2]$ está dada por el valor de la integral de la función de densidad de probabilidad en dicho intervalo. En símbolos: $P(y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} f_Y(y)dy$. En la Figura 3.2 se indica el área que representa esta probabilidad bajo la función de densidad de probabilidad de Y.

Teniendo en cuenta lo explicado, la probabilidad se puede interpretar como *la proporción de unidades de la población para las cuales la característica Y toma valores entre* y_1 *e* y_2.

Además, según se definió anteriormente, $P(Y = y) = P(Y \in \{y\}) = P(y \leq Y \leq y) = \int_{y}^{y} f_Y(s)ds = 0$. Es decir que si bien idealmente la variable Y podría tomar el valor y, la probabilidad de que eso ocurra es nula.

De lo anterior se deduce que $P(y_1 \leq Y < y_2) = P(y_1 < Y \leq y_2) = P(y_1 < Y < y_2) = P(y_1 \leq Y \leq y_2)$.

[2] Para que el área de cada barra sea igual a la frecuencia relativa de la clase a la que está asociada, la altura de la barra debe ser igual al cociente entre la frecuencia relativa que representa la barra y el ancho de la misma. Dicho cociente recibe el nombre de densidad de la frecuencia relativa en dicha clase.

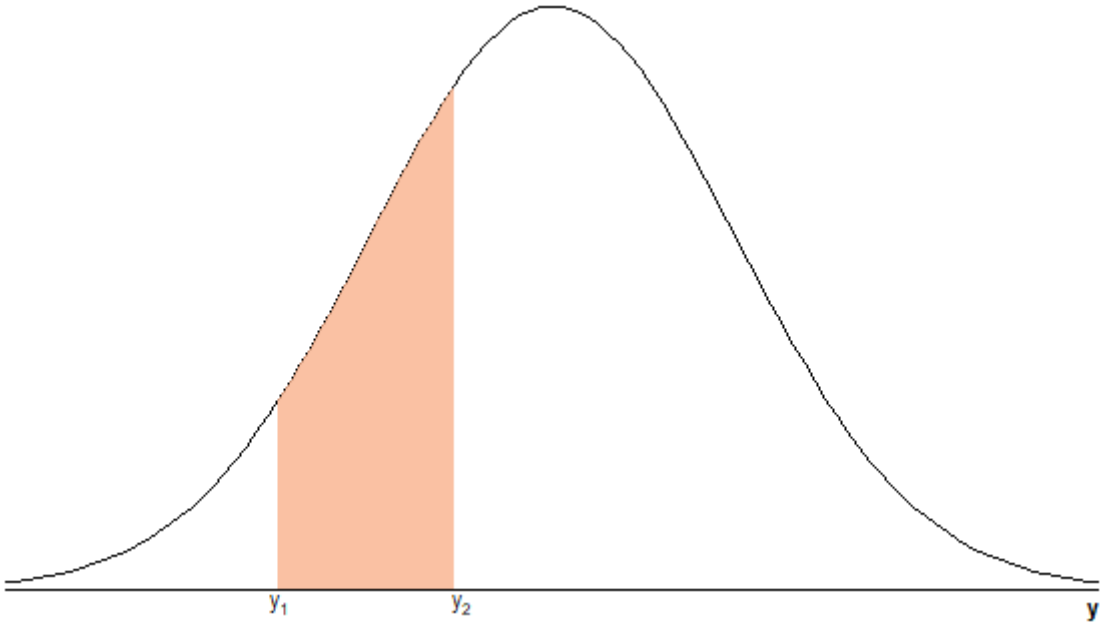


Figura 3.2 *Función de densidad de probabilidad de Y. El área de la región en naranja indica la probabilidad de que Y tome valores entre y_1 e y_2*

Problema 1 (pág. 7): *En la gerencia de la empresa metalúrgica están interesados en convertirse en proveedores de la automotriz y quieren saber cuál es la proporción de barras que cumplen con el requerimiento impuesto para las longitudes.*
Suponga que se conoce que el comportamiento de las longitudes de las barras se puede describir mediante la función:

$$f_Y(y) = \begin{cases} (y-249) & si \;\; 249 < y \leq 250 \\ (251-y) & si \;\; 250 < y \leq 251 \\ 0 & en\ otro\ caso. \end{cases}$$

La proporción de barras con longitud entre 249,4 mm y 250,6 mm se calcula de la siguiente forma:

$P(249,4 \leq Y \leq 250,6) = P(249,4 \leq Y \leq 250) + P(250 < Y \leq 250,6) =$

$= \int_{249,4}^{250}(y-249)dy + \int_{250}^{250,6}(251-y)dy =$

$= (\frac{y^2}{2} - 249y)|_{249,4}^{250} + (251y - \frac{y^2}{2})|_{250}^{250,6} =$

$= [(\frac{250^2}{2} - 249 \cdot 250) - (\frac{249,4^2}{2} - 249 \cdot 249,4)] + [(251 \cdot 250,6 - \frac{250,6^2}{2}) - (251 \cdot 250 - \frac{250^2}{2})] =$

$= [31000 - (-31000,42)] + [31500,42 - 31500] = 0,42 + 0,42 = 0,84.$

En la Figura 3.3 se representa la función de densidad de probabilidad de Y y la probabilidad calculada.

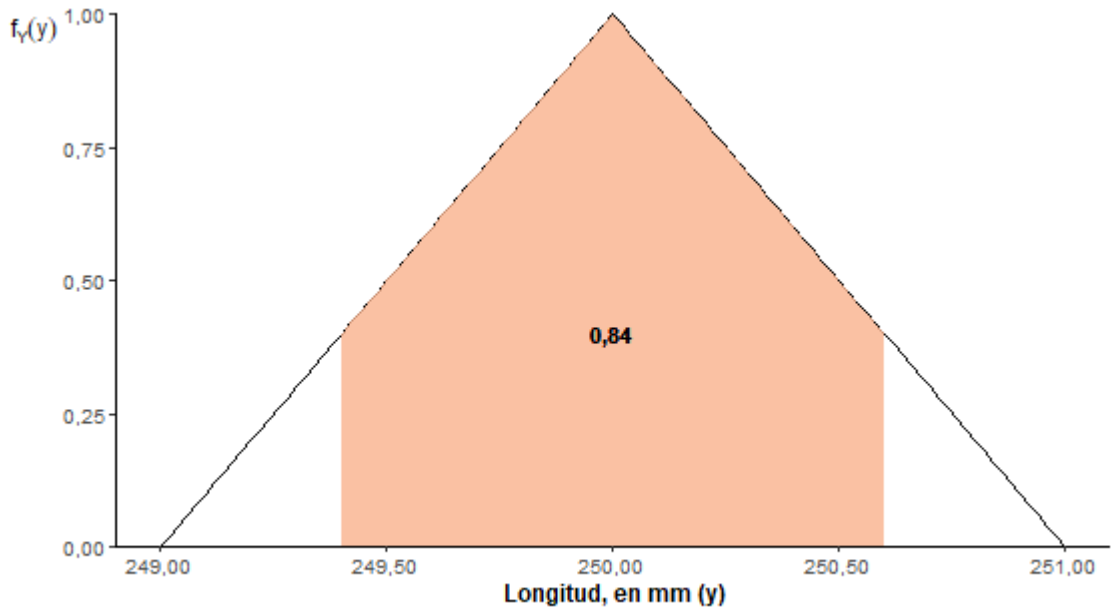


Figura 3.3 *Función de densidad de probabilidad de Y, longitud de las barras. El área de la región en naranja indica la probabilidad de que Y tome valores entre 249,4 mm y 250,6 mm*



Con lo visto hasta ahora, se puede resolver la Actividad 1, Sección 3.5.

3.2.2 Variable aleatoria discreta. Función de probabilidad puntual

A modo de ejemplo, en el Problema 5 se desea evaluar el comportamiento de *Y: cantidad de imperfecciones de un rollo de 100 m de cable unipolar de 2,5 mm*.

La variable aleatoria Y asocia a cada rollo el valor de la cantidad de imperfecciones que este posee. Se trata, en este caso, de una variable aleatoria discreta.

Análogamente al caso continuo, a medida que aumenta el tamaño de la muestra, las distribuciones de frecuencias relativas convergen a una distribución límite, la **distribución de probabilidades**.

La función p (suele simbolizarse también p_Y) recibe el nombre de **función de probabilidad puntual** de Y. Es una representación matemática (o modelo matemático) del comportamiento de una variable aleatoria discreta Y en la población.

La función de probabilidad puntual verifica las siguientes propiedades que se derivan de su relación con las frecuencias relativas muestrales:

- $p_Y(y) \geq 0 \; \forall \, y$
- $\sum_{y \in R_Y} p_Y(y) = 1$ (propiedad conocida como **condición de cierre**: establece que la suma de las probabilidades puntuales asociadas a todos los valores de una variable aleatoria discreta es uno).

La proporción de unidades de la población con valores de la variable discreta en un intervalo cualquiera $[y_1, y_2]$ está dada por la suma de las probabilidades puntuales asociadas a los valores de la variable

pertenecientes a en dicho intervalo. Esta frecuencia relativa límite o proporción poblacional recibe el nombre de **probabilidad**.

Se puede decir, entonces, que la probabilidad (P) de que la variable aleatoria Y tome valores en el intervalo $[y_1, y_2]$ está dada por la sumatoria de la función de probabilidad puntual de los valores de la variable pertenecientes a dicho intervalo. En símbolos: $P(y_1 \leq Y \leq y_2) = \sum_{y=y_1}^{y_2} p_Y(y)$.

La representación gráfica más conveniente de la distribución de probabilidades de una variable aleatoria discreta es, como en el caso de las distribuciones de frecuencias, un gráfico de bastones. En la Figura 3.4 se indican los bastones que representan esta probabilidad en la función de probabilidad puntual de Y.

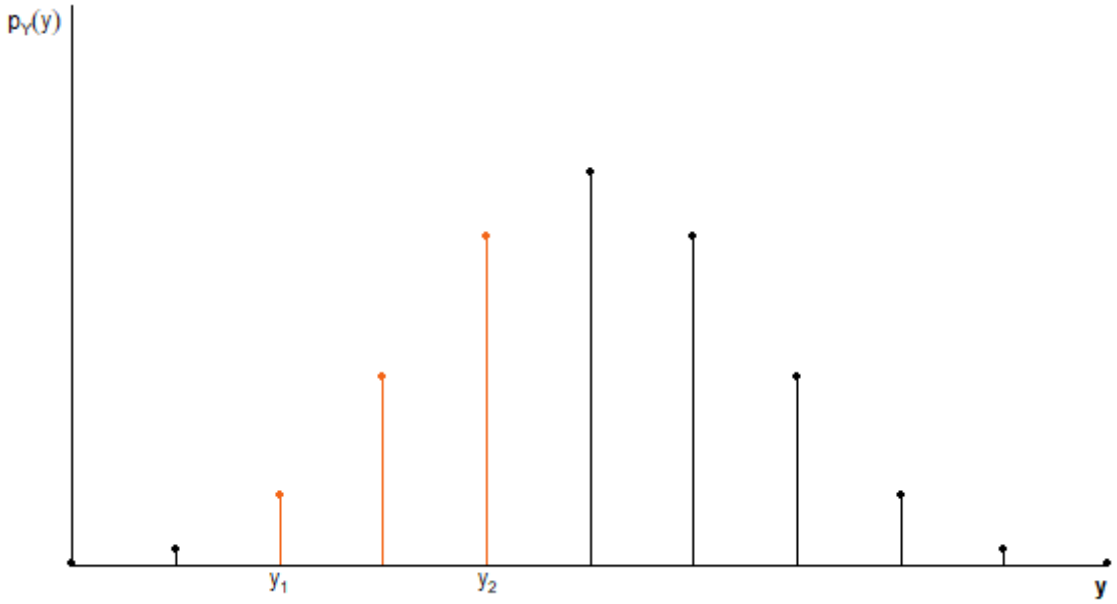


Figura 3.4 *Función de probabilidad puntual de Y. La suma de las alturas de los bastones en naranja indica la probabilidad de que Y tome valores entre y_1 e y_2, ambos inclusive*

La valores de la función de probabilidad puntual para cada valor del recorrido de una variable aleatoria discreta se representan por un punto. Los bastones se trazan para visualizar mejor el comportamiento de dicha variable en la población.

Teniendo en cuenta lo explicado, $P(y_1 \leq Y \leq y_2)$ se puede interpretar como *la proporción de unidades de la población en las cuales la característica Y toma valores entre y_1 e y_2, ambos inclusive*.

Note que $P(y_1 \leq Y < y_2)$, $P(y_1 < Y \leq y_2)$, $P(y_1 < Y < y_2)$ y $P(y_1 \leq Y \leq y_2)$ no son siempre iguales ya que $P(Y = y_1)$ y $P(Y = y_2)$ pueden ser distintas de 0.

La función de probabilidad puntual, p_Y, de una variable aleatoria discreta Y indica, para cada y, la probabilidad de que Y asuma el valor y. Luego, $p_Y(y) = P(Y = y)$.

Problema 5 (pág. 8): *Un cliente, interesado en comprar rollos de 100 m de cable unipolar de 2,5 mm, pretende que la proporción de rollos con más de dos imperfecciones sea la mínima posible y admite, como máximo, que esta sea 0,05. Suponga que se conoce que el comportamiento de las cantidades de imperfecciones en rollos de este tipo se puede describir mediante la función:*

$$p_Y(y) = \begin{cases} 0,7 & si\ y = 0 \\ 0,18 & si\ y = 1 \\ 0,09 & si\ y = 2 \\ 0,03 & si\ y = 3 \\ 0 & en\ otro\ caso. \end{cases}$$

La proporción de rollos con más de 2 imperfecciones se calcula de la siguiente forma:

$P(Y > 2) = P(Y = 3) = 0,03$

Debido a que la proporción calculada es menor a 0,05, en principio no habría inconvenientes en abastecer a este cliente.
En la Figura 3.5 se representa la función de probabilidad puntual de Y y la probabilidad calculada.

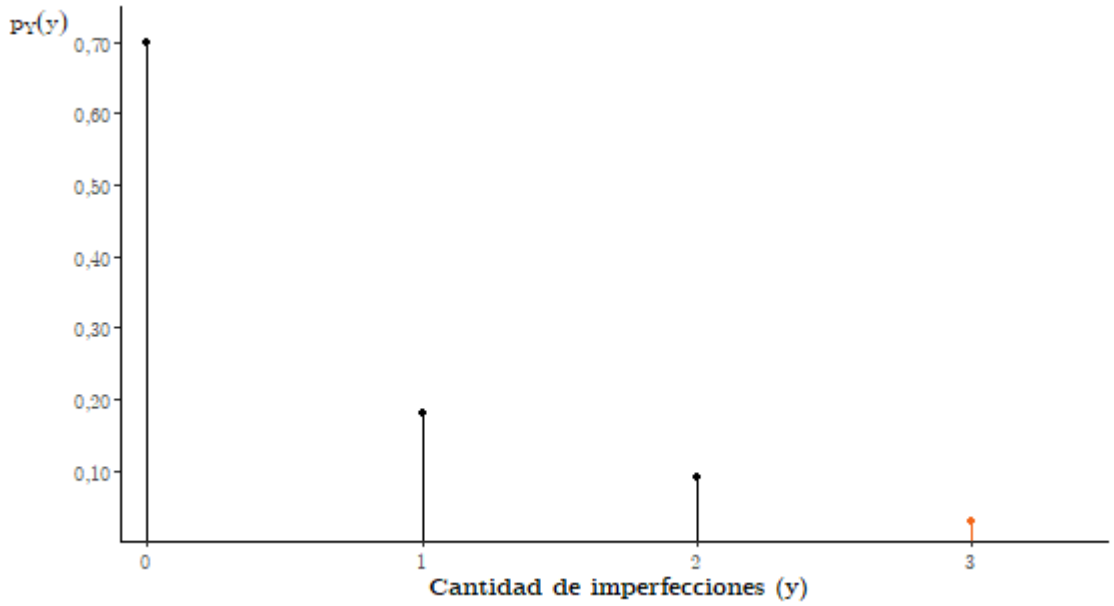


Figura 3.5 *Función de probabilidad puntual de Y, cantidad de imperfecciones por rollo. La altura del bastón en naranja indica la probabilidad de que Y sea mayor a 2*

Con lo visto hasta ahora, se puede resolver la Actividad 2, Sección 3.5.

3.2.3 Función de distribución acumulada

Para una variable aleatoria Y, tanto discreta como continua, la **función de distribución acumulada** F_Y se define de la siguiente manera: $F_Y(y) = P(Y \leq y)\ \ \forall\, y \in \mathbb{R}$.

Esta función especifica para cada valor y, la probabilidad de que la variable Y tome valores menores o iguales que él. Dependiendo de si la variable aleatoria es continua o discreta, F_Y se puede obtener a partir de su distribución de probabilidades de la siguiente forma:

Función de distribución acumulada	Variable continua	Variable discreta
$F_Y(y)$	$\int_{-\infty}^{y} f_Y(s)ds$	$\sum_{s\leq y} p_Y(s)$

La función F_Y goza de las siguientes propiedades:

- es monótona no decreciente en y;
- su imagen se encuentra en el intervalo [0,1].

Debido a que las funciones de densidad de probabilidad y de probabilidad puntual mantienen una relación con la función de distribución acumulada, ambas se pueden obtener a partir de esta de la siguiente manera:

- cuando Y es una variable continua: $f_Y(y) = \frac{d}{dy}F_Y(y)$ para todo valor y en el cual F sea derivable;
- cuando Y es una variable discreta: $p_Y(y_i) = F_Y(y_i) - F_Y(y_{i-1})$ para cada i.

Definida F_Y, cualquier probabilidad se puede expresar en función de la misma. Por ejemplo:

- $P(Y \leq y) = F_Y(y)$
- $P(Y > y) = 1 - P(Y \leq y) = 1 - F_Y(y)$
- $P(y_1 < Y \leq y_2) = P(Y \leq y_2) - P(Y \leq y_1) = F_Y(y_2) - F_Y(y_1)$



Problema 1 (pág. 7): *Suponiendo que la función de densidad de probabilidad de la variable aleatoria Y:* ***longitud*** *es la mencionada en la página 148, se puede calcular su función de distribución acumulada, F_Y, como:*

- *si* $y \leq 249 : F_Y(y) = 0$;
- *si* $249 < y \leq 250 : F_Y(y) = \int_{249}^{y}(s-249)ds =$

$= (\frac{s^2}{2} - 249s)|_{249}^{y} =$

$= (\frac{y^2}{2} - 249y) - (\frac{249^2}{2} - 249^2) = \frac{y^2}{2} - 249y + \frac{249^2}{2}$;

- *si* $250 < y \leq 251 : F_Y(y) = \int_{249}^{250}(s-249)ds + \int_{250}^{y}(251-s)ds =$

$= (\frac{s^2}{2} - 249s)|_{249}^{250} + (251s - \frac{s^2}{2})|_{250}^{y} =$

$= (\frac{250^2}{2} - 249 \cdot 250) - (\frac{249^2}{2} - 249^2) + (251y - \frac{y^2}{2}) - (251 \cdot 250 - \frac{250^2}{2}) =$

$= \frac{250^2}{2} - 249 \cdot 250 - \frac{249^2}{2} + 249^2 + 251y - \frac{y^2}{2} - 251 \cdot 250 + \frac{250^2}{2} =$

$= -\frac{y^2}{2} + 251y - \frac{62999}{2}$;

- *si* $y > 251 : F_Y(y) = 1$.

En forma resumida resulta:

$$F_Y(y) = \begin{cases} 0 & si\ \ y \leq 249 \\ \frac{y^2}{2} - 249\,y + \frac{249^2}{2} & si\ \ 249 < y \leq 250 \\ -\frac{y^2}{2} + 251\,y - \frac{62999}{2} & si\ \ 250 < y \leq 251 \\ 1 & si\ \ y > 251. \end{cases}$$

En la Figura 3.6 se puede observar esta función de Y.

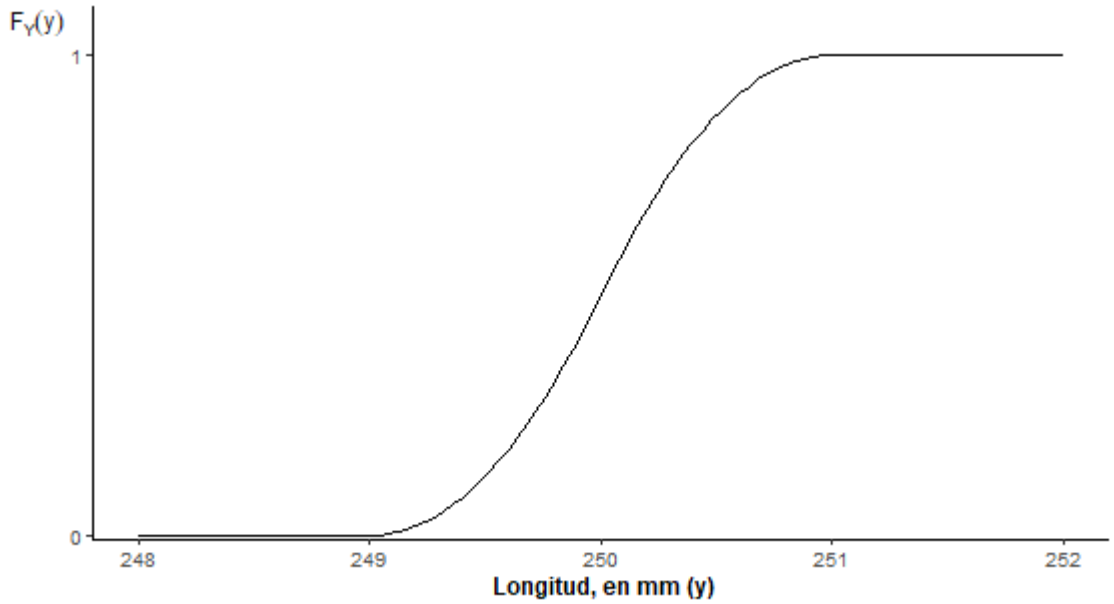


Figura 3.6 *Función de distribución acumulada de Y, longitud de las barras*

Note que en la función de densidad de probabilidad, $P(Y \leq y)$ está representada por una área mientras que, en la función de distribución acumulada, esa probabilidad está representada por la ordenada de un punto. A modo de ejemplo, en la Figura 3.7 se representan $P(Y \leq 250)$ en las funciones de densidad de probabilidad y de distribución acumulada.

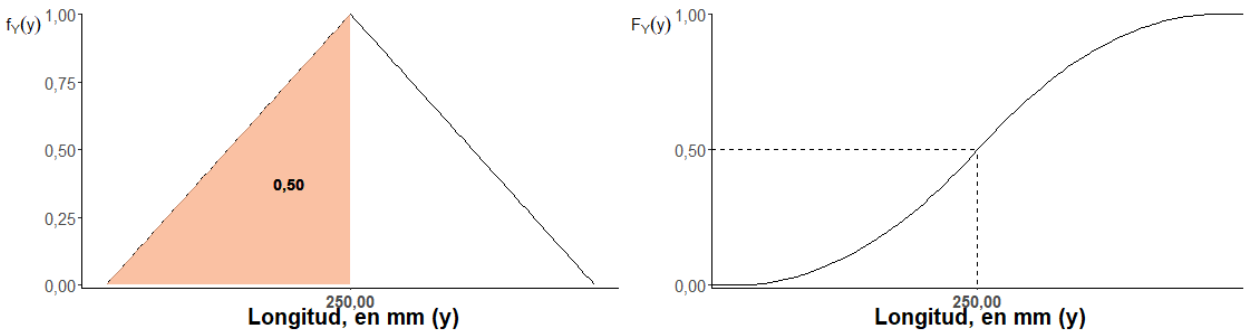


Figura 3.7 *Probabilidad de que la longitud, Y, sea a lo sumo 250 mm*

La probabilidad de que la longitud de una barra se encuentre entre 249,4 mm y 250,6 mm se expresa a través de la función de distribución hallada de la siguiente manera:

$$P(249,4 \leq Y \leq 250,6) = P(Y \leq 250,6) - P(Y < 249,4) = P(Y \leq 250,6) - P(Y \leq 249,4) =$$
$$= F_Y(250,6) - F_Y(249,4) =$$
$$= (-\frac{250,6^2}{2} + 251 \cdot 250,6 - \frac{62999}{2}) - (\frac{249,4^2}{2} - 249 \cdot 249,4 + \frac{249^2}{2}) = 0,92 - 0,08 = 0,84.$$

Problema 5 (pág. 8): *Suponiendo que la función de probabilidad puntual de la variable aleatoria Y* ***: cantidad de imperfecciones por rollo*** *es la mencionada anteriormente se puede calcular su función de distribución acumulada como:*

$$F_Y(y) = \sum_{s \leq y} p_Y(s) \begin{cases} 0 & si\ y < 0 \\ 0,7 & si\ 0 \leq y < 1 \\ 0,88 & si\ 1 \leq y < 2 \\ 0,97 & si\ 2 \leq y < 3 \\ 1 & si\ y \geq 3 \end{cases}$$

En la Figura 3.8 se puede observar esta función de Y.

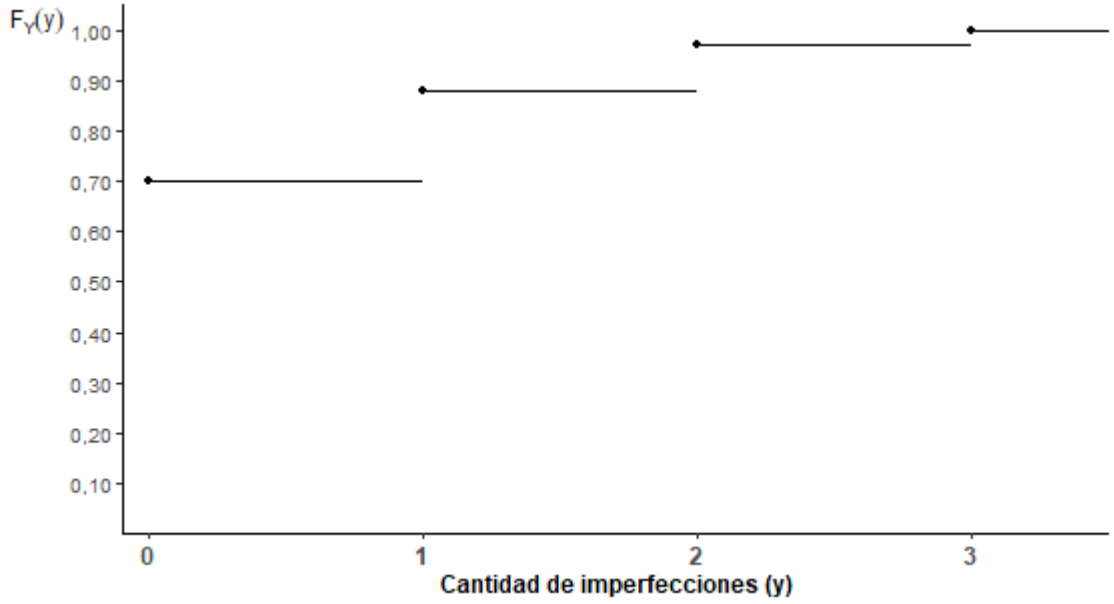


Figura 3.8 *Función de distribución acumulada de Y, cantidad de imperfecciones por rollo*

Note que en la función de probabilidad puntual, $P(Y \leq y)$ está representada por la suma de las alturas de los bastones, mientras que en la función de distribución acumulada, esa probabilidad está representada por la ordenada de un punto. A modo de ejemplo, en la Figura 3.9 se presenta $P(Y \leq 1)$ en las funciones de probabilidad puntual y de distribución acumulada.

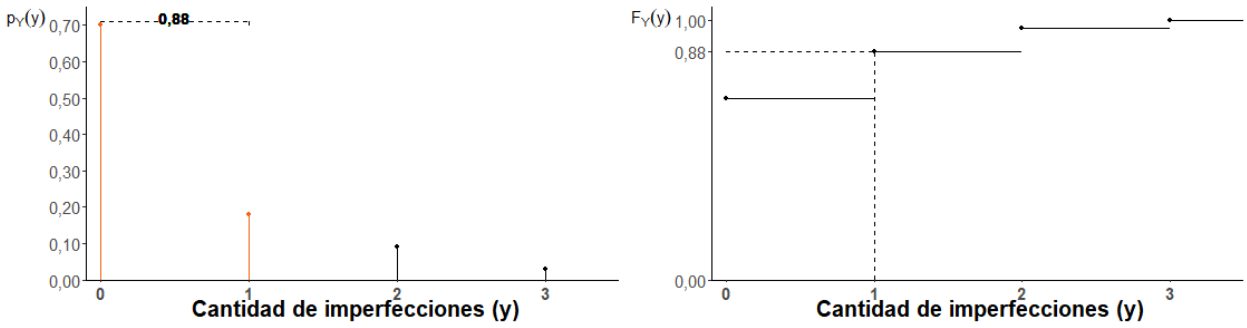


Figura 3.9 *Probabilidad de que la cantidad de imperfecciones por rollo, Y, sea a lo sumo 1*

La probabilidad de que la cantidad de imperfecciones de un rollo sea mayor a 2 se expresa a través de la función de distribución hallada de la siguiente manera:

$P(Y > 2) = 1 - P(Y \leq 2) = 1 - F_Y(2) = 1 - 0,97 = 0,03.$

Con lo visto hasta ahora, se pueden resolver las Actividades 3 a 6, Sección 3.5.

3.3 Parámetros de una variable aleatoria

Dado que las distribuciones de probabilidades describen el comportamiento de una variable aleatoria en la población, cualquier medida de resumen que se obtenga a partir de ellas se denomina **parámetro**.

En esta sección se presentan medidas que indican la localización o posición y la dispersión o variabilidad de los valores de una variable aleatoria, obtenidas a partir de su distribución de probabilidades.

Algunos de los parámetros más utilizados que caracterizan a esa distribución son la media y el desvío estándar. También pueden interesar la moda y diferentes percentiles. Estas y otras medidas, se pueden calcular a partir de la distribución de probabilidades de la variable.
A continuación, se muestra cómo obtener estos parámetros a partir de la función de densidad de probabilidad o de probabilidad puntual según corresponda.

3.3.1 Parámetros de localización o posición

En el Capítulo 2 se presenta la siguiente expresión para la media muestral: $\bar{y} = \sum_{j=1}^{k} y_j f_j$ (página 73). En el presente capítulo se vio que, si el tamaño de muestra (n) tiende a infinito, la frecuencia relativa muestral tiende a la frecuencia relativa poblacional o probabilidad y se tiene entonces la expresión de la media poblacional para el caso discreto. Para el caso continuo, se define de manera análoga considerando la función de densidad de probabilidad.

Se denomina **media poblacional** o **esperanza matemática** de una variable aleatoria Y, y se simboliza con $E(Y)$ o μ_Y indistintamente, al valor que se obtiene de la siguiente manera:

- cuando Y es continua, $E(Y) = \mu_Y = \int_{y \in R_Y} y\, f_Y(y) dy$;
- cuando Y es discreta, $E(Y) = \mu_Y = \sum_{y \in R_Y} y\, p_Y(y)$.

En el Capítulo 2 se define a la mediana muestral como el valor de la variable que acumula el 50% de las observaciones ordenadas. Si se piensa en la mediana poblacional, entonces, inmediatamente se la asocia con el valor de la variable que acumula una probabilidad igual a 0,50, es decir, con el valor de la variable (y^*) tal que $F_Y(y*) = 0,50$. Sin embargo, es posible que este valor no exista, es decir que no se verifique la igualdad mencionada para ningún y* perteneciente al recorrido de Y, especialmente en el caso de variables aleatorias discretas. Por ese motivo, se busca el menor valor de la variable que acumula como mínimo una probabilidad de 0,50 y consecuentemente se define a la mediana poblacional.

Se denomina **mediana poblacional** de una variable aleatoria Y, y se simboliza $\widetilde{\mu}_Y$, al valor que se obtiene de la siguiente manera:
$\widetilde{\mu}_Y = \text{mín}\{y : F_Y(y) \geq 0,50\}$

- Cuando Y es una variable aleatoria continua, $\widetilde{\mu}_Y = \text{mín}\{y : \int_{s \leq y} f_Y(s)ds \geq 0,50\}$
- Cuando Y es una variable aleatoria discreta, $\widetilde{\mu}_Y = \text{mín}\{y : \sum_{s \leq y} p_Y(s) \geq 0,50\}$

En forma análoga al cálculo de la mediana se pueden obtener otros percentiles, como por ejemplo, el **cuartil 1** (Q_{1Y}) y el **cuartil 3** (Q_{3Y}), que se definen a continuación:

- Se denomina **primer cuartil poblacional** de una variable aleatoria Y, y se simboliza Q_{1Y} , al valor que se obtiene de la siguiente manera:
  $Q_{1Y} = \text{mín}\{y : F_Y(y) \geq 0,25\}$.
  - Cuando Y es una variable aleatoria continua, $Q_{1Y} = \text{mín}\{y : \int_{s \leq y} f_Y(s)ds \geq 0,25\}$.
  - Cuando Y es una variable aleatoria discreta, $Q_{1Y} = \text{mín}\{y : \sum_{s \leq y} p_Y(s) \geq 0,25\}$.
- Se denomina **tercer cuartil poblacional** de una variable aleatoria Y, y se simboliza Q_{3Y} , al valor que se obtiene de la siguiente manera:
  $Q_{3Y} = \text{mín}\{y : F_Y(y) \geq 0,75\}$.
  - Cuando Y es una variable aleatoria continua, $Q_{3Y} = \text{mín}\{y : \int_{s \leq y} f_Y(s)ds \geq 0,75\}$.
  - Cuando Y es una variable aleatoria discreta, $Q_{3Y} = \text{mín}\{y : \sum_{s \leq y} p_Y(s) \geq 0,75\}$.

En el Capítulo 2 se define a la moda muestral como el valor de la variable que ocurre con mayor frecuencia. En la población, la moda se puede pensar de forma similar.

Se denomina **moda poblacional** de una variable aleatoria Y, y se simboliza Mo[4], al valor donde la función de densidad de probabilidad en el caso continuo, o la función de probabilidad puntual en el caso discreto, asume su valor máximo.

[4] Como ya se expresó en el Capítulo 2, no existe un criterio unificado para simbolizar a la moda. En este texto se va a utilizar *mo* en el caso de muestras y *Mo* en el caso poblacional.

La moda no necesariamente es única, puede haber más de un máximo local dando lugar a más de una moda. Por ejemplo, si Y tiene dos modas, se dice que tiene una distribución de probabilidades *bimodal*.

3.3.2 Parámetros de dispersión o variabilidad

Para dar una idea de la variabilidad de Y se sugiere que una medida de posición se acompañe con una medida de dispersión que esté relacionada con la misma. La media poblacional se acompaña, en general, con el desvío estándar poblacional y la mediana, con el rango intercuartílico. El desvío estándar indica la dispersión de los valores de Y alrededor de su media.

El **desvío estándar poblacional** de la variable aleatoria Y se simboliza con $D(Y)$ o σ_Y.

Este parámetro es la raíz cuadrada positiva de la variancia, es decir, de la esperanza matemática de los desvíos con respecto a la media poblacional elevados al cuadrado:

- cuando Y es continua, $D(Y) = \sigma_Y = \sqrt{V(Y)} = \sqrt{E(Y-\mu_Y)^2} = \sqrt{\int_{y\in R_Y}(y-\mu_Y)^2 f_Y(y)dy} =$
  $= \sqrt{\int_{y\in R_Y} y^2 f_Y(y)dy - \mu_Y^2} = \sqrt{E(Y^2)-\mu_Y^2}$;
- cuando Y es discreta, $D(Y) = \sigma_Y = \sqrt{V(Y)} = \sqrt{E(Y-\mu_Y)^2} = \sqrt{\sum_{y\in R_Y}(y-\mu_Y)^2 p_Y(y)} =$
  $= \sqrt{\sum_{y\in R_Y} y^2 p_Y(y) - \mu_Y^2} = \sqrt{E(Y^2)-\mu_Y^2}$.

Observe que μ_Y^2 es el valor del promedio poblacional de la variable aleatoria Y, elevado al cuadrado; mientras que $E(Y^2)$ es el promedio poblacional del cuadrado de la variable aleatoria Y.
En el Apéndice (Sección 3.7.1) se obtiene la expresión de la variancia poblacional, $V(Y)$, en función de $E(Y)$ (o μ_Y) y de $E(Y^2)$.

El **rango intercuartílico poblacional** de la variable aleatoria Y se simboliza $RIC(Y)$ y se obtiene haciendo la diferencia $Q_{3Y} - Q_{1Y}$.

Problema 1 (pág. 7): *Suponiendo que la función de densidad de probabilidad de la variable aleatoria Y:* ***longitud****, es la mencionada en la página 148 se calculan a continuación algunas medidas de interés.*

$E(Y) = \int_{249}^{251} y\, f_Y(y)dy = \int_{249}^{250} y\,(y-249)dy + \int_{250}^{251} y\,(251-y)dy =$

$= (\frac{y^3}{3} - \frac{249y^2}{2})|_{249}^{250} + (\frac{251y^2}{2} - \frac{y^3}{3})|_{250}^{251} = 124,833 + 125,167 = 250.$

$Q_{1Y} = \min\{y : F_Y(y) \geq 0,25\} = 249 + \frac{1}{\sqrt{2}} \approx 249,7$, *ya que* $\int_{249}^{249+\frac{1}{\sqrt{2}}} (y-249)dy = 0,25.$

$\widetilde{\mu}_Y = \min\{y : F_Y(y) \geq 0,50\} = 250$, *ya que* $\int_{249}^{250}(y-249)dy = 0,5.$

$Q_{3Y} = \min\{y : F_Y(y) \geq 0,75\} = 250 + \frac{1}{\sqrt{2}} \approx 250,7$, *ya que* $= \int_{249}^{250}(y-249)dy + \int_{250}^{250+\frac{1}{\sqrt{2}}} (251-y)dy = 0,75.$

$Mo = 250$ *ya que la función de densidad presenta su máximo en ese punto.*

$D(Y) = \sqrt{\int_{249}^{251} y^2 f_Y(y)dy - \mu_Y^2} = \sqrt{[\int_{249}^{250} y^2\,(y-249)dy + \int_{250}^{251} y^2\,(251-y)dy] - \mu_Y^2} =$

$= \sqrt{[(\frac{y^4}{4} - \frac{249y^3}{3})|_{249}^{250} + (\frac{251y^3}{3} - \frac{y^4}{4})|_{250}^{251}] - 250^2} = \sqrt{\frac{1}{6}} \approx 0,41.$

$RIC(Y) = Q_{3Y} - Q_{1Y} = \sqrt{2} \approx 1,4.$

A partir de los resultados obtenidos, la distribución de las longitudes de las barras puede describirse del siguiente modo.
La longitud media es de 250 mm, valor que coincide con la mediana, de modo que el 50 % de las barras presenta longitudes menores o iguales a 250 mm, mientras que el 50 % restante presenta longitudes mayores o iguales a ese valor.
El 25 % de las barras tiene longitudes menores o iguales a aproximadamente 249,7 mm, mientras que el 75 % no supera aproximadamente los 250,7 mm.
En cuanto a la variabilidad, el rango intercuartílico es de aproximadamente 1,4 mm, lo que indica la máxima diferencia en el 50 % central de las longitudes, mientras que el desvío estándar es de aproximadamente 0,41 mm.
Finalmente, la distribución presenta su mayor concentración de densidad en torno a 250 mm, valor que coincide con el punto donde la función de densidad alcanza su máximo, lo cual es consistente con una distribución simétrica centrada en ese valor.



Problema 5 (pág. 8): *Suponiendo que la función de probabilidad puntual de la variable aleatoria* Y***: cantidad de imperfecciones por rollo****, es la mencionada en la página 151 se puede calcular su media y su desvío estándar como:*

$$E(Y) = \sum_{y=0}^{3} y\, p_Y(y) =$$

$$= 0 \cdot 0,7 + 1 \cdot 0,18 + 2 \cdot 0,09 + 3 \cdot 0,03 = 0,45$$

$Q_{1Y} = \text{mín}\{y : F_Y(y) \geq 0,25\}.$
Como $F_Y(0) = 0,70 \geq 0,25$*, se tiene que* $Q_{1Y} = 0.$

$\widetilde{\mu}_Y = \text{mín}\{y : F_Y(y) \geq 0,50\}.$
Como $F_Y(0) = 0,70 \geq 0,50$*, se tiene que* $\widetilde{\mu}_Y = 0.$

$Q_{3Y} = \text{mín}\{y : F_Y(y) \geq 0,75\}.$
Como $F_Y(1) = 0,88 \geq 0,75$*, se tiene que* $Q_{3Y} = 1.$

$Mo = 0$*, ya que es el valor para el que se obtiene la mayor probabilidad.*

$$D(Y) = \sigma_Y = \sqrt{\sum_{y=0}^{3} y^2\, p_Y(y) - \mu_Y^2} =$$

$$= \sqrt{0^2 \cdot 0,7 + 1^2 \cdot 0,18 + 2^2 \cdot 0,09 + 3^2 \cdot 0,03 - 0,45^2} =$$
$$= \sqrt{0,81 - 0,2025} = 0,78.$$

$RIC(Y) = Q_{3Y} - Q_{1Y} = 1.$

A partir de los resultados obtenidos, la distribución de las cantidades de imperfecciones en los rollos de 100 m de cable unipolar de 2,5 mm puede describirse del siguiente modo.
La cantidad media es de 0,45 imperfecciones por rollo.
El primer cuartil y la mediana coinciden en el valor 0, que además es la moda. Esto significa que la mayoría de los rollos de ese tipo (70 %) no presentan imperfecciones. El tercer cuartil indica que el 75 % de los rollos tienen como máximo 1 imperfección.

En cuanto a la variabilidad, el rango intercuartílico es de 1 imperfección, mientras que el desvío estándar es de aproximadamente 0,78 imperfecciones.
La distribución es asimétrica a la derecha.



Con lo visto hasta ahora, se pueden resolver las Actividades 7 a 11, Sección 3.5.

3.3.3 Parámetros que resultan cuando se aplica una transformación lineal

Suponga que a una variable aleatoria Y se le aplica una transformación lineal tal que surge una nueva variable aleatoria $X = a + bY$, con a y b pertenecientes a los números reales.

Entonces, conociendo el valor de la media y el desvío estándar poblacional de Y y aplicando propiedades de las mismas, pueden obtenerse los valores de dichas medidas para la variable aleatoria transformada (X), de la siguiente manera: $E(X) = a + bE(Y)$ y $D(X) = |b|D(Y)$.

La misma expresión propuesta para obtener el valor de $E(X)$ a partir del valor de $E(Y)$, se puede aplicar para obtener los valores de otros parámetros de posición como la moda, la mediana y cualquier otro percentil; mientras que la misma expresión propuesta para obtener el valor de $D(X)$ a partir del valor de $D(Y)$, se puede aplicar para obtener el valor de otros parámetros de dispersión como el rango y el rango intercuartílico. Para el caso de la variancia puede decirse que $V(X) = b^2 V(Y)$.



En el Apéndice (Sección 3.7.2) se demuestran las expresiones presentadas para $E(X)$, $V(X)$ y $D(X)$.



Problema 1 (pág. 7): *Reconsidere el problema de las longitudes, medidas en mm, de las barras producidas en una empresa metalúrgica (Y). Suponga que se conoce que la longitud promedio de las barras de la empresa metalúrgica, $E(Y)$, es 250 mm y que el desvío estándar de la longitud, $D(Y)$, vale 0,2 mm.*

- *Un cliente en particular requiere la información sobre las longitudes, medidas en pulgadas.*

Para este cliente, se define a la variable aleatoria X: longitud de una barra, en pulgadas. Dado que 1 mm $= 0,0394$" [5], la longitud en pulgadas se obtiene multiplicando la longitud en mm por el factor de conversión 0,0394 $\frac{"}{mm}$. Por lo tanto, $X = 0,0394\frac{"}{mm}Y$.

Definida X como función lineal de Y, y conocidos los valores de $E(Y)$ y de $D(Y)$, se pueden obtener los valores de $E(X)$ y de $D(X)$ de la siguiente manera:

$E(X) = 0,0394\frac{"}{mm}E(Y) = 0,0394\frac{"}{mm} \cdot 250\ mm = 9,85"$.
$D(X) = |0,0394|\frac{"}{mm}D(Y) = 0,0394\frac{"}{mm} \cdot 0,2\ mm = 0,0079"$.

- *Considerando a las longitudes de las barras medidas en pulgadas, un encargado del proceso advierte que desde hace muchísimo tiempo estas se obtienen con un equipo que mide por exceso 0,5". Si se corrigiera este error, ¿se modificarían la longitud promedio y el desvío estándar?*

En esta situación, se define la variable aleatoria X': longitud corregida de una barra, en ". Observe que $X' = X - 0,5" = 0,0394\frac{"}{mm}Y - 0,5"$.

Definida X' como función lineal de Y, y conocidos los valores de $E(Y)$ y de $D(Y)$, se pueden obtener los valores de $E(X')$ y de $D(X')$ de la siguiente manera:

$E(X') = 0,0394\frac{"}{mm}\,E(Y) - 0,5" = 0,0394\frac{"}{mm} \cdot 250\ mm - 0,5" = 9,35"$;

$D(X') = |0,0394|\frac{"}{mm}\,D(Y) = 0,0394\frac{"}{mm} \cdot 0,2\ mm = 0,0079"$.

Observe que $D(X') = D(X)$, ya que cuando a una variable aleatoria se le suma o resta una constante, los valores de los parámetros de dispersión no se modifican.
En cambio, $E(X') = E(X) - 0,5$, ya que, cuando a una variable aleatoria se le suma o resta una constante, los valores de los parámetros de localización de la nueva variable resultan de sumar o restar esa misma constante a los valores de los parámetros de localización de la variable original.

[5] En este caso se simplifica la conversión de mm a pulgadas, multiplicando por el factor 0,0394

Ejemplo 3.1 Suponga que se tiene una variable aleatoria Y, de la que se conoce que $E(Y) = \mu$ y $D(Y) = \sigma$ y se desea aplicar la siguiente transformación lineal: $Z = \frac{Y-\mu}{\sigma}$. Esta transformación se denomina **estandarización**.

Z es una nueva variable aleatoria que resulta ser una función lineal de Y. Esta variable se denomina **variable estandarizada** y es adimensional, ya que las cantidades en el numerador y el denominador tienen la misma unidad de medida.

La variable Z puede expresarse de la siguiente manera: $Z = \frac{1}{\sigma}Y - \frac{\mu}{\sigma}$.

Conocidos los valores de $E(Y)$ y $D(Y)$, se pueden obtener los valores de $E(Z)$ y $D(Z)$ como se describe a continuación:

$E(Z) = E(\frac{1}{\sigma}Y - \frac{\mu}{\sigma}) = \frac{1}{\sigma}E(Y) - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$;

$D(Z) = D(\frac{1}{\sigma}Y - \frac{\mu}{\sigma}) = |\frac{1}{\sigma}|\,D(Y) = \frac{\sigma}{\sigma} = 1$.

Observe que $E(Z) = 0$ y $D(Z) = 1$, independientemente de los valores que asuman $E(Y)$ y $D(Y)$.

En la Tabla 3.1 se muestran, a modo de ejemplo, algunos valores de una variable aleatoria Y, con $E(Y) = 30$ y $D(Y) = 2$ y los valores correspondientes estandarizados.

Observe que el valor de la variable Z indica a qué distancia está el valor de la variable Y respecto de la media μ, en términos de σ. Para el caso de $y = 31$, $z = 0,5$ indica que se trata de un valor que está $0,5$ desvíos a la derecha del promedio ($0,5\,\sigma = 0,5 \cdot 2 = 1$ unidad a la derecha de $\mu = 30$).

Tabla 3.1 *Algunos valores de una variable aleatoria Y y valores de la variable Z correspondientes*

Valores de Y	**Valores de** $Z = \frac{Y-\mu}{\sigma}$
$y = 30$ $(y = \mu)$	$z = 0$
$y = 32$ $(y = \mu + \sigma)$	$z = 1$
$y = 28$ $(y = \mu - \sigma)$	$z = -1$
$y = 31$ $(y = \mu + 0,5\,\sigma)$	$z = 0,5$
$y = 25$ $(y = \mu - 2,5\,\sigma)$	$z = -2,5$

La transformación lineal presentada en el Ejemplo 3.1 constituye la forma más usual de estandarización de una variable, y consiste en ajustar sus valores para que el nuevo promedio sea igual a 0 y el nuevo desvío estándar sea igual a 1. Esta estandarización de variables se aplica en algunas operaciones relacionadas con la distribución Normal (que se presenta en el Capítulo 5, Sección 5.2.1) y en la definición de los estadísticos pivote (que se presenta en el Capítulo 7, Sección 7.4.3). También se puede aplicar en el análisis de datos cuando se consideran dos o más variables que tienen rangos de variación muy diferentes: estandarizarlas previamente las hace comparables.

Con lo visto hasta ahora, se puede resolver la Actividad 12, Sección 3.5.

3.3.4 Algunas propiedades de los procedimientos de medición: parámetros asociados

Cuando el objetivo de un estudio estadístico consiste en evaluar el desempeño de un procedimiento de medición, se requiere realizar mediciones repetidas de una misma magnitud.

La población se define como la totalidad de mediciones de una magnitud δ que se pueden realizar, con el procedimiento en cuestión, en condiciones similares. Se trata de una población conceptual e infinita compuesta por unidades intangibles. La variable aleatoria asociada, X, se define como el resultado de una medición.

En el resultado de cada medición pueden estar presentes el error sistemático (o sesgo) y el error aleatorio. El primero se simboliza con k y representa la parte del error que es constante en cada medición; el segundo se simboliza con ε y varía entre mediciones, es decir, es una variable aleatoria. Dado que pueden producirse errores aleatorios tanto por defecto como por exceso, es razonable pensar que, a largo plazo, los mismos se compensarán, por lo que puede suponerse que $E(\varepsilon) = 0$.

La variable aleatoria X, se puede expresar entonces de la siguiente manera:

$X = \delta + k + \varepsilon$

Como se expresó anteriormente, el error sistemático o sesgo no varía de medición a medición; mientras que el error aleatorio es de magnitud variable. A su vez, el valor real de la magnitud δ es constante. Es decir que la variable "Resultado de una medición" (X), puede expresarse como la suma de una variable aleatoria, ε y una constante, $(\delta + k)$.

Por lo visto en la sección anterior, al ser X una nueva variable aleatoria, función lineal de la variable aleatoria ε, se puede obtener $E(X)$, $D(X)$ a partir de los valores de $E(\varepsilon)$ y $D(\varepsilon)$, de la siguiente manera:

$E(X) = E(\delta + k + \varepsilon) = \delta + k + E(\varepsilon) = \delta + k$ (ya que $E(\varepsilon)$ se puede considerar nula)

$D(X) = D(\delta + k + \varepsilon) = D(\varepsilon)$.

Interesan algunas propiedades del procedimiento de medición. En particular, su exactitud y su precisión.

Un procedimiento de medición se considera exacto cuando el sesgo o error sistemático es nulo o muy pequeño.

Si el error sistemático es muy pequeño, el valor del promedio de todas las posibles mediciones repetidas de una magnitud δ, $E(X)$, es muy próximo al valor de dicha magnitud. Si el error sistemático es nulo, las mediciones repetidas de cierta magnitud, coinciden, en promedio, con el valor real de la misma, es decir, $E(X) = \delta$.

La precisión del procedimiento de medición se asocia al grado con que tienden a coincidir las mediciones repetidas de la misma magnitud. Si las mediciones repetidas resultan cercanas entre sí, la precisión es alta. Si son muy dispersas, la precisión es baja. La precisión se determina mediante la desviación estándar, parámetro al que con frecuencia ingenieros y científicos refieren como "incertidumbre estadística" del proceso de medición.

Un procedimiento de medición se considera preciso si el valor del desvío estándar, $D(X)$, es pequeño.

En la Figura 3.10 se muestran algunos resultados obtenidos de aplicar cuatro procedimientos diferentes para medir una magnitud δ.

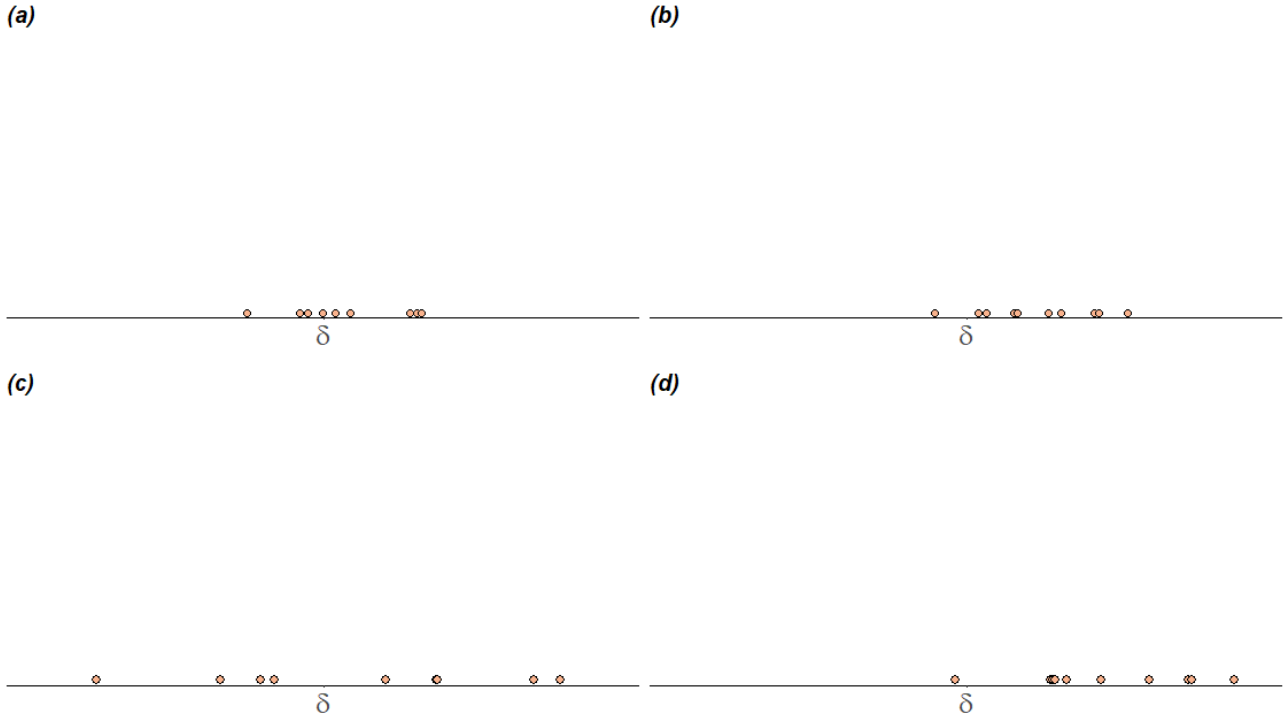


Figura 3.10 *Algunas mediciones de una magnitud δ obtenidas con diferentes procedimientos*

Observe que para los procedimientos a y c, las mediciones obtenidas se ubican alrededor de δ, lo que hace pensar que el promedio poblacional de todas las mediciones repetidas de dicha magnitud es cercano a ese valor y concluir que ambos procedimientos son exactos. En cambio, para los procedimientos b y d, las mediciones obtenidas se ubican alrededor de un valor mayor que δ, lo que hace pensar que el promedio poblacional de todas las mediciones repetidas de dicha magnitud está alejado de ese valor y concluir que ambos procedimientos no son exactos.

Por otra parte, para los procedimientos a y b, la variabilidad que presentan las mediciones obtenidas parece pequeña, lo que hace pensar que el desvío estándar poblacional de todas las mediciones repetidas de la magnitud δ también lo será y concluir que ambos procedimientos son precisos. En cambio, para los procedimientos c y d, la variabilidad que presentan las mediciones obtenidas parece grande, lo que hace pensar que el desvío estándar poblacional de todas las mediciones repetidas de la magnitud δ también lo será y concluir que ambos procedimientos no son precisos.

En relación a los procedimientos que se utilizaron para obtener las mediciones presentadas en la Figura 3.10, es importante definir previamente qué maximo valor se admite para el error sistemático, k y qué valores se consideran pequeños para el desvío poblacional, σ, para luego definir si dichos procedimientos son exactos y/o precisos.

Por otra parte, las conclusiones para cada procedimiento se obtienen a partir de una muestra de mediciones y por lo tanto se deben considerar **preliminares**: hace falta aplicar herramientas inferenciales para concluir sobre su exactitud y su precisión.



Problema 7 (pág. 8): *Una empresa desea evaluar el desempeño de un multímetro digital para medir tensión en corriente continua (en voltios). Para ello lo envía a un laboratorio especializado en el que van a realizar mediciones repetidas de una tensión conocida, con el mencionado equipo.*
Aunque el multímetro mida muchas veces la misma tensión conocida, todas las mediciones serán diferentes y es probable que ninguna coincida con la magnitud real ya que, en cada medición se pueden presentar errores.
Si el multímetro no está calibrado adecuadamente, cada medición estará lejos de su valor real en cierta cantidad fija. Es decir que una calibración imperfecta aporta errores de la misma magnitud en cada medición. Problemas en la calibración u otros que den como resultado errores que se mantengan constantes de medición a medición, dan lugar a un error sistemático de magnitud k.
Existen, además, otros errores posibles en cada medición debidos a conexiones defectuosas, baterías dañadas, fallas del operador del equipo, interferencias, cambios en la temperatura ambiente, etc. Estos errores pueden variar entre una medición y otra y se consideran en la variable aleatoria ε.
En este contexto, ***la población está definida por todas las mediciones que se pueden realizar con cierto procedimiento (que incluye el uso del multímetro), de una tensión conocida δ, en condiciones similares*** *y* ***la variable de interés X es el resultado de una medición de la tensión δ con ese procedimiento****.*
Considerando los posibles errores de este procedimiento de medición de distancias, la variable X se puede expresar de la siguiente manera: $X = \delta + k + \varepsilon$, de modo que:
$E(X) = \delta + k$
$D(X) = D(\varepsilon) = \sigma_\varepsilon$.

Suponga ahora que $\delta = 12$ V y que se admite, como máximo, un error sistemático de 0,05 V. Es decir que, si

$k \leq 0,05$ V, el procedimiento de medición de tensión con el multímetro se puede considerar exacto. Suponga además que se pretende que el desvío estándar sea inferior al 2% del valor de δ para que el procedimiento de medición se pueda considerar preciso.
Con estos criterios, se puede afirmar que:

- *El procedimiento se considera exacto si $11,95\ V \leq E(X) \leq 12,05\ V$.*
- *El procedimiento se considera preciso si $D(X) < 0,24\ V$.*

Suponga ahora que $\delta = 20$ V y que se admite, como máximo, un error sistemático de 0,5 V. Es decir que, si $k \leq 0,5$ V, el procedimiento de medición de tensión con el multímetro se puede considerar exacto. Suponga, además, que se pretende que el desvío estándar sea inferior a 1 V para que el procedimiento se pueda considerar preciso.
Con estos criterios, se puede afirmar que:

- *El procedimiento se considera exacto si $19,5\ V \leq E(X) \leq 20,5\ V$.*
- *El procedimiento se considera preciso si $D(X) < 1\ V$.*



Como se mencionó para los procedimientos presentados en la Figura 3.10 y en el ejemplo del procedimiento de medición de la tensión con cierto multímetro, en cada situación es importante definir el máximo valor que se admite para el error sistemático, k y para el desvío estándar, $D(\varepsilon)$, para luego poder analizar si un procedimiento es exacto y/o preciso. Independientemente de ello, se puede afirmar que cuanto más pequeño sea k, el procedimiento de medición será más exacto y que cuanto más pequeño sea $D(X)$, será más preciso.



Problema 1 (pág. 7): *Reconsidere la situación de las longitudes de las barras, medidas en pulgadas. Un encargado del proceso advierte que desde hace muchísimo tiempo estas se obtienen con un equipo que mide por exceso 0,5".*
En este caso, se puede decir que el equipo al que hace referencia el encargado, presenta un error sistemático, k, de 0,5".
¿Considera que el procedimiento de medición de las longitudes de las barras con dicho equipo es exacto? ¿De qué depende su respuesta?



Con lo visto hasta ahora, se pueden resolver las Actividades 13 y 14, Sección 3.5.

3.4 Síntesis

Definidas las variables aleatorias como criterios que permiten asociar un número real a cada unidad de la población, las distribuciones de probabilidades constituyen modelos que describen su comportamiento en esa población y permiten obtener el valor de diferentes parámetros.

En el caso de las variables aleatorias continuas la distribución se representa a través de la función de densidad de probabilidad y en el caso de las variables aleatorias discretas a través de la función de probabilidad puntual. En ambos casos se puede definir la función de distribución acumulada.

Las probabilidades y el resto de las medidas definidas constituyen parámetros ya que brindan información poblacional y se utilizan no solo para describir el comportamiento de la variable en la población sino para tomar decisiones.

En la siguiente tabla se indica cómo obtener probabilidades y otros parámetros de interés, tanto para el caso de variables continuas como discretas.

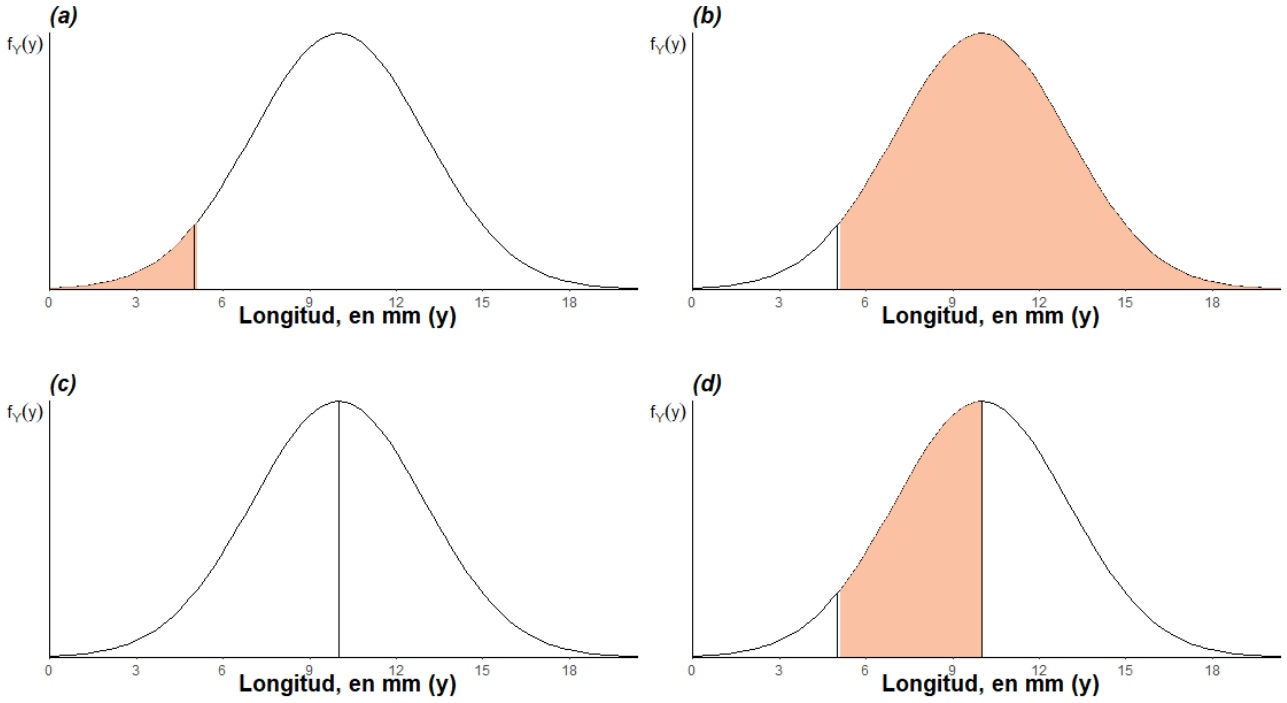
	Variable continua	**Variable discreta**
Distribución de probabilidades	f_Y: función de densidad de probabilidad	p_Y: función de probabilidad puntual
Cálculo de probabilidades con f_Y o p_Y	$P(Y = y_1) = 0$ $P(Y < y_1) = \int_{-\infty}^{y_1} f_Y(y)dy$ $P(Y > y_1) = \int_{y_1}^{+\infty} f_Y(y)dy$ $P(y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} f_Y(y)dy$	$P(Y = y_1) = p_Y(y_1)$ $P(Y < y_1) = \sum_{y<y_1} p_Y(y)$ $P(Y > y_1) = \sum_{y>y_1} p_Y(y)$ $P(y_1 \leq Y \leq y_2) = \sum_{y=y_1}^{y_2} p_Y(y)$
$F_Y(y)$	$\int_{-\infty}^{y} f_Y(s)ds$	$\sum_{s\leq y} p_Y(s)$
$E(Y) = \mu_Y$	$\int_{y\in R_Y} y\, f_Y(y)dy$	$\sum_{y\in R_Y} y\, p_Y(y)$
$D(Y) = \sigma_Y$	$\sqrt{\int_{y\in R_Y} y^2 f_Y(y)d_y - \mu_Y^2}$	$\sqrt{\sum_{y\in R_Y} y^2 p_Y(y) - \mu_Y^2}$

Conocida una variable aleatoria Y, en algunas situaciones puede ser de utilidad definir otra variable aleatoria X, como función lineal de la misma. Definida X, se pueden obtener los valores de algunos parámetros de interés, a partir de los valores de los parámetros correspondientes a la variable Y, como se muestra en la Sección 3.3.3.

El error aleatorio de una medición (ε), así como el resultado de la misma (X), son variables aleatorias de interés en la evaluación de los procedimientos de medición. Los parámetros $E(X)$ y $D(\varepsilon)$ dan información sobre el desempeño de un procedimiento de medición cuando se realizan mediciones repetidas de una magnitud δ, en condiciones similares. Específicamente, $E(X)$ informa sobre la exactitud y $D(\varepsilon)$ informa sobre la precisión. Dados los valores de estos parámetros, se necesita un criterio, definido por los usuarios del procedimiento, para concluir si el mismo es exacto y/o preciso.

3.5 Actividades propuestas

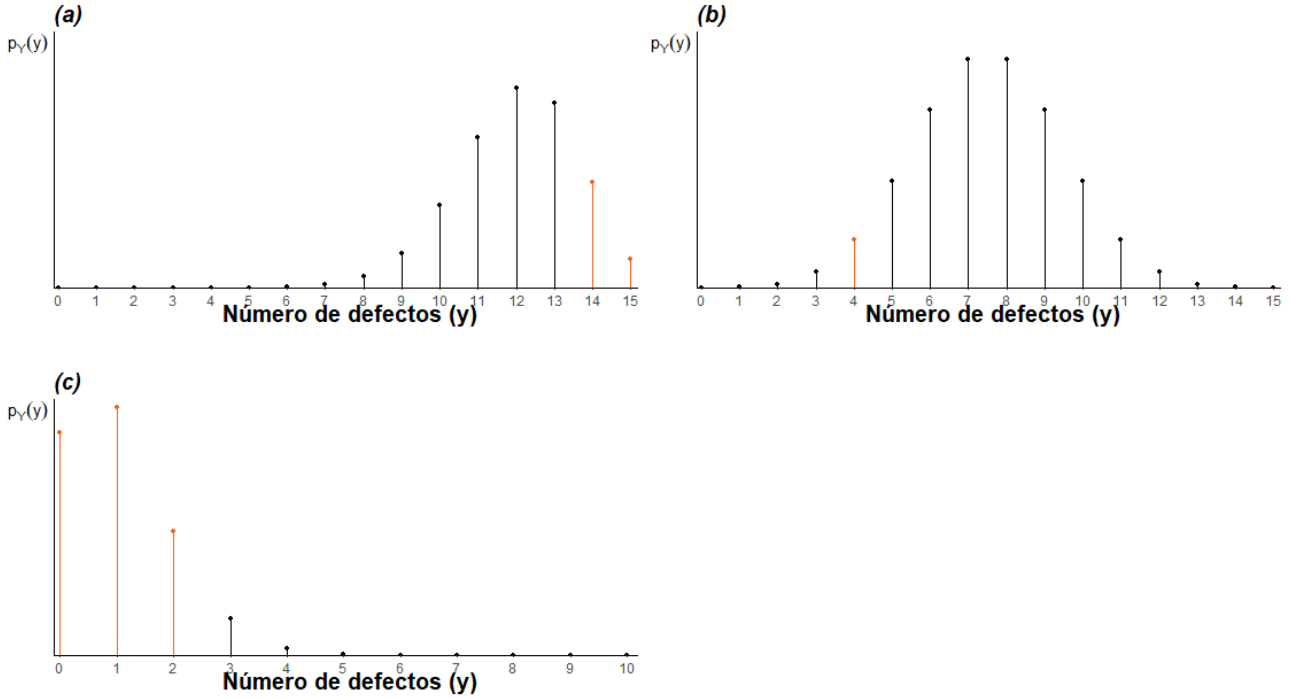
1. En cada caso, se representa la función de densidad de probabilidad de la longitud de ciertas piezas metálicas necesarias para el armado de electrodomésticos (en milímetros):



a) Indique la población en estudio.
b) Exprese analiticamente la proporción que está sombreada en cada caso.
c) Asigne un gráfico a cada uno de los siguientes enunciados, teniendo en cuenta la probabilidad que representan:
   - La proporción de piezas metálicas con longitudes entre 5 mm y 10 mm.
   - La proporción de piezas con longitud de 10 mm.
   - La proporción de piezas con una longitud de a lo sumo 5 mm.
   - La proporción de piezas metálicas que tienen longitudes superiores a 5 mm.

d) ¿Alguna de las proporciones representadas es nula? Explique.
e) La proporción de piezas con longitud de al menos 5 mm e inferior a 10 mm, ¿coincide con alguna de las anteriores? Justifique.

2. En cada caso, se representa la distribución de probabilidades del número de defectos presentes en cajas de cartón utilizadas para el embalaje de electrodomésticos:



a) Indique la población en estudio.

b) Exprese analiticamente la proporción que se encuentra coloreada en cada caso.

c) Asigne un gráfico a cada uno de los siguientes enunciados, según corresponda:

- La distribución es simétrica.
- La proporción de cajas con al menos 3 defectos es inferior a $0,1$.
- El $50\,\%$ de las cajas presentan a lo sumo 7 defectos.
- La distribución es asimétrica por derecha.
- El número de defectos por caja con mayor frecuencia es 12.
- La distribución es asimétrica por izquierda.

3. Una empresa produce piezas de grifería, a las cuales recubre con una capa de cromado. Para un tipo de pieza en particular, se define a la variable aleatoria *C: Espesor de la capa de cromado (en milímetros)*. Por estudios anteriores, se sabe que $f(c) = kc$, con $0,03 \leq c \leq 0,05$, modela la función de densidad de probabilidad de la variable en estudio.

   a) Calcule el valor de k que hace que f sea una función de densidad de probabilidad para la variable en estudio.

   b) Calcule e interprete $P(0,035 \leq C \leq 0,04)$ en el contexto del problema.

   c) ¿Cuánto vale $P(C = 0,035)$? ¿Cómo se interpreta este resultado?

   d) Calcule $P(0,035 < C \leq 0,04)$. ¿Qué relación existe entre el resultado obtenido en este ítem y en el apartado *b*?

   e) Cuando el espesor de la capa de cromado en este tipo de piezas es mayor que 0,048 mm, aumenta considerablemente el costo de las mismas, de modo que se pretende que esto ocurra como máximo en el $10\,\%$ de las piezas. ¿Considera Ud. que se cumple con lo pretendido por la empresa en relación a esto? Justifique.

4. Para rollos de alambre de cierto tipo interesa la variable *Y: Cantidad de defectos*, cuya función de probabilidad puntual es:

$$P(Y=y)=p_Y(y)=\frac{2^y e^{-2}}{y!} \text{ con } y\in\mathbb{N}_0$$

   a) Identifique la población en estudio.
   b) ¿Qué proporción de rollos de alambre no tienen defectos?
   c) ¿Qué proporción de rollos tienen a lo sumo 2 defectos?
   d) ¿Qué proporción de rollos tienen por lo menos 3 defectos?
   e) Determine si la siguiente igualdad es VERDADERA o FALSA, justificando apropiadamente: $F_Y(6)-F_Y(2)=P(2\leq Y\leq 6)$
   f) Para la empresa, los rollos de alambre con más de 4 defectos se consideran de segunda calidad, por lo que se venden a menor precio. Si el porcentaje de rollos de segunda calidad es superior al 5 %, se debe ajustar el proceso de producción ya que la venta a menor precio generaría pérdidas. ¿Qué recomendación daría a la empresa?

5. En una planta industrial dedicada a la producción de piezas fundidas de aluminio, las mismas se envían a los clientes en lotes con una gran cantidad de unidades. Cada pieza debe cumplir con una serie de requisitos para ser considerada buena por los clientes. En cada lote, interesa la variable aleatoria *Y: proporción de piezas buenas*. La función de densidad de probabilidad de esta variable es:

$$f(y)=\begin{cases}3y^2 & \text{si } 0<y<1\\ 0 & \text{en otro caso}\end{cases}$$

   a) Defina la población y la variable de interés adecuadamente y verifique que f es una función de densidad de probabilidad.
   b) Calcule $P\left(Y<\frac{1}{2}\right)$ e interprete el resultado obtenido en el contexto del problema.
   c) Grafique la función de densidad de probabilidad, f, y represente gráficamente el valor obtenido en el ítem anterior.
   d) Halle la función de distribución acumulada para la variable Y.
   e) Exprese $P(Y\geq 0,80)$ utilizando la función de distribución acumulada. Calcule dicha probabilidad e interprete el resultado obtenido en el contexto del problema.

6. En una empresa comercializan chips en cajas de 100 unidades. En estas cajas interesa la variable aleatoria: *X: Cantidad de chips defectuosos*, que presenta la siguiente distribución:

x	$P(X=x)$
0	t
1	0,05
2	0,03
3	0,02
En otro caso	0

a) Calcule el valor de t de manera tal que la distribución presentada en la tabla represente efectivamente una distribución de probabilidades para la variable X.
b) Explique cómo cree que se obtuvieron las proporciones de la tabla.
c) ¿Qué proporción de cajas contienen a lo sumo un chip defectuoso?
d) ¿Qué proporción de cajas contienen por lo menos 2 chips defectuosos?
e) ¿Qué proporción de cajas contienen entre 1 y 2 chips defectuosos, ambos extremos incluidos?
f) Obtenga y grafique la función de distribución acumulada para la variable X.
g) Exprese a los valores obtenidos en los ítems c, d y e en función de F_X.

7. Considere la variable aleatoria *W: Cantidad de veces que un proceso de producción necesita recalibrarse por semana*, cuya función de distribución acumulada verifica:

w	$F_W(w)$
$w < 0$	0
$0 \leq w < 1$	0,15
$1 \leq w < 2$	0,55
$2 \leq w < 3$	0,85
$3 \leq w < 4$	0,98
$w \geq 4$	1

a) ¿Qué proporción de semanas el proceso requiere recalibrarse exactamente una vez?
b) ¿Qué proporción de semanas el proceso requiere recalibrarse a lo sumo 2 veces?
c) Grafique la función de probabilidad puntual asociada a la variable aleatoria W.
d) Calcule la esperanza y la desviación estándar de la variable W e interprete los resultados obtenidos en el contexto del problema planteado.
e) Calcule e interprete los cuartiles de la distribución.
f) Las medidas calculadas en los apartados anteriores ¿son parámetros o estadísticos? Justifique.

8. El tiempo (en horas) requerido para el armado de un artículo electrónico es una variable aleatoria y su función de densidad de probabilidad es $f_Y(y) = 3y^{-4}$ para $1 < y$.
a) Halle la función de distribución acumulada para la variable en estudio.
b) Calcule la proporción de artículos que requieren entre 2 h y 4 h para ser armados, utilizando la función de densidad de probabilidad y la función de distribución acumulada.
c) Halle el valor de q tal que $P(Y \leq q) = 0,5$. ¿Qué representa el valor hallado? Interprételo en el contexto del problema.
d) Utilice la función obtenida en (a) para obtener el percentil 99 de los tiempos. Interprételo en el contexto del problema.
e) Halle la esperanza y la desviación estándar de la variable en estudio. Interprete los valores hallados en el contexto del problema.

9. Cierto tipo de máquina falla diariamente a lo sumo dos veces. Se conoce que en el 70 % de los días tiene a lo sumo una falla y que el número promedio de fallas diarias es 0,85.

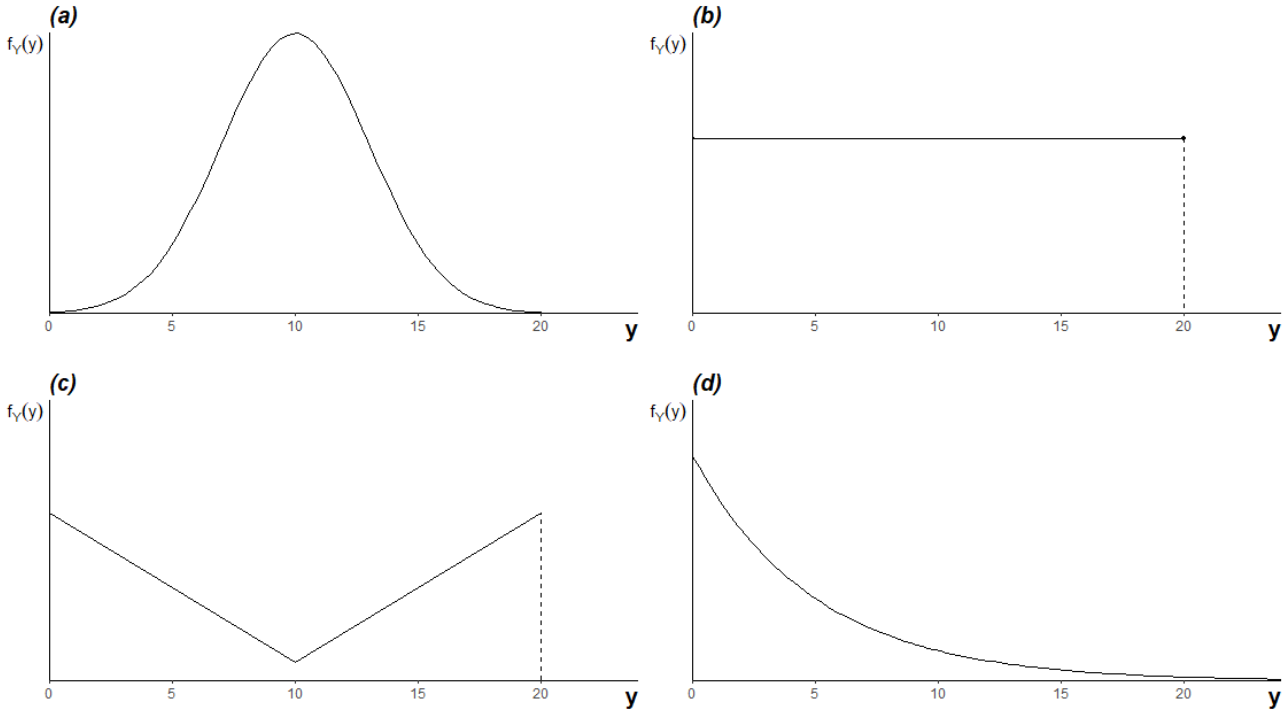
a) Indique cómo pudo llegarse a determinar el porcentaje indicado.

b) Se define la variable aleatoria *X: Cantidad de fallas por día*. Indique cuál es la población en estudio. Obtenga la distribución de probabilidades de X a partir de los datos dados y represéntela gráficamente.

c) Calcule la mediana de la cantidad de fallas. ¿Es simétrica la distribución? Justifique.

d) Obtenga el desvío estándar de X e interprételo en el contexto del problema. ¿Este valor corresponde a un parámetro o a un estadístico? Justifique.

10. En cada caso, se representa la función de densidad de probabilidad de una variable aleatoria continua:



a) Indique el valor aproximado del promedio de la variable Y a partir de lo que se observa en la gráfica de la función de densidad de probabilidad.

b) Indique si las siguientes afirmaciones son VERDADERAS o FALSAS justificando adecuadamente en cada caso:

1) La distribución representada en el gráfico (d) es asimétrica a derecha.
2) En la distribución (b) los tres cuartiles valen 5, 10 y 15 unidades respectivamente.
3) La distribuciones representadas en los gráficos (a) y (b) son simétricas.
4) El valor de la moda para la distribución (d) es mayor que cualquiera de las modas de la distribución (c).
5) En la distribución (d), promedio y mediana coinciden.
6) Las distribuciones (a), (c) y (d) son unimodales.
7) La distribución representada en el gráfico (c) es asimétrica a izquierda.

c) Ordene (en orden creciente) a las distribuciones representadas en los gráficos (a), (b) y (c) según su variabilidad, considerando como medida de variabilidad:

1) al rango.
2) a la variancia.

11. La velocidad (en kilómetros por hora) de los autos que pasan por un determinado punto de control de la autopista Rosario-Córdoba es una variable aleatoria con función densidad de probabilidad:

$$f(x) = \begin{cases} \frac{x}{10000} & \text{si } 0 < x \leq 100 \\ 0,02 - \frac{x}{10000} & \text{si } 100 < x \leq 200 \\ 0 & \text{en caso contrario} \end{cases}$$

a) Represente gráficamente la función densidad de probabilidad, f.
b) ¿Qué proporción de vehículos circulan a menos de 100 km/h? Represente el resultado obtenido en el gráfico construido en el apartado anterior.
c) En el puesto de control hay un radar que controla la velocidad de los vehículos. Si la velocidad es inferior a 100 km/h el importe de la multa es de \$0 (no hay multa), en cambio si la velocidad está comprendida entre 100 y 120 km/h la multa es de \$8500 y si la velocidad supera los 120 km/h la multa es de \$12000. Halle la distribución de probabilidades de la variable aleatoria *I: Importe de la multa que tiene que pagar un vehículo elegido al azar (en \$)*.
d) Calcule e interprete el valor de la esperanza matemática y de la desviación estándar de la variable aleatoria I.

12. Una empresa fabrica placas con un determinado tipo de acero y la característica más crítica de estas es la resistencia a la fractura, medida en megapascales. De esta variable aleatoria, R, se conocen los valores de los siguientes parámetros:
$Q_{1R} = 74$, $Q_{2R} = 76,20$, $Q_{3R} = 79$, $E(R) = 80$ y $D(R) = 3,2$.
Suponga que se defina una variable aleatoria X, de la siguiente manera: $X = a + bR$, con $b \geq 0$.

a) Indique, en cada caso, cuál fue la transformación lineal aplicada, si se conocen los siguientes parámetros correspondientes a la variable X. Justifique.
1) $Q_{3X} = 89$ y $RIC(X) = 5$.
2) $E(X) = 72$ y $D(X) = 2,88$.
3) $Q_{1X} = 70$ y $D(X) = 3,52$.
b) Definida la transformación correspondiente al ítem a-1, indique el nuevo valor de Q_{1X}, Q_{2X}, $E(X)$ y $D(X)$. Justifique.
c) Indique si para alguna de las transformaciones realizadas, el coeficiente de variación de la variable X coincide con el de la variable R. Justifique.

13. Considere la realización de mediciones repetidas de la solubilidad de una determinada sustancia mediante un cierto procedimiento. Sea Y la variable aleatoria asociada a dichas mediciones, en gramos por litro (g/L) de agua a 25 °C. Suponga que se conoce por experiencia que $E(Y) = 0,8$ y que $D(Y) = 0,02$.

a) Defina a la población y a la variable aleatoria de interés.
b) Suponga que la solubilidad de dicha sustancia es 0,9 g/L a 25 °C. Exprese a la variable aleatoria Y en función de este valor y de los posibles errores de medición.

c) ¿El procedimiento se puede considerar exacto? Justifique su respuesta y, si esta es negativa, indique el valor del error sistemático.

d) Suponga que el procedimiento se considera preciso si el parámetro correspondiente es menor que el 3 % de la solubilidad de la sustancia. Concluya sobre la precisión del procedimiento.

14. En una empresa cuentan con dos procedimientos, A y B, para medir longitudes. Suponga que se realizan mediciones repetidas de una longitud δ, en milímetros, tanto con el procedimiento A como con el B.

Se definen las variables aleatorias ε_A: error aleatorio de una medición de la longitud δ, con el procedimiento A y ε_B: error aleatorio de una medición de la longitud δ, con el procedimiento B.

Se conocen los valores de los siguientes parámetros:

$E(\varepsilon_A) = E(\varepsilon_B) = 0$, $D(\varepsilon_A) = 0,8$ y $D(\varepsilon_B) = 0,5$.

Suponga que el procedimiento A tiene error sistemático nulo; mientras que el procedimiento B tiene un error sistemático, k = -0,2.

Dadas las variables Y_A e Y_B, definidas como el resultado de una medición de la longitud δ mediante el procedimiento A o B, según corresponda:

a) Indique los valores de la esperanza y del desvío estándar para cada una de ellas. Justifique.

b) Suponga que un procedimiento se considera preciso si el desvío estándar de una medición de cierta magnitud δ, es menor al 10 % del valor de dicha magnitud. Proponga un posible valor para δ, de tal manera que se concluya que uno de los procedimientos es preciso y el otro no. Justifique su propuesta.

c) Indique por qué las siguientes afirmaciones son verdaderas:

- El procedimiento A es más exacto que el B.
- Es posible que el procedimiento B se considere exacto.
- Los resultados de las mediciones de la longitud δ realizadas mediante el procedimiento B se encuentran alrededor de $\delta - 0,2$ mm.

3.6 Estadística con R

3.6.1 Variables aleatorias continuas

1. Funciones de densidad de probabilidad y de distribución acumulada

En R es posible definir funciones de densidad de probabilidad para variables aleatorias. Una vez definida la función de densidad de probabilidad, se puede integrar a fin de obtener probabilidades. Así mismo, se la puede graficar utilizando ggplot2. Por ejemplo, si se está trabajando con la variable Y cuya distribución de probabilidades es la presentada en la página 148, su función de densidad de probabilidad se puede expresar como:

```
f <- function(y) {
ifelse(y < 249 | y > 251,0,
ifelse(y >= 249 & y <= 250, y-249, 251-y))
}
```

Con `ifelse(y < 249 | y > 251,0)` se indica que fuera del intervalo [249 ; 251], la función de densidad de probabilidad `f` toma el valor cero, mientras que con `ifelse(y >= 249 & y <= 250, y-249, 251-y)` se indica que en el intervalo [249 ; 250] la función `f` es `y-249`

Para verificar que efectivamente se trata de una función de densidad de probabilidad, se puede calcular el área del recinto limitado por la curva que representa esta función y el eje de las abscisas y corroborar si vale 1. Para ello, se ejecuta:

```
area <- integrate(f, lower = 249, upper = 251)
area
```

Si, por ejemplo, se pretende obtener $P(249,4 < Y < 250,6)$, se podrá ejecutar:

```
integrate(f, lower = 249.4, upper = 250.6)
```

Esto resulta igual a 0,84.

Para realizar la gráfica de la función de densidad de probabilidad, se trabaja en forma similar a lo visto para otras distribuciones continuas, indicando en `stat_function(fun= )` el nombre de la función creada. Por ejemplo:

```
ggplot(data.frame(x = c(248.5, 251.5)), aes(x = x)) +
    stat_function(fun = f)+
    #Nombre de los ejes
```

```
    labs(x = "\n Longitud, en mm (y)", y = expression(f[Y](y))) +
    #Configuraciones de formato
    #Estilo
    theme_classic()+
    #Fuente para los ejes y nombre eje abscisas centrado
    theme(axis.title.x = element_text(face="bold", colour="black", size = 12,
    hjust = 0.5),
    axis.title.y = element_text(face="bold", colour="black", size = 12))+
    scale_y_continuous(expand=c(0,0),
                       labels = scales::label_number(accuracy = 0.01,
                                                     decimal.mark = ','))  +
      scale_x_continuous(expand=c(0,0),
                       labels = scales::label_number(accuracy = 1,
                                                     decimal.mark = ','))
```

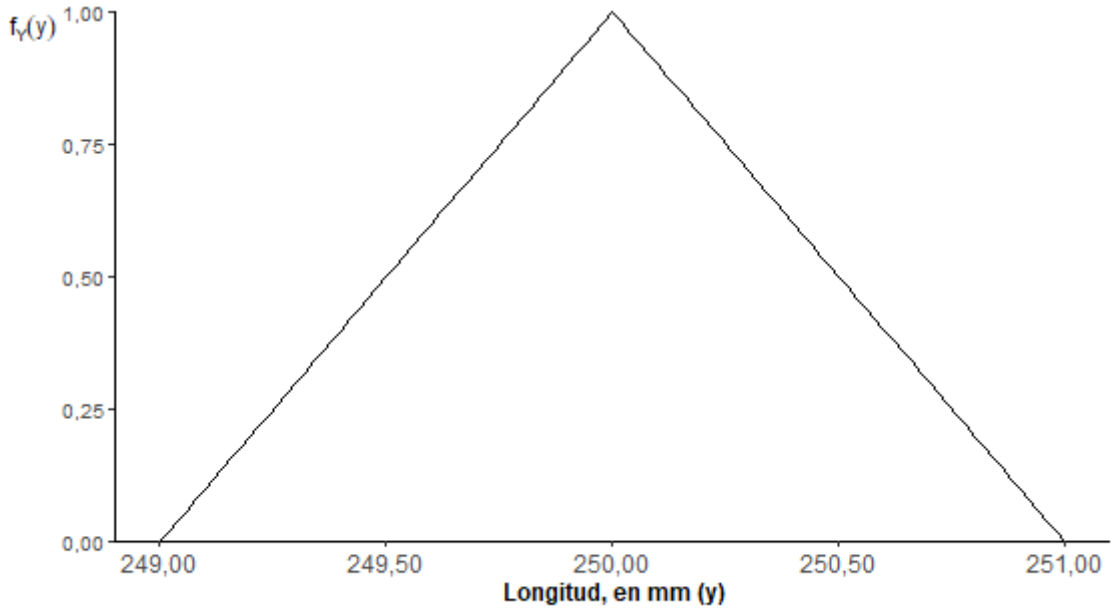


Figura 3.11 *Función de densidad de probabilidad de Y, longitud de las barras, página 149*

De manera similar a lo actuado para la función de densidad de probabilidad, es posible trabajar con la función de distribución acumulada. Para el ejemplo presentado en la página 148, la función de distribución acumulada resulta:

```
a <- 249
b <- 251
c <- 250

F <- function(y) {
  ifelse(y <= a, 0,
```

```
        ifelse(y > a & y <= c, ((y-a)^2)/((b-a)*(c-a)),
               ifelse(y > c & y < b, 1-((b-y)^2)/((b-a)*(b-c)), 1)))
}
```

Para graficar esta función, se ejecuta:

```
ggplot(data.frame(x = c(248.5, 251.5)), aes(x = x)) +
    stat_function(fun = F)+
    #Nombre de los ejes
    labs(x = "\n Longitud, en mm (y)", y = expression(F[Y](y))) +
    #Configuraciones de formato
    #Estilo
    theme_classic()+
    #Fuente para los ejes
    theme(axis.title.x = element_text(face="bold", colour="black", size = 12,
    hjust = 0.5),
      axis.title.y = element_text(face="bold", colour="black", size = 12))+
      scale_y_continuous(expand=c(0,0),
                         labels = scales::label_number(accuracy = 0.01,
                                                       decimal.mark = ','))  +
      scale_x_continuous(expand=c(0,0),
                         labels = scales::label_number(accuracy = 1,
                                                       decimal.mark = ','))
```

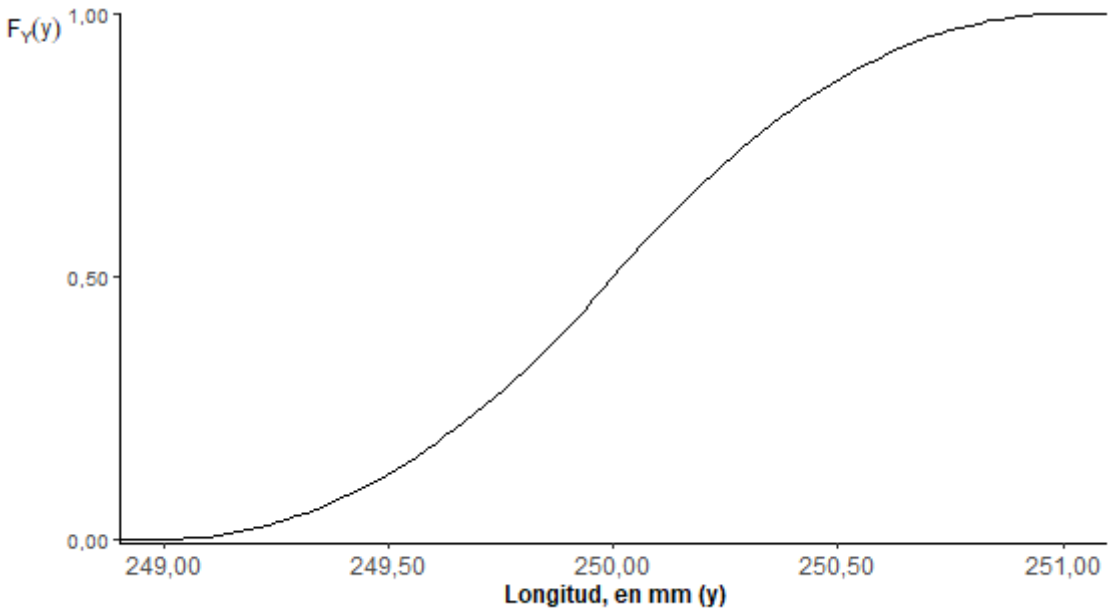


Figura 3.12 *Función de distribución acumulada de Y, longitud de las barras, página 153*

2. Medidas de resumen

Teniendo en cuenta las definiciones dadas para la media, la varianza y el desvío estándar de variables aleatorias, es posible definirlas como funciones en R y trabajar adecuadamente para obtener sus valores a partir de la función de densidad.

Para el caso presentado en la página 148, la forma de obtener la media, la varianza y el desvío estándar es la siguiente:

```
FMedia <- function(x) {x * f(x)}

Media <- integrate(FMedia,
  lower=249,
  upper=251)
print(Media)

FVar <- function(x) {f(x)*(x-Media$value)^2}

Varianza <- integrate(FVar,
                  lower=249,
                  upper=251)
print(Varianza)

Desvio <- sqrt(Varianza$value)
print(Desvio)
```

Siempre es importante definir correctamente los límites de integración `lower` y `upper`, de acuerdo al recorrdio de la variable en cuestión. Si la variable está definida para todos los reales, se puede indicar `lower = -Inf` y `upper = Inf`.

Para encontrar la moda poblacional de la variable aleatoria en estudio, se debe conocer el valor donde la función de densidad alcanza su máximo. Una posible manera de realizar esto es creando un vector que contenga el valor que toma dicha función para distintos valores de la variable, generados mediante una secuencia dentro del correspondiente recorrido. Por ejemplo:

```
x <- seq(249,251,by=0.000001)
p <- f(x)
```

En el vector `x` se incluyen todos los valores entre 249 y 251 cada 0,000001 unidades, y en `p` se calcula el valor que toma `f` en cada uno de ellos. Luego, resta buscar el valor de la variable para el cual la función de densidad toma su máximo, para esto:

```
moda <- x[which.max(p)]
print(moda)
```

Para el cálculo de percentiles, se requiere conocer el comportamiento de la función de distribución acumulada dentro del recorrido de la variable para luego buscar a qué valor de la variable corresponde una probabilidad acumulada específica. A continuación se ejemplifica la búsqueda de la mediana y los cuartiles 1 y 3, en base a la función de distribución acumulada `F` y al vector `x`, creados anteriormente:

```
facum <- F(x)

mediana <- x[which(facum>=0.5)][1]
print(mediana)


Q1 <- x[which(facum>=0.25)][1]
print(Q1)

Q3 <- x[which(facum>=0.75)][1]
print(Q3)
```

En el ejemplo anterior, en el vector `facum` se generan las probabilidades acumuladas para los distintos valores del vector `x`, a partir de la función de distribución acumulada `F`. Luego, para calcular el valor de la mediana, se busca el valor de `x` tal que `facum` asuma un valor igual o superior a 0,5 *por primera vez*. De manera análoga se buscan los cuartiles.

3.6.2 Variables aleatorias discretas

1. Funciones de probabilidad puntual y de distribución acumulada

Las funciones de probabilidad puntual pueden ser definidas en R, mediante la generación de un vector con los valores de la variable y otro con las probabilidades puntuales correspondientes. Esta información, debe ser almacenada en formato tabular a fin de utilizarla posteriormente para realizar gráficos y cálculos. A continuación, se ejemplifica la definición de la función de probabilidad puntual presentada en la página 151, junto con la generación de la función de distribución acumulada (`acum`):

```
y <- seq(0,3,by=1)
p <- c(0.7, 0.18, 0.09, 0.03)
acum <- cumsum(p)

tabla <- data.frame(cbind(y,p, acum))
```

Es posible utilizar la tabla de distribución y consultar probabilidades de interés. Por ejemplo, para obtener $P(Y \leq 1)$ basta con utilizar la sentencia `acum[which(y==1)]`. Si se quiere obtener $P(Y > 2)$ se hace `1 - acum[which(y==2)]`.

Para representar gráficamente la función de probabilidad puntual mediante un diagrama de bastones y la distribución acumulada mediante un gráfico escalonado, se procede de la misma forma que se presentó en el Capítulo 2.

```
ggplot(data=tabla) +
  geom_hline(aes(yintercept=0)) +
  geom_segment(aes(y,p,xend=y,yend=p-p)) +
  geom_point(aes(y,p),size=1.5) +
  labs(x = "\n Cantidad de imperfecciones (y)", y = expression(p[Y](y))) +
  theme_classic()+
  scale_x_continuous(expand=c(0,0), limits = c(-0.1,3.5)) +
  scale_y_continuous(expand=c(0,0), limits = c(0,0.75), breaks=seq(0.1,0.7,0.1),
  labels = scales::label_number(accuracy = 0.01, decimal.mark = ','))+
  theme(axis.title.x = element_text(face="bold", colour="black", size = 11,
  hjust = 0.5, vjust=10),
        axis.title.y = element_text(face="bold", colour="black", size = 11,
        angle = 0, vjust = 1))
```

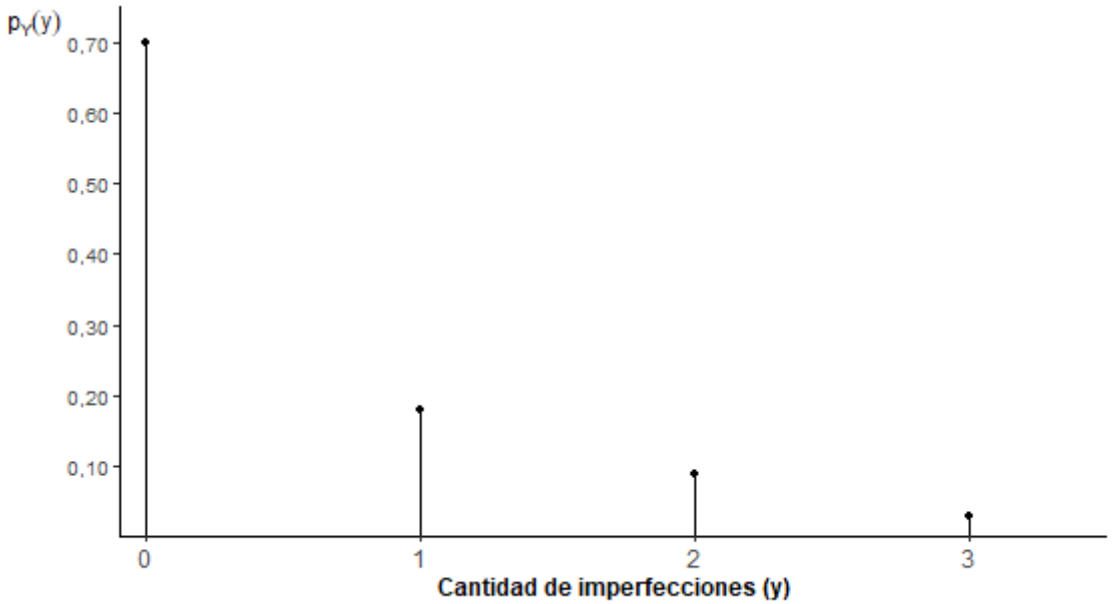


Figura 3.13 *Función de probabilidad puntual de Y, cantidad de imperfecciones por rollo, página 151*

```
ggplot(data=tabla) +
  geom_hline(aes(yintercept=0)) +
  geom_segment(aes(y,acum,xend=y+1,yend=acum)) +
  geom_segment(aes(-0.1,0,xend=0,yend=0)) +
  geom_segment(aes(3,1,xend=3.5,yend=1)) +
  geom_point(aes(y,acum),size=1.5, shape=1) +
  labs(x = "\n Cantidad de imperfecciones (y)", y = expression(F[Y](y))) +
  theme_classic()+
  scale_x_continuous(expand=c(0,0), limits = c(-0.1,3.5)) +
  scale_y_continuous(expand=c(0,0), limits = c(0,1.05), breaks=seq(0.1,1,0.1),
  labels = scales::label_number(accuracy = 0.01, decimal.mark = ','))+
  theme(axis.title.x = element_text(face="bold", colour="black", size = 11,
  hjust = 0.5,  vjust=10),
```

```
        axis.title.y = element_text(face="bold", colour="black", size = 11,
        angle = 0, vjust = 1))
```

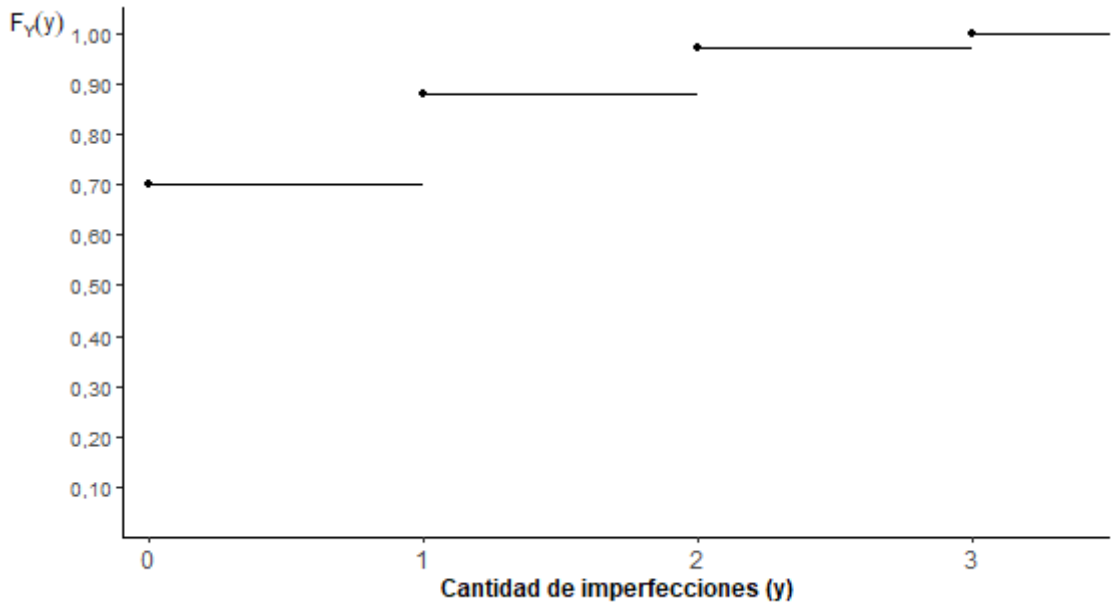


Figura 3.14 *Función de distribución acumulada de Y, cantidad de imperfecciones por rollo, página 154*

2. Medidas de resumen

Para obtener las medidas de resumen de una variable aleatoria discreta, se procede de manera similar a lo visto en el caso continuo, solo que en este caso ya se cuenta con los valores de la función de probabilidad puntual y de distribución acumulada para los distintos valores de la variable.

```
Media <- sum(y*p)
print(Media)

Varianza <- sum(p*(y-Media)^2)
print(Varianza)

Desvio <- sqrt(Varianza)
print(Desvio)

moda <- y[which.max(p)]
print(moda)

mediana <- y[which(acum>=0.5)][1]
print(mediana)

Q1 <- y[which(acum>=0.25)][1]
print(Q1)

Q3 <- y[which(acum>=0.75)][1]
print(Q3)
```

3.7 Apéndice

3.7.1 Sobre la variancia de una variable aleatoria

En la Sección 3.3 se presentan algunos parámetros de una variable aleatoria, Y, como $E(Y)$ y $D(Y)$, entre otros. Este último se define como la raíz positiva de la variancia. En esta sección se define a la variancia, $V(Y)$, y se obtiene su expresión en función de $E(Y)$ (o μ_Y) y de $E(Y^2)$.

En el Capítulo 2 se presenta la siguiente expresión para la variancia muestral (pág. 85):

$$s^2 = \frac{\sum_{j=1}^{k}(y_j - \bar{y})^2 n_j}{n-1}$$

Si se multiplica a esta expresión por el cociente $\frac{n}{n}$, lo cual no altera la igualdad, queda:

$$s^2 = \frac{\sum_{j=1}^{k}(y_j - \bar{y})^2 n_j}{n-1}\frac{n}{n}$$

Trabajando y reordenando, resulta:

$$s^2 = \frac{n}{n-1}\sum_{j=1}^{k}(y_j - \bar{y})^2 \frac{n_j}{n} = \frac{n}{n-1}\sum_{j=1}^{k}(y_j - \bar{y})^2 f_j$$

Si el tamaño de muestra, n, tiende a infinito, la frecuencia relativa muestral tiende a la frecuencia relativa poblacional o probabilidad, la media muestral tiende a la media poblacional y el cociente $\frac{n}{n-1}$ tiende a 1. Queda entonces la expresión de la variancia poblacional para el caso discreto. Para el caso continuo, se define de manera análoga considerando la función de densidad de probabilidad.

- Cuando Y es una variable aleatoria continua, $V(Y) = \int_{y\in R_Y}(y-\mu)^2 f_Y(y)dy$;
- cuando Y es una variable aleatoria discreta, $V(Y) = \sum_{y\in R_Y}(y-\mu)^2 p_Y(y)$

A continuación, se busca obtener la expresión de la variancia poblacional en función de $E(Y)$ y de $E(Y^2)$.

Para ello se recuerda la definición de $E(Y)$, esperanza o promedio poblacional de la variable aleatoria Y. También se la simboliza μ_Y.

- Cuando Y es una variable aleatoria continua, $E(Y) = \mu_Y = \int_{y\in R_Y} y\, f_Y(y)dy$;
- cuando Y es una variable aleatoria discreta, $E(Y) = \mu_Y = \sum_{y\in R_Y} y\, p_Y(y)$.

Análogamente, se puede definir la esperanza de cualquier función de la variable aleatoria Y, $H(Y)$:

- Cuando Y es una variable aleatoria continua, $E[H(Y)] = \int_{y\in R_Y} h(y)\, f_Y(y)dy$;
- cuando Y es una variable aleatoria discreta, $E[H(Y)] = \sum_{y\in R_Y} h(y)\, p_Y(y)$.

En particular, $V(Y)$, se puede definir como el promedio poblacional o esperanza de los desvíos cuadrados de los valores de la variable Y respecto de su media (μ_Y o $E(Y)$). Es decir, se puede escribir de la siguiente manera:

$$V(Y) = E(Y-\mu_Y)^2 = E(Y-E(Y))^2$$

.

Si se define $H(Y) = (Y-E(Y))^2$, la variancia poblacional se puede expresar como:

- cuando Y es una variable aleatoria continua, $V(Y) = E(Y-E(Y))^2 = \int_{y\in R_Y}(y-E(Y))^2 f_Y(y)dy$;
- cuando Y es una variable aleatoria discreta, $V(Y) = E(Y-E(Y))^2 = \sum_{y\in R_Y}(y-E(Y))^2 p_Y(y)$.

A continuación se trabaja algebraicamente para obtener una expresión más simple de la variancia poblacional, para el caso de una variable aleatoria continua. Análogamente se puede obtener la misma expresión para el caso de una variable aleatoria discreta, reemplazando la integral por la sumatoria.

$E(Y-E(Y))^2 = \int_{y\in R_Y}(y-E(Y))^2 f_Y(y)dy =$

$\int_{y\in R_Y}(y^2 - 2yE(Y) + E(Y)^2) f_Y(y)dy =$

$\int_{y\in R_Y} y^2 f_Y(y)dy - 2E(Y)\int_{y\in R_Y} y f_Y(y)dy + E(Y)^2 \int_{y\in R_Y} f_Y(y)dy =$

$E(Y^2) - 2\cdot E(Y)E(Y) + E(Y)^2 = E(Y^2) - E(Y)^2 = E(Y^2) - \mu^2$



Recuerde que:

- $E(Y) = \mu$ es una constante (y por ello puede sacarse fuera de la integral o de la sumatoria si fuera una variable aleatoria discreta).
- $\int_{y\in R_Y} f_Y(y)dy = 1$ por condición de cierre de una distribución de probabilidades.
- Y^2 es una función de Y: $H(Y) = Y^2$ y por lo tanto, $\int_{y\in R_Y} y^2 f_Y(y)dy = E(Y^2)$.

3.7.2 Sobre la esperanza y la variancia de una variable aleatoria X, función lineal de Y

Suponga que a partir de una variable aleatoria Y se le aplica una transformación lineal tal que $X = a+bY$, con a y b pertenecientes a los números reales.

En la Sección 3.3.3 se presentan expresiones para obtener el valor de $E(X)$ y de $D(X)$, conocidos los valores de la media y el desvío estándar poblacional de Y, $E(Y)$ y $D(Y)$. En esta sección se deducen estas fórmulas para el caso de variables aleatorias continas. Análogamente se pueden obtener las expresiones para el caso de variables aleatorias discretas, reemplazando la integral por la sumatoria.

$$X = H(Y) = a+bY$$

Por lo visto en la Sección 3.7.1, $E(X) = E(H(Y)) = E(a+bY) = \int_{y\in R_Y}(a+by) f_Y(y)dy =$

$\int_{y\in R_Y} a f_Y(y)dy + \int_{y\in R_Y} by f_Y(y)dy = a\int_{y\in R_Y} f_Y(y)dy + b\int_{y\in R_Y} y f_Y(y)dy = a+bE(Y)$.

Es decir,

$$E(X) = E(a+bY) = a+bE(Y).$$

Recuerde que:

- $\int_{y \in R_Y} f_Y(y)dy = 1$ por condición de cierre de una distribución de probabilidades.
- $\int_{y \in R_Y} y\, f_Y(y)dy = E(Y)$.

Por lo visto en la Sección 3.7.1, $V(X) = E(X^2) - E(X)^2$.

- $E(X^2) = E(a+bY)^2 = E(a^2 + 2abY + b^2Y^2) =$
  $\int_{y \in R_Y} (a^2 + 2aby + b^2y^2)\, f_Y(y)dy =$
  $a^2 \int_{y \in R_Y} f_Y(y)dy + 2ab \int_{y \in R_Y} y\, f_Y(y)dy + b^2 \int_{y \in R_Y} y^2 f_Y(y)dy =$
  $a^2 + 2abE(Y) + b^2E(Y^2)$
- $E(X) = a + bE(Y)$, de donde:
  $E(X)^2 = [a + bE(Y)]^2 = a^2 + 2abE(Y) + b^2E(Y)^2$

Recuerde que:

- $\int_{y \in R_Y} f_Y(y)dy = 1$ por condición de cierre de una distribución de probabilidades.
- $\int_{y \in R_Y} y\, f_Y(y)dy = E(Y)$
- $\int_{y \in R_Y} y^2 f_Y(y)dy = E(Y^2)$

Habiendo obtenido separadamente las expresiones de $E(X^2)$ y de $E(X)^2$, a partir de los valores de $E(Y)$ y de $E(Y^2)$, a continuación se obtiene la expresión de $V(X)$.

$V(X) = E(X^2) - E(X)^2 =$

$[a^2 + 2abE(Y) + b^2E(Y^2)]$ - $[a^2 + 2abE(Y) + b^2E(Y)^2] =$

$b^2E(Y^2) - b^2E(Y)^2 =$

$b^2[E(Y)^2 - E(Y)^2] = b^2V(Y)$.

Es decir:

$$V(X) = V(a+bY) = b^2V(Y)$$

.

$D(X)$ se define como la raíz cuadrada positiva de $V(X)$.
Dado que $V(X) = b^2V(Y)$, $D(X) = |b|D(Y)$.

En resumen, definida $X = a + bY$, con a y b pertenecientes a los números reales:

- $E(X) = a + bE(Y)$;
- $V(X) = b^2V(Y)$;
- $D(X) = |b|D(Y)$.



4. Experimentos aleatorios y probabilidad

4.1 Introducción

En el Capítulo 3 se presenta la noción de experimento aleatorio y se define a la probabilidad como la frecuencia relativa de valores de la variable en la población, o proporción poblacional de unidades con ciertos valores de la variable.

Por ejemplo, suponga que para la población de barras de acero de la empresa metalúrgica se obtiene la probabilidad de que la longitud esté comprendida entre 249,4 mm y 250,6 mm y esta resulta igual a 0,84. Se puede decir, entonces, que para un gran número de barras, aproximadamente el 84 % de las mismas tiene longitud mayor a 249,4 mm y menor a 250,6 mm. Ahora bien, si se selecciona una barra al azar de esa población, ¿cuál es la "chance" de que esta barra mida más que 249,4 mm y menos que 250,6 mm? Claramente la respuesta es 0,84.

A partir de estas ideas, en este capítulo se desarrollan las herramientas para el estudio de los fenómenos aleatorios y la cuantificación de la incertidumbre mediante probabilidades.

Los objetivos de este capítulo son:

- Presentar las definiciones de experimento aleatorio; espacio muestral y suceso, para luego definir formalmente a la probabilidad.
- Brindar algunos enfoques que permiten asignar probabilidades a los diversos sucesos.
- Introducir los conceptos de probabilidad condicional y de independencia entre sucesos.
- Presentar una aplicación del Teorema de la Probabilidad Total.

4.2 Algunas definiciones preliminares

Se considera **experimento aleatorio** al proceso de seleccionar una unidad al azar de una población determinada y observar en ella una variable de interés.

Un **experimento aleatorio** es uno no determinista, es decir, uno en el que no puede determinarse a priori cuál será el resultado de una realización particular, y que cumple además con las siguientes características:

- Se puede establecer el conjunto de sus posibles resultados.
- Se puede repetir una gran cantidad de veces en igualdad de condiciones.

A continuación se presentan algunos ejemplos de experimentos aleatorios, relacionados con los problemas presentados en la Sección 1.4 (página 7).



Problema 4 (pág. 8): *En este problema la población está compuesta por todas las devoluciones de productos vendidos en línea por la empresa. Un experimento aleatorio consiste en seleccionar al azar una devolución y observar el motivo principal de la misma.*



Problema 2 (pág. 7): *En este problema la población está compuesta por todas las placas de madera de un lote. Un experimento aleatorio consiste en seleccionar una placa al azar y registrar la cantidad de defectos.*



Problema 1 (pág. 7): *En este problema la población está compuesta por todas las barras de cierto tipo que se producen (en condiciones similares) en la empresa metalúrgica. Un experimento aleatorio puede ser seleccionar al azar una barra de la producción y registrar su longitud, en mm.*



Problema 7 (pág. 8): *En este problema una de las poblaciones está compuesta por todas las mediciones que se pueden realizar de una resistencia conocida, con cierto equipo. Un experimento aleatorio puede ser realizar una medición y observar el valor obtenido de la resistencia, en* Ω.

Observe que en el caso del Problema 7, la población tiene existencia conceptual y la selección de un elemento de la población es en realidad la realización de una medición bajo ciertas condiciones.

Los experimentos mencionados anteriormente y otros similares tienen en común algunas características:

- Se realizan de acuerdo a reglas definidas.
- Los resultados varían de una repetición a otra de manera imprevisible.
- Los resultados individuales parecen ocurrir de forma arbitraria; sin embargo, cuando el experimento se repite un gran número de veces, aparece un modelo definido de regularidad. Por ejemplo, suponga que un experimento consiste en lanzar un dado equilibrado y observar el número que sale en la cara superior. Si este experimento se repite n veces, para valores de n convenientemente grandes, se observa que la proporción de unos es aproximadamente igual a $\frac{1}{6}$, o también que la proporción de números pares (o de impares) es igual a 0,5.

Se denomina **espacio muestral** al conjunto de todos los resultados posibles para un experimento aleatorio. Usualmente se lo simboliza con la letra S.

Cada uno de los elementos del espacio muestral recibe el nombre de **resultado** o **punto muestra**. Estos resultados verifican las siguientes propiedades:

- Siempre ocurre alguno de ellos al llevar a cabo el experimento.
- Son mutuamente excluyentes (la ocurrencia de uno implica la no ocurrencia de los demás).

En la Tabla 4.1 se ejemplifican estos conceptos para los problemas considerados.

Tabla 4.1 *Población, experimento aleatorio y espacio muestral para los problemas*

Problema	**Población**	**Experimento aleatorio (E)**	**Espacio muestral (S)**
Problema 1	Todas las barras de cierto tipo que produce una empresa metalúrgica.	E_1: Seleccionar al azar una barra y observar su longitud, en mm.	$S_1 = [a, b]$ Subconjunto de los números reales positivos.
Problema 2	Todas las placas de madera de un lote.	E_2: Seleccionar al azar una placa y observar la cantidad de defectos que posee.	$S_2 = \{0, 1, 2, ...\}$ Conjunto de los números enteros no negativos.
Problema 4	Todas las devoluciones de productos vendidos en línea.	E_4: Seleccionar al azar una devolución y registrar el motivo principal de la misma.	$S_4 =$ {error en la facturación, pedido equivocado, precio final no coincidente, producto dañado, producto mal embalado, producto vencido o en mal estado, retraso en la entrega, otro}
Problema 7	Todas las mediciones que se pueden realizar de una resistencia conocida, con cierto equipo.	E_7: Realizar una medición y registrar el resultado de la misma, en Ω.	$S_7 = [a, b]$ Subconjunto de los números reales positivos.

El subíndice utilizado en E y S se corresponde con el número de problema al que hacen referencia. S_4 está asociado a una variable cualitativa o atributo, S_2, a una cuantitativa discreta y S_1 y S_7 a una cuantitativa continua.

Observe que las cardinalidades de los espacios muestrales pueden ser variadas, más precisamente S_4 es finito, S_2 es infinito (numerable) y S_1 y S_7 son infinitos (no numerables).

Cualquier subconjunto del espacio muestral se denomina **suceso** o **evento**. Se suelen simbolizar con letras mayúsculas (A, B, C por ejemplo). Si un suceso está compuesto por un único elemento, se lo denomina **suceso elemental**, y en particular se lo simboliza $A = \{a\}$.

Observe lo siguiente en relación a las definiciones introducidas:

1. Todo suceso expresa una condición (por medio de una proposición), que el resultado del experimento puede verificar o no. Por lo tanto, un suceso se identifica con un subconjunto de S, el subconjunto formado por aquellos resultados que lo verifican.
2. El espacio muestral (S) se considera un suceso y recibe el nombre de **suceso seguro**, porque siempre se verifica una vez realizado el experimento (tener en cuenta que en S se encuentran todos los resultados posibles del mismo).
3. El conjunto vacío se considera un suceso y recibe el nombre de **suceso imposible** porque nunca se verifica.

Algunos sucesos de interés para los experimentos considerados podrían ser, a modo de ejemplo, los que se presentan en la Tabla 4.2.

Tabla 4.2 *Algunos sucesos, a modo de ejemplo*

Experimento aleatorio	Espacio muestral	Ejemplos de sucesos
E_1	S_1 = [a, b]	F (la barra seleccionada tiene longitud dentro de especificaciones) = $[249,4\ ;\ 250,6]$, con $F \subseteq S_1$ G (la barra seleccionada tiene longitud menor que 250 mm)= $[a\ ;\ 250)$, con $G \subseteq S_1$
E_2	$S_2 = \{0,1,2,....\}$	B (la placa elegida tiene a los sumo 2 defectos) = $\{0,1,2\}$ M(la placa elegida tiene 5 o más defectos) = $\{5,6,7,.....\}$
E_4	S_4 [1]	A (el motivo principal de la devolución seleccionada es el retraso en la entrega) = {retraso en la entrega} C (el motivo principal de la devolución seleccionada no es el error en la facturación) = {pedido equivocado, precio final no coincidente, producto dañado, producto mal embalado, producto vencido, retraso en la entrega, otro}
E_7	S_7 = [a, b]	H (la medición realizada da un valor menor que 100 Ω)= $[a\ ;\ 100)$, con $H \subseteq S_7$ J (la medición realizada da un valor mayor que 100,2 Ω)= $(100,2\ ;\ b]$, con $J \subseteq S_7$

Si se define una variable aleatoria Y asociada a un experimento aleatorio dado, el espacio muestral es el **recorrido**, R_Y, de dicha variable y los sucesos pueden expresarse en función de Y. Por ejemplo, $M = \{Y : Y \leq 2\}$, $N = \{Y : 1 \leq Y < 8\}$, etc.

[1] S_4 = { error en la facturación, pedido equivocado, precio final no coincidente, producto dañado, producto mal embalado, producto vencido o en mal estado, retraso en la entrega, otro}

Problema 1 (pág. 7): *En este problema se puede definir la variable aleatoria Y: longitud de una barra producida. Si se considera el suceso F (la barra seleccionada tiene longitud dentro de especificaciones), y recordando que la empresa automotriz exige que la longitud de las barras se encuentre en el intervalo* $(250 \pm 0,6)$ *mm, dicho suceso puede expresarse como* $F = [249,4; 250,6]$. *Otro suceso de interés puede ser G (la barra seleccionada tiene longitud menor que 250 mm), el cual puede expresarse como* $G = [a; 250)$.

Con lo visto hasta ahora se pueden resolver las Actividades 1 y 2, Sección 4.8.

Se introducen a continuación algunas otras definiciones, para sucesos A y B de un mismo espacio S, que permiten generar otros sucesos del mismo espacio muestral:

Siendo A y B sucesos de un mismo espacio muestral S, se define:

- El **complemento** de un suceso A está formado por todos los elementos de S que no pertenecen a A. Se suele simbolizar $\bar{A}$.
  Se dice que $\bar{A}$ ocurre (o, equivalentemente, que A no ocurre) cuando al realizar el experimento aleatorio se obtiene un resultado de $\bar{A}$, es decir un resultado que no está en A.
- La **unión** de dos sucesos A y B está formada por todos los elementos de S que pertenecen a A, a B o a ambos. Se simboliza $A \cup B$.
  Se dice que $A \cup B$ ocurre cuando al realizar el experimento aleatorio, se obtiene un resultado de A que no está en B, uno de B que no está en A o uno que está tanto en A como en B.
- La **intersección** de dos sucesos A y B está formada por todos los elementos de S que pertenecen a ambos conjuntos. Se simboliza $A \cap B$.
  Se dice que $A \cap B$ ocurre cuando al realizar el experimento aleatorio, se obtiene un resultado que está en A como así también en B.

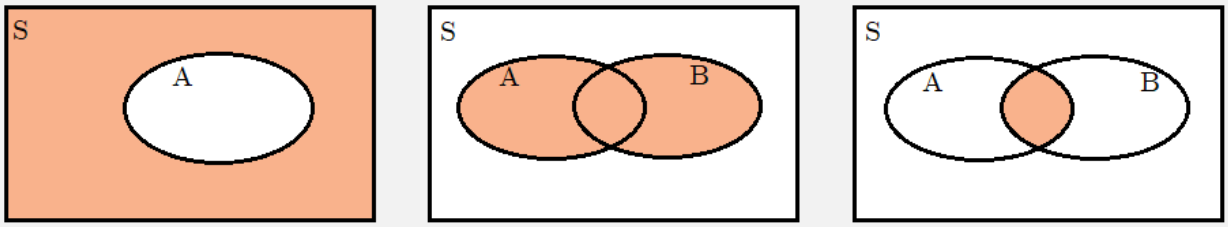


Figura 4.1 *Representación de los sucesos* $\bar{A}$, $A \cup B$ *y* $A \cap B$, *respectivamente*

- Dos sucesos A y B se dicen **mutuamente excluyentes** si carecen de resultados en común, es decir, si su intersección es el conjunto vacío. Cuando A y B son mutuamente excluyentes, $A \cap B$ es el suceso imposible y se dice que A y B no pueden ocurrir a la vez. En particular, los sucesos A y su complemento, son mutuamente excluyentes.

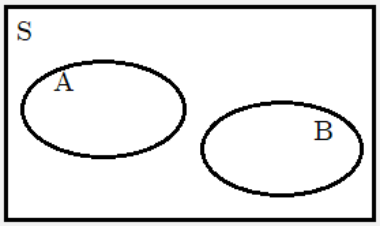


Figura 4.2 *Representación del suceso imposible como la intersección vacía entre A y B. En símbolos:* $A \cap B = \emptyset$

Problema 2 (pág. 7): *Reconsidere este problema, en relación al experimento aleatorio de seleccionar una placa de madera de un determinado lote y observar la cantidad de defectos que esta posee.*

Se consideran buenas las placas con a lo sumo 2 defectos, regulares las que tienen 3 o 4 defectos, y malas las restantes. Si Y es la variable aleatoria "cantidad de defectos en la placa seleccionada", se definen entonces, sobre el espacio muestral de todas las placas (o lo que es lo mismo sobre R_Y*), los siguientes sucesos:*

- $B: Y \leq 2$ *(la placa elegida al azar es buena, es decir, tiene a lo sumo 2 defectos).*
- $R: 3 \leq Y \leq 4$ *(la placa elegida al azar es regular, es decir, tiene 3 o 4 defectos).*
- $M: Y \geq 5$ *(la placa elegida al azar es mala, es decir, tiene 5 o más defectos).*

Claramente los sucesos B, R y M son mutuamente excluyentes dos a dos, es decir, los sucesos $B \cap R$, $B \cap M$ *y* $R \cap M$ *son todos iguales al suceso imposible.*

Si por ejemplo una placa va a ser descartada cuando tiene más de 2 defectos, también puede interesar pensar en el suceso $\{Y > 2\}$*. Note que por ejemplo se puede expresar a este suceso como* $\bar{B}$*, o también como* $R \cup M$*. En otras palabras, las placas que no son buenas son aquellas regulares o malas.*

Se define ahora el siguiente suceso sobre R_Y*:*

$D: 1 \leq Y \leq 4$ *(la placa elegida al azar presenta 1, 2, 3 o 4 defectos).*

Observe que D y B no son mutuamente excluyentes, ya que su intersección no coincide con el conjunto vacío (y lo mismo vale para los sucesos D y R). En efecto, $D \cap B$ *es el suceso* $1 \leq X \leq 2$ *(la placa elegida al azar tiene 1 o 2 defectos).*

El suceso $D \cap B$ *hace referencia al cumplimiento simultáneo de la condición definida por el suceso D y de la condición definida por el suceso B.*

Por otro lado, el suceso $D \cup B$*:* $Y \leq 4$ *hace referencia a que la placa elegida al azar tenga como máximo 4 defectos.*

El suceso $D \cup B$ *hace referencia al cumplimiento de al menos una de las dos condiciones: cumplen con la condición definida por el suceso D (la placa tiene más de 1 defecto pero no más de 4) o con la condición definida por el suceso B (la placa tiene como máximo 2 defectos) o con ambas condiciones.*

Observe que $R \subseteq D$*, de modo que el suceso* $D \cap R$ *coincide con el suceso R y el suceso* $D \cup R$ *coincide con el suceso D.*

4.3 La probabilidad de un suceso

Considere un experimento aleatorio cualquiera y un suceso A (subconjunto de su espacio muestral S). Se pretende definir un número asociado a A, $P(A)$, como una medida de la chance o posibilidad de que este suceso se verifique (u ocurra) en una realización particular del experimento. Suele decirse que la probabilidad cuantifica o mide la incertidumbre respecto de que ocurra un suceso al realizar el experimento aleatorio.

4.3.1 ¿Cómo se pueden asignar probabilidades?

Las probabilidades de los sucesos se pueden determinar de la siguiente manera:

1. **Encontrando, a partir de la naturaleza del experimento, relaciones que permitan determinar las probabilidades de cada uno de los elementos del espacio muestral**

El caso más simple es el de la equiprobabilidad, que se da cuando el experimento aleatorio tiene asociado un espacio muestral finito, con resultados elementales igualmente probables.

Si en el espacio muestral existen k sucesos elementales equiprobables, la probabilidad de cada uno de ellos será entonces $\frac{1}{k}$, para asegurar que la suma total resulte 1. Luego, la probabilidad de un suceso compuesto A (que contiene c sucesos elementales) será igual a $\frac{c}{k}$, lo que da lugar a la regla:

$$P(A) = \frac{\text{número de resultados (o casos) favorables al suceso A}}{\text{número de casos o resultados posibles del experimento aleatorio}} = \frac{c}{k}.$$

Esta regla, referenciada como *definición clásica de probabilidad*, solo debe utilizarse cuando la equiprobabilidad de los resultados elementales esté confirmada por el mecanismo generador de los mismos o por la evidencia empírica. Históricamente, esta fue la primera definición de probabilidad utilizada, en relación a los juegos de azar, cuyo estudio dio lugar al nacimiento del Cálculo de Probabilidades.

Ejemplo 4.1 Se pretende realizar una inspección de maquinarias en una fábrica que cuenta con seis máquinas del mismo tipo (numeradas del 1 al 6). Como resulta demasiado costoso inspeccionar a todas ellas, se elegirá una al azar y se llevará a cabo la inspección. Solo la fábrica tiene el conocimiento de que las máquinas 2 y 5 están momentáneamente funcionando mal. Se pretende entonces averiguar qué chance se tiene de que la máquina elegida sea una de estas dos.

Un espacio muestral asociado a este experiento aleatorio de elegir una máquina al azar entre las seis y observar cuál de ellas se inspeccionará, es uno finito con seis resultados posibles, es decir:

$$S = \{\text{máquina } 1, \ldots, \text{máquina } 6\}.$$

Como se supone una elección al azar, los 6 sucesos elementales son equiprobables, es decir $P(\text{sale máquina } 1) = P(\text{sale máquina } 2) = P(\text{sale máquina } 3) = P(\text{sale máquina } 4) = P(\text{sale máquina } 5) = P(\text{sale máquina } 6)$.

Además, $P(\text{sale máquina 1}) + \cdots + P(\text{sale máquina 6}) = 1$.

Luego,

$$P(\text{sale máquina 1}) = \cdots = P(\text{sale máquina 6}) = \frac{1}{6}.$$

Entonces por ejemplo, la probabilidad del suceso de interés A: sale la máquina 2 o la 5, se puede calcular de la siguiente manera:

$$P(A) = P(\text{sale máquina 2 ó sale máquina 5 }) = \frac{1}{6} + \frac{1}{6}.$$

Note que A puede pensarse como un suceso compuesto por la unión de dos sucesos elementales, y por lo tanto mutuamente excluyentes.

2. Conociendo (al menos aproximadamente) la ley de probabilidad

Si se conoce la distribución de probabilidades de la variable aleatoria en cuestión (función de probabilidad puntual o función de densidad de probabilidad, según corresponda), o la distribución de frecuencias relativas de un atributo o variable categórica en la población, se las puede utilizar para obtener probabilidades.



Problema 2 (pág. 7): *Suponga que se cuenta ahora con la distribución de frecuencias de la cantidad de defectos por placa, para un lote en particular (Lote 1), esta es:*

Tabla 4.3 *Distribución de frecuencias de la cantidad de defectos por placa, para el Lote 1*

Cant. de defectos por placa	*0*	*1*	*2*	*3*	*4*	*5 ó más*
Cant. de placas	*10*	*29*	*26*	*10*	*4*	*1*

Como se cuenta con información de toda la población, se puede construir la distribución de probabilidades de Y: cantidad de defectos por placa.

Tabla 4.4 *Distribución de probabilidades de la cantidad de defectos por placa para el Lote 1*

Cant. de defectos por placa	*0*	*1*	*2*	*3*	*4*	*5 ó más*
Probabilidad puntual	*0,125*	*0,3625*	*0,325*	*0,125*	*0,05*	*0,0125*

Considere el experimento aleatorio E_2, definido en la Tabla 4.1, que consiste en extraer una placa al azar del Lote 1 y observar la cantidad de defectos que tiene. Un espacio muestral asociado a esta experiencia es el recorrido de la variable Y, es decir $S_2 = R_Y = \{0,1,2,\cdots\}$. A partir de la función de probabilidad puntual presentada en la tabla, se pueden obtener probabilidades de diferentes sucesos como por ejemplo:

- $P(Y \geq 1) = 1 - P(Y = 0) = 1 - 0,125 = 0,875$. *Esta cantidad representa la probabilidad de que la placa elegida tenga al menos un defecto. También representa la proporción poblacional de placas que tienen al menos un defecto en el Lote 1.*
- $P(Y \leq 2) = P(Y = 0) + P(Y = 1) + P(Y = 2) = 0,125 + 0,3625 + 0,325 = 0,8125$. *Esta cantidad representa la probabilidad de que la placa elegida tenga no más de dos defectos. También, representa la proporción poblacional de placas que tienen como máximo dos defectos en el Lote 1.*

3. Estudiando las frecuencias relativas al repetir el experimento en condiciones similares

Este método solo es factible en ocasiones en que es posible una experimentación continuada. Esta noción ya ha sido introducida y analizada en la Sección 3.2.1 del Capítulo 3.

Informalmente se puede decir que las probabilidades se obtienen a partir de las frecuencias relativas de los sucesos de interés, para un número grande de repeticiones (o ensayos) de la experiencia.

Denotando con n al número de repeticiones de la experiencia, y con $n(A)$ al número de veces que ocurre el evento A entre esas n repeticiones, resulta:

$$P(A) = lim_{n\to\infty}\frac{n(A)}{n} = lim_{n\to\infty} f(A).$$

En general, si se desea obtener la probabilidad de un determinado suceso, se debería repetir el experimento, pero es entendible que no se puede llevar a la práctica la idea de considerar un número **infinito** de repeticiones, sino un número suficientemente grande factible en cada problema. Se puede afirmar en general que con un número finito de repeticiones es posible hacer una "estimación" de la probabilidad de un evento y valorar la aproximación a ese nivel. De esta manera se tiene una estimación de la probabilidad buscada, que será mejor o peor dependiendo de cuántas repeticiones se realicen.

Surgen algunas preguntas como: ¿cuántas repeticiones se deben realizar?, ¿si aumenta el número de repeticiones, se obtendrá una mejor aproximación al valor de la probabilidad? Esto se va a profundizar en capítulos posteriores cuando se aborde el estudio de la Inferencia Estadística.



Problema 7 (pág. 8): *En esta situación, puede considerarse que una de las poblaciones de interés está compuesta por todas las mediciones de determinada resistencia que se podrían realizar con el multímetro. Recuerde el experimento aleatorio E_7, definido en la Tabla 4.1, que consiste en realizar una medición de la resistencia y observar el resultado de la misma.*

Ahora bien, se puede imaginar que se realiza un número relativamente grande (por ejemplo 100, aunque en general no es posible conseguir realizar un número demasiado grande de repeticiones) de mediciones de esa resistencia, en Ω. Si se considera al suceso A (la medición realizada se encuentra entre dos valores de interés, a y b) y se corrobora que de las 100 mediciones, 82 se encuentran entre esos dos valores, se puede considerar que $\frac{82}{100} = 0,82$ constituye una "estimación" de $P(A)$ (probabilidad de que una medición realizada se encuentre entre esos valores). Se retomará este ejemplo en el Capítulo 8, cuando esta estimación se interprete en el contexto de la inferencia estadística.

En algunas otras aplicaciones es difícil imaginar la repetición del experimento. Por ejemplo, cuando un gerente calcula la probabilidad de que un nuevo producto tenga éxito en el mercado, el producto solo puede ser introducido una única vez.

¿Cuál es entonces el significado que debe dársele a la probabilidad en estos casos? Si se dice que la probabilidad de que el producto tenga éxito es 0,9, esta es una valoración **personal o subjetiva** ya que es una cuestión de opinión. Más aún, dos personas pueden pensar en valoraciones distintas y asignar diferentes probabilidades al mismo resultado. Tales valoraciones también pueden observarse en proyecciones económicas que realizan los economistas para años venideros, en el análisis de inversiones, en el estudio de finanzas, etc.



A esta altura se puede ver entonces que ya se ha asignado un valor a $P(A)$, probabilidad de que el suceso A ocurra. La pregunta que surge es: ¿cómo se puede interpretar esto?

Se lo puede hacer de dos maneras:

1. como una medida de la chance o posibilidad de que A se verifique (u ocurra) en una realización particular del experimento,
2. como la proporción o frecuencia relativa de elementos de la población que tienen valores (o niveles) de la característica en estudio según se enuncian en A. Esta interpretación de la probabilidad ya se consideró en el Capítulo 3, para el caso de variables aleatorias continuas y discretas.

4.3.2 La probabilidad en la toma de decisiones

La probabilidad cumple un rol especial en la vida cotidiana porque se la utiliza para medir la incertidumbre y para tomar decisiones de diferente índole. Suponga que una persona debe viajar a una zona determinada en el mes de enero y se conoce que en esa zona hay una alta probabilidad de lluvias durante ese mes. ¿Qué precauciones deberá tomar? Seguramente incluirá botas de lluvia e impermeable en su valija, entre otras cosas.

En la resolución de problemas ingenieriles de naturaleza estadística, muchas decisiones se deben tomar en contextos de incertidumbre y la probabilidad resulta de gran ayuda.

Considere las siguientes situaciones:

- Suponga que una empresa cuenta con dos posibles proveedores de un mismo insumo, clave para cierto proceso de producción, el cual se comercializa en tanques de 100 litros. Suponga además que para ambos proveedores, el precio de venta y las condiciones de pago y plazos de entrega son similares; pero se sabe que la probabilidad que un tanque no cumpla con los requerimientos de calidad para uno de los proveedores es 0,15 y para el otro es 0,03. ¿A qué proveedor seleccionarán en la empresa?
- En una distribuidora de productos electrónicos de la ciudad de Rosario se conoce, por estudios previos, que la probabilidad de que la demanda semanal de cierto producto supere las 6 unidades es 0,00046. Usted, ¿recomendaría tener más de 6 artículos en stock?
- Una selectora de personal analiza los antecedentes, las entrevistas y otra información de cierto postulante para un puesto y considera que esta persona tiene una alta probabilidad de desempeñarse bien en la función que debe cumplir. ¿Cree Ud. que la recomendará para dicho puesto?

En la primera situación es razonable elegir como proveedor al que tiene la menor probabilidad de que un tanque no cumpla con los requerimientos [2]; mientras que en la segunda situación no se recomendaría tener más de 6 artículos en stock (especialmente si esto es costoso para la empresa), ya que es muy poco probable que la demanda supere las 6 unidades. En la tercera situación, el hecho de que sea muy

[2] Observe que además de la probabilidad, en este caso se consideró otra información para decidir, como los costos, los plazos de entrega,etc.

alta la probabilidad de que el postulante se desempeñe bien en el puesto, hace que en la selectora lo recomienden para el mismo.

Independientemente de cómo se haya obtenido el valor de la probabilidad, según cuál sea la decisión a tomar, algunas veces es razonable optar por la alternativa asociada a una probabilidad baja de ocurrencia (como el caso de la elección del proveedor) y otras veces es razonable la alternativa asociada a una probabilidad alta (como el caso de la selección del postulante).

La probabilidad resulta clave, además, cuando en el marco de la resolución de un problema ingenieril de naturaleza estadística se decide seleccionar una muestra aleatoria de la población de interés. En primer lugar, cada unidad de la población tiene cierta probabilidad de ser seleccionada. Luego, una vez analizados los datos, existe una cierta probabilidad de que las conclusiones que se obtengan para la población no sean correctas, ya que para obtenerlas no se estudió a la totalidad de unidades de la población. En los procedimientos de inferencia estadística se busca que esta probabilidad sea lo más baja posible. Estas cuestiones se abordan en los Capítulos 8 y 9, relativos a la inferencia estadística.

En muchas situaciones, se utiliza la expresión "riesgo" para hacer referencia a la probabilidad de un evento que tiene consecuencias negativas. Por ejemplo, volviendo a la situación de la distribuidora, la probabilidad de que la demanda semanal de cierto producto supere las 6 unidades es 0,00046. Si en la empresa siempre tienen 6 productos en stock al inicio de cada semana, esa probabilidad es el riesgo de que en una semana cualquiera, la demanda no sea satisfecha.

En las herramientas de inferencia, las probabilidades de error también se pueden denominar riesgos de obtener inferencias erróneas.

4.4 Probabilidad. Definición axiomática. Algunas consecuencias

Se define a la probabilidad sobre un espacio muestral S a través de las siguientes propiedades (comúnmente denominadas axiomas), inspiradas en las propiedades de las frecuencias relativas:

Axiomas de la probabilidad de un suceso

1. $P(A) \geq 0$ para cualquier suceso A.
2. $P(S) = 1$.
3. Si A y B son sucesos mutuamente excluyentes y se considera la unión de ellos llamando C al nuevo suceso (es decir $C = A \cup B$), resulta $[P(C) = P(A \cup B) = P(A) + P(B).]$

A partir de estos axiomas se puede desprender el cálculo de las probabilidades de otros sucesos (su demostración puede verse en el Apéndice (Sección 4.9)):

1. La probabilidad del suceso complemento:

$$P(\bar{A}) = 1 - P(A).$$

2. La probabilidad de la unión de dos sucesos cualesquiera (es decir, no necesariamente excluyentes):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Problema 2 (pág. 7): *Reconsidere este problema, donde se contabilizaron los defectos de ciertas placas en un lote. Considere los siguientes sucesos, ya definidos:*

- $B:\ Y \leq 2$ *(una placa elegida al azar tiene a lo sumo 2 defectos),*
- $R:\ 3 \leq Y \leq 4$ *(una placa elegida al azar tiene 3 o 4 defectos),*
- $M:\ Y \geq 5$ *(una placa elegida al azar tiene 5 o más defectos)y*
- $D:\ 1 \leq Y \leq 4$ *(una placa elegida al azar tiene 1, 2 3 o 4 defectos).*

Entre otras, se pueden obtener las siguientes probabilidades:

1. *La probabilidad de que una placa elegida al azar resulte buena:* $P(B) = \frac{65}{80} = 0,8125$.
2. *La probabilidad de que una placa elegida al azar resulte regular:* $P(R) = \frac{14}{80} = 0,175$.
3. *La probabilidad de que una placa al azar resulte mala:* $P(M) = \frac{1}{80} = 0,0125$.
4. *Como R y M son mutuamente excluyentes (más aún, es claro que B, R y M son mutuamente excluyentes dos a dos), la probabilidad de que una placa elegida al azar no resulte buena se puede calcular de la siguiente manera:* $[P(R \cup M) = P(R) + P(M) = \frac{14}{80} + \frac{1}{80} = \frac{15}{80} = 0,1875]$
   o, equivalentemente como $1 - P(B)$*, ya que* $R \cup M = \bar{B}$.
5. *La probabilidad de que una placa elegida al azar tenga 1 o más defectos pero no más que 4:* $P(D) = \frac{69}{80} = 0,8625$.

Note que D y B no son mutuamente excluyentes pues $D \cap B = \{1,2\}$ *y se tiene que* $P(B \cap D) = \frac{55}{80} = 0,6875$. *Entonces la probabilidad de que se presente al menos uno de estos sucesos se obtiene de la siguiente manera:*

$$P(D \cup B) = P(D) + P(B) - P(B \cap D) = \frac{69}{80} + \frac{65}{80} - \frac{55}{80} = \frac{79}{80} = 0,9875.$$

Con lo visto hasta ahora se pueden resolver las Actividades 3 a 5, Sección 4.8.

4.5 Probabilidad condicional. Independencia entre sucesos

Problema 2 (pág. 7): *Suponga que se cuenta con dos lotes (Lotes 1 y 2) y que las placas de ambos se combinan en un solo lote de 160 unidades. Suponga también que de cada una de estas placas se puede conocer el lote de donde proviene y su calidad.*

La información sobre la calidad de las placas está resumida en la Tabla 4.5, de acuerdo al lote:

Tabla 4.5 *Distribución de frecuencias de la calidad de las placas, por lote*

Lote	***Calidad de la placa***		
	Buena (suceso B)	***Regular (suceso R)***	***Mala (suceso M)***
Lote 1 (suceso L_1)	*65*	*14*	*1*
Lote 2 (suceso L_2)	*68*	*10*	*2*

La información del Lote 1 se tomó de la Tabla 4.3.
Se selecciona una placa al azar y se observa que proviene del Lote 1.

Es claro que el conocimiento de que la placa proviene del Lote 1 lleva a tener en cuenta solo al subconjunto de las 80 placas del Lote 1, y entonces se puede aventurar que el resultado de si es buena por ejemplo, estará sujeto a esta "reducción". Más precisamente, se dice que la ocurrencia del suceso B está condicionada a la ocurrencia del suceso L_1.

La probabilidad de que la placa extraída, de la que se tiene el conocimiento de que proviene del Lote 1, sea buena es igual a $\frac{65}{80}$*; equivalentemente es igual a*

$$\frac{65}{80} = \frac{65/160}{80/160}.$$

Note que la fracción en el numerador representa a la probabilidad de que la placa "sea buena y pertenezca al Lote 1 *en relación a la totalidad (las 160)" y por otro lado, que el denominador representa la probabilidad de que la placa "provenga del Lote 1 en relación a la totalidad (las 160)".*

Más formalmente y en un contexto general más allá de este ejemplo, se introduce la siguiente definición:

La probabilidad de un suceso A condicionada a otro suceso B del mismo espacio muestral, denotada por $P(A|B)$ y denominada **probabilidad condicional**, es la probabilidad del suceso A condicionada a la ocurrencia del suceso B (es decir, en un espacio muestral restringido por el suceso B):

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ para } P(B) \neq 0.$$



Problema 2 (pág. 7): *Utilizando la notación presentada anteriormente, la probabilidad obtenida se simboliza:*

$$P(B|L_1) = \frac{P(B \cap L_1)}{P(L_1)} = \frac{65/160}{80/160} = 0,8125.$$

¿Cómo se interpreta esta probabilidad?

- *Si se selecciona una placa del Lote 1, la chance de que esta resulte buena es 0,8125.*
- *Pensando en todas las placas del Lote 1, la proporción de placas buenas es 0,8125.*



La probabilidad condicional se interpreta como la proporción poblacional de unidades que cumplen con cierta condición (en el ejemplo, son placas buenas), pero no en la población total sino en una subpoblación definida por el suceso "condicionante" (en el ejemplo, esa subpoblación está integrada por las placas del Lote 1).



Problema 1 (pág. 7): *Como se mencionó en la situación planteada en la página 76, en la empresa metalúrgica, además de las barras que interesan a la industria automotriz (barras tipo A), se producen barras roscadas (barras tipo B) y barras laminadas (barras tipo C). Suponga que, por experiencia, se conoce*

que el % de barras para cada tipo que no cumplen con las especificaciones en relación a la longitud (barras no conformes) es 3 %, 5 % y 6 % respectivamente. ¿Se puede expresar la información brindada en función de probabilidades condicionales?
Se definen los siguientes sucesos:

- *A: la barra es de tipo A,*
- *B: la barra es de tipo B,*
- *C: la barra es de tipo C y*
- *D: la barra es no conforme.*

Si se piensa en la población formada por la totalidad de las barras que se producen en la empresa, en particular, los sucesos A, B y C, definen subpoblaciones de dicha población.
La información brindada se puede escribir formalmente utilizando probabilidades condicionales, de la siguiente manera:

- *P(D|A) = 0,03 que indica que en la subpoblación de barras de tipo A, aproximadamente el 3 % de las mismas es no conforme.*
- *P(D|B) = 0,05 que indica que en la subpoblación de barras de tipo B, aproximadamente el 5 % de las mismas es no conforme.*
- *P(D|C) = 0,06 que indica que en la subpoblación de barras de tipo C, aproximadamente el 6 % de las mismas es no conforme.*

Es decir, se utilizan las probabilidades condicionales para reflejar proporciones dentro de subpoblaciones particulares, en este caso, definidas por el tipo de barra.

De la definición de probabilidad condicional introducida más arriba, es claro que se puede "despejar" el valor de $P(A \cap B)$, siendo este igual a $P(A|B)\,P(B)$, para B tal que $P(B) \neq 0$.

Una consecuencia inmediata de la definición de independencia entre sucesos que se verá a continuación indica que la probabilidad de la ocurrencia conjunta (a la vez) de dos sucesos cualesquiera es igual al producto de las probabilidades individuales de estos sucesos.

Se dice que A y B son **independientes** si la ocurrencia de uno de ellos no modifica las probabilidades de ocurrencia del otro.

Es decir, si $P(A|B) = P(A)$ (ó $P(B|A) = P(B)$), ya que si se da una de estas igualdades, se verifica también la otra).

Cuando se verifica la independencia entre A y B, resulta $P(A \cap B) = P(A)\,P(B)$.

Cuando dos sucesos A y B son **independientes entre sí**, también lo son A y $\bar{B}$, $\bar{A}$ y B y $\bar{A}$ y $\bar{B}$. Por ende resulta:
$P(A \cap \bar{B}) = P(A)\,P(\bar{B})$;
$P(\bar{A} \cap B) = P(\bar{A})\,P(B)$ y
$P(\bar{A} \cap \bar{B}) = P(\bar{A})\,P(\bar{B})$.

El siguiente es un ejemplo que muestra la forma de analizar la independencia de ciertos sucesos a partir de la verificación de la propiedad (definición) introducida recién:

Problema 2 (pág. 7): *Continuando con el ejemplo de las placas de dos lotes, se definen los siguientes sucesos:*

L_1: la placa elegida pertenece al Lote 1, y L_2: la placa elegida pertenece al Lote 2.
Claramente, $P(L_1) = P(L_2) = 0,5$.

Se puede analizar por ejemplo si los sucesos L_1 y B son independientes. Para ello se puede comparar $P(B \cap L_1)$ con $P(B)\,P(L_1)$.

Como se cuenta con la correspondiente información en una tabla de doble entrada, para calcular $P(B \cap L_1)$ se puede por ejemplo calcular de allí la frecuencia relativa (o proporción poblacional) de elementos que reunen ambas características, esta es $\frac{65}{160}$. De otra manera, se puede hallar primero $P(B|L_1)$, que representa a la probabilidad de que la placa sea buena sabiendo que corresponde al Lote 1, y multiplicarla por $P(L_1)$. Entonces

$$P(B \cap L_1) = P(L_1)\,P(B|L_1) = 0,5 \cdot 0,8125 = 0,4062.$$

Por otra parte, $P(B) = \frac{133}{160} = 0,831$ y entonces es claro que

$$P(B \cap L_1) \neq P(B)\,P(L_1),$$

es decir se concluye que B y L_1 no son sucesos independientes. Se puede decir entonces que el conocimiento de la procedencia (Lote 1) de la placa, alteró la probabilidad de que la placa sea buena, respecto de la probabilidad original de que lo sea. Note que se puede concluir más rápidamente lo mismo a partir de que $P(B|L_1) \neq P(B)$, es decir, viendo que la probabilidad de B cambió ante el conocimiento (o condicionamiento) de que pertenece al Lote 1.

Con lo visto hasta ahora se pueden resolver las Actividades 6 a 9, Sección 4.8.

Problema 2 (pág. 7):

1. Si interesa extraer placas de manera sucesiva y sin reposición del total de las 160 placas del lote combinado (ver Tabla 4.5, página 194), claramente se puede aventurar que el resultado (buena, regular o mala) de la placa extraída en segundo lugar, dependerá del resultado de la primer extracción, ya que si esta última fuera buena por ejemplo, la conformación del lote antes de la segunda extracción será de 132 buenas, 24 regulares y 3 malas, y por ende, las chances de obtener una de cada tipo cambian (observar que además se cuenta con una placa menos en total). Entonces se puede calcular, por ejemplo, la probabilidad de que el suceso: la placa extraída en segundo lugar es buena, pero claramente se tendrá que contar con la información del resultado de la primer extracción, es decir, si la placa extraída en primer lugar fue buena, regular o mala.
Se dice que la ocurrencia en la primer extracción de cierto suceso, condiciona la ocurrencia de dicho suceso en la segunda extracción.
Se considera el siguiente experimento aleatorio: extraer sucesivamente y sin reposición dos placas al azar del total de 160 placas y observar la cantidad de defectos que tiene cada una.
Se pueden definir, a modo de ejemplo, los sucesos B_1 y B_2 de la siguiente manera:

B_1: la placa seleccionada en primer lugar es buena y B_2: la placa seleccionada en segundo lugar es buena.
$P(B_1) = \frac{133}{160} = 0,831$ *indica la probabilidad de que la placa seleccionada en primer lugar sea buena (es decir, que tenga hasta 2 defectos).*

$P(B_2|B_1) = \frac{132}{159} = 0,83$ *indica la probabilidad de que la placa seleccionada en segundo lugar sea buena, sabiendo que la primera placa seleccionada fue buena.*

$P(B_1 \cap B_2) = P(B_1)\,P(B_2|B_1) = 0,831 \cdot 0,83 = 0,6897$ *indica la probabilidad de que las dos placas seleccionadas sean buenas.*

2. Note que la situación cambia completamente si las extracciones de placas se realizan con reposición, es decir de manera tal que, luego de extraer cualquiera de ellas, la misma es devuelta al lote: en la siguiente extracción no es necesario preocuparse por el resultado obtenido ya que se está en las mismas condiciones que al inicio. Esto es equivalente a pensar que la selección de las unidades se realiza a partir de una población infinita.
Se considera ahora el siguiente experimento aleatorio: extraer sucesivamente y con reposición dos placas al azar del total de 160 placas y observar el número de defectos que tiene cada una.
Para los sucesos B_1 y B_2, definidos anteriormente, se pueden obtener las siguientes probabilidades:

$P(B_1) = \frac{13}{160} = 0,831$ *indica la probabilidad de que la placa seleccionada en primer lugar sea buena (es decir, que tenga hasta 2 defectos).*

$P(B_2|B_1) = P(B_2) = \frac{133}{160} = 0,831$ *indica la probabilidad de que la placa seleccionada en segundo lugar sea buena, sabiendo que la primera placa seleccionada fue buena. En este caso, la probabilidad de que una placa sea buena es la misma para cualquier extracción, ya que estas se realizan con reposición.*

$P(B_1 \cap B_2) = P(B_1)\,P(B_2|B_1) = P(B_1)\,P(B_2) = 0,831 \cdot 0,831 = 0,6906$ *indica la probabilidad de que las dos placas seleccionadas sean buenas.*

En resumen, en el punto 1 se considera una secuencia de sucesos (en el ejemplo, de dos sucesos) que no resultan independientes entre sí y en el punto 2, se considera una secuencia de sucesos que sí resultan independientes entre sí y, para obtener la probabilidad de cada secuencia, se multiplican las probabilidades de los sucesos correspondientes.



Con lo visto hasta ahora se pueden resolver las Actividades 10 a 12, Sección 4.8.

En el siguiente ejemplo se puede pensar si es razonable la independencia de ciertos sucesos de interés y, si es así, utilizarla. Por ende, para estos sucesos se cumplirán las igualdades en la definición de independencia y esto permitirá el cálculo de otras probabilidades asociadas al problema:

Ejemplo 4.2 Suponga que se tiene un sistema armado con dos componentes, conectados en serie (Figura 4.3.a). En este caso, el sistema funciona si ambas componentes funcionan.



Figura 4.3 *Dos componentes conectadas a) en serie y b) en paralelo*

Se sabe además, que el funcionamiento (o no) de una de las componentes no altera la probabilidad de funcionamiento de la otra componente. Esto lleva a poder considerar que ambas componentes funcionan de "manera independiente".

Si se conoce que la probabilidad de que la primera componente funcione luego de t horas es 0,80 y que la segunda lo haga es 0,90, se puede entonces calcular la probabilidad de que el sistema funcione luego de t horas. Para ello se definen los sucesos:

A: la primera componente funciona luego de t horas

B: la segunda componente funciona luego de t horas.

En ese caso, el sistema funciona luego de t horas si ambas componentes funcionan. Es decir, interesa considerar el suceso $A \cap B$, que ocurre cuando ambas componentes funcionan a la vez.

Se tiene luego

$$P(A \cap B) = P(A)\,P(B) = 0,80 \cdot 0,90 = 0,72$$

ya que, según lo dicho en el enunciado, se considera que las componentes funcionan independientemente una de otra.

¿Cómo se obtendrá la probabilidad de que el sistema funcione, pero considerando que las componentes están conectadas en paralelo? (Figura 4.3.b).

En ese caso, el sistema funciona luego de t horas si cualquiera de las dos componentes lo hace. Es decir, interesa considerar el suceso $A \cup B$, que ocurre cuando alguno o ambos, A ó B, ocurre. Entonces se tiene

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0,80 + 0,90 - 0,72 = 0,98.$$

Se puede observar lo siguiente:

1. Para el cálculo de $P(A \cap B)$ se utilizó el conocimiento de la independencia entre A y B.
2. Los sucesos A y B del Ejemplo 4.2 no son mutuamente excluyentes porque es posible que ambas componentes funcionen simultáneamente.
3. El sistema ahora, con sus componentes conectadas en paralelo, tiene mayor probabilidad de funcionar luego de t horas, es decir mayor *confiabilidad*.

Con lo visto hasta ahora se pueden resolver las Actividades 13 y 14, Sección 4.8.

4.6 Teorema de la Probabilidad Total

Se propone reflexionar sobre la siguiente situación:

Problema 1 (pág. 7): *Como se vio en la página 195, en la empresa metalúrgica se producen tres tipos de barras (A, B y C). Se conoce que el porcentaje de unidades que no cumple con las especificaciones en relación a la longitud (barras no conformes) para A, B y C es de 3 %, 5 % y 6 % respectivamente.*
Suponga que ahora se conoce también que el 50 % de las barras que se producen son de tipo A; mientras que de los tipos B y C se produce un 25 % para cada uno. ¿Permite esta información determinar el porcentaje de barras no conformes en el total? Equivalentemente, ¿se podrá hallar la probabilidad de que una barra producida en la empresa sea no conforme?
La empresa metalúrgica produce tres tipos de barras para la industria automotriz, que se pueden denominar A, B y C.
¿Permite esta información determinar el porcentaje de barras no conformes en el total? Equivalentemente, ¿se podrá hallar la probabilidad de que una barra producida en la empresa sea no conforme?
Si se denota con x al total de barras, entonces $x \cdot 0,5$ *representa el total de barras de tipo A que se producen y, de manera similar,* $x \cdot 0,5 \cdot 0,03$ *representa el total de barras de tipo A que son 'no conformes'.*
De manera análoga se establece que $x \cdot 0,25 \cdot 0,05$ *y* $x \cdot 0,25 \cdot 0,06$ *representan el número de barras de tipos B y C respectivamente que se producen en la empresa. En consecuencia* $x \cdot 0,5 \cdot 0,03 + x \cdot 0,25 \cdot 0,05 + x \cdot 0,25 \cdot 0,06$ *representa el total de barras no conformes producidas.*
La proporción poblacional de barras no conformes, o probabilidad de hallar una barra no conforme en la producción total, $P(D)$*, se obtiene dividiendo el número total de barras no conformes por el total de barras producidas, x, obteniéndose:*
$P(D) = 0,5 \cdot 0,03 + 0,25 \cdot 0,05 + 0,25 \cdot 0,06 = 0,0425.$
Se concluye que aproximadamente el $4,25\,\%$ *de la totalidad de las barras es no conforme.*
Se ha determinado la proporción de barras no conformes de la producción total, a partir de la información de las proporciones de barras no conformes de cada tipo y de la proporción de barras producidas de cada tipo.
La información se puede volcar en un diagrama de árbol (ver cómo construir un diagrama como este en el Apéndice, Sección 4.9.2) como el que se presenta en la Figura 4.4:

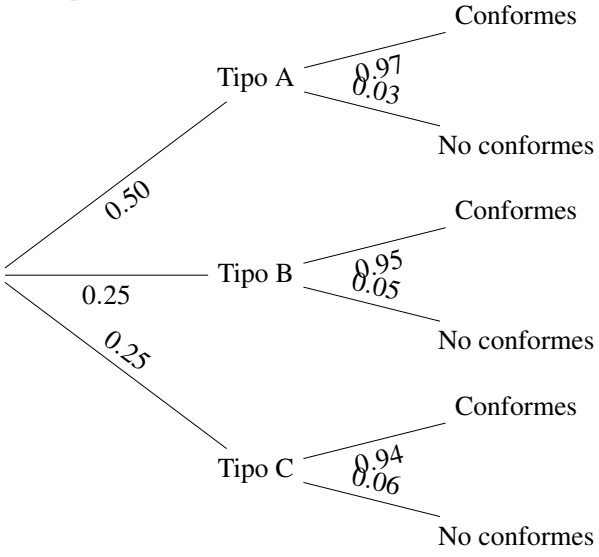


Figura 4.4 *Esquema de árbol para la situación de las barras de la empresa metalúrgica.* $P(D)$ *puede hallarse sumando los productos de las probabilidades sobre las correspondientes ramas del árbol*

Si se denota con:
A: una barra elegida al azar es de tipo A,
B: una barra elegida al azar es de tipo B,
C: una barra elegida al azar es de tipo C, se tiene

$$P(D) = P(A)\,P(D|A) + P(B)\,P(D|B) + P(C)\,P(D|C).$$

El resultado obtenido constituye un caso particular del Teorema de la Probabilidad Total *que se considera en el Apéndice (Sección 4.9.3).*

Retomando la situación planteada sobre las barras, suponga que se elige una barra al azar de la producción total y esta resulta ser no conforme. Qué probabilidad tiene de ser de alguno de los tipos A, B o C? Es decir, se desea hallar $P(A|D)$, $P(B|D)$ y $P(C|D)$, o solo alguna de ellas.

Para ello se puede proceder de la siguiente manera:

A partir de la definición de probabilidad condicional, se sabe que: $P(A|D) = \frac{P(A\cap D)}{P(D)}$, y del resultado recién visto (Teorema de la Probabilidad Total), se sabe cómo hallar $P(D)$.

Luego se obtiene:

$$P(A|D) = \frac{P(A)\,P(D|A)}{P(A)\,P(D|A) + P(B)\,P(D|B) + P(C)\,P(D|C)} = \frac{0,5\cdot 0,03}{0,0425} = 0,353$$

.

Este resultado se puede interpretar de la siguiente manera:

- Si se selecciona al azar una barra y esta es no conforme, la probabilidad o chance de que sea del tipo A es 0,353.
- Pensando en una gran cantidad de barras no conformes, aproximadamente el 35,3 % de ellas es de tipo A. Es decir, en la subpoblación formada por todas las barras no conformes, las de tipo A representan aproximadamente el 35,3 %.

Este último resultado se conoce con el nombre de **Teorema de Bayes**.



Con lo visto hasta ahora se pueden resolver las Actividades 15 a 19, Sección 4.8.

A modo de resumen, se enuncian a continuación definiciones y propiedades desarrolladas hasta este punto:

- Si A y B son sucesos cualesquiera, resulta $P(A\cup B) = P(A) + P(B) - P(A\cap B)$.
  - En particular, si A y B son sucesos mutuamente excluyentes (ver definición en página 187), resulta $P(A\cup B) = P(A) + P(B)$.
- $P(A) = 1 - P(\bar{A})$.
- $P(A|B) = \frac{P(A\cap B)}{P(B)}$, para $P(B) \neq 0$.
- $P(A\cap B) = P(A|B)\,P(B)$, para $P(B) \neq 0$.
  - En particular, si A y B son independientes (ver definición en página 196), resulta $P(A\cap B) = P(A)\,P(B)$.

- Teorema de la Probabilidad Total (ver página 211): Si B es un suceso cualquiera de S, $\{A_1, A_2, \cdots A_r\}$ una partición de S y se conocen $P(B|A_i)$ y $P(A_i)$ para todo $i = 1, 2, \cdots r$, entonces

$$P(B) = P(A_1)\,P(B|A_1) + P(A_2)\,P(B|A_2) + \cdots + P(A_r)\,P(B|A_r).$$

4.7 Síntesis

La Teoría de la Probabilidad proporciona las bases matemáticas y el lenguaje para la descripción de la variación implícita en los experimentos aleatorios.

No se puede decir a priori si un suceso (de un espacio muestral asociado a un experimento aleatorio) ocurrirá o no al realizar el experimento; pero sí se le podrá asociar un número que mida de alguna manera, la posibilidad que tiene dicho suceso de ocurrir. Esto conduce al concepto de probabilidad.

Hay diferentes maneras de obtener ese número, como se ha visto en este capítulo: a través de la definición clásica (restringida a espacios muestrales finitos en los cuales todos los sucesos elementales son equiprobables), conociendo la distribución de probabilidades de la variable aleatoria de interés o bien, estudiando las frecuencias relativas al repetir el experimento aleatorio un cierto número de veces, en condiciones similares (en esta última se obtiene una estimación o valor aproximado de la probabilidad).

Al formalizar las propiedades que cumple la frecuencia relativa surge la definición axiomática de probabilidad, la cual generaliza la definición clásica. Esta formalización, sin embargo, no especifica de qué manera asignar probabilidades a los sucesos.

Independientemente de cómo se obtenga, la probabilidad de un suceso admite dos interpretaciones: como una medida de la chance o posibilidad de que dicho suceso se verifique (u ocurra) en una realización particular del experimento, o, como la proporción o frecuencia relativa de elementos de la población que tienen valores (o niveles) de la característica en estudio según se enuncian en el suceso.

La probabilidad permite cuantificar la incertidumbre en una amplia variedad de situaciones, tanto en el ámbito científico como en la vida cotidiana. En particular, en la resolución de problemas ingenieriles de naturaleza estadística, los profesionales suelen enfrentarse a la toma de decisiones en contextos donde la incertidumbre, el azar y el riesgo son factores relevantes. En estos casos, la probabilidad actúa como una guía para la toma de decisiones fundamentadas. Cuando se aplican herramientas inferenciales (es decir, cuando se busca generalizar las conclusiones obtenidas a partir de una muestra hacia toda la población), la probabilidad permite cuantificar el riesgo asociado a posibles errores en ese proceso.

4.8 Actividades propuestas

1. Considere los problemas de la Sección 1.4, Capítulo 1 (página 7), y complete una tabla como la Tabla 4.1, para los problemas no considerados en la misma.

2. Para cada uno de los experimentos definidos en la Actividad 1:
   a) Defina una variable de interés e indique el recorrido de la misma.
   b) Defina dos sucesos cualesquiera y exprese a uno de ellos como proposición y al otro, en función de la variable aleatoria definida anteriormente.

3. En un laboratorio de calidad de una empresa metalúrgica se inspeccionan piezas metálicas especiales. En dicha inspección se estudian ciertas condiciones de las piezas que pueden requerir retrabajo antes de su aprobación final. Se definen los siguientes sucesos:
   A: la pieza requiere retrabajo por desviación dimensional, y
   B: la pieza requiere retrabajo por acabado superficial fuera de especificación.
   A partir de datos históricos del proceso se dispone de la siguiente información:
   $P(\bar{A}) = 0,88$; $P(B) = 0,08$ y $P(A \cap B) = 0,03$.
   a) Defina un experimento aleatorio adecuado para esta situación, indicando la población de referencia.
   b) Interprete las probabilidades mencionadas en el enunciado, en contexto.
   c) Debido a los costos asociados a los retrabajos, la empresa considera aceptable el proceso de producción si menos del 18 % de las piezas requiere al menos un tipo de retrabajo. Determine si el proceso resulta aceptable según este criterio. Justifique su respuesta y explicite los cálculos realizados.
   d) Dé un ejemplo de sucesos mutuamente excluyentes asociados a la situación.

4. En una línea de ensamblaje de equipos electromecánicos se evalúa la calidad de los componentes electrónicos que se incorporan en la etapa final. En cada componente interesa analizar si presenta defectos superficiales y/o soldaduras defectuosas.
   Por experiencia se conoce que:
   - la proporción de componentes con defectos superficiales es 0,10;
   - la probabilidad de que un componente seleccionado al azar presente al menos uno de los problemas es 0,22;
   - la chance de que un componente presente ambos problemas es 0,04.

   a) Defina el experimento aleatorio.
   b) Exprese formalmente la información brindada en el enunciado.
   c) Determine la probabilidad de que un componente seleccionado al azar presente soldaduras defectuosas. Interprete dicho valor en el contexto de la situación.
   d) Los sucesos A: el componente presenta defectos superficiales y B: el componente presenta soldaduras defectuosas, ¿son mutuamente excluyentes? Justifique su respuesta.

5. Un laboratorio de resistencia de materiales, después de realizar ensayos de elongación y torsión sobre varillas de un nuevo material, informa que la probabilidad de que una varilla resulte con fallas de elongación es 0,06; con fallas de torsión es 0,04 y con al menos uno de los dos tipos de fallas es 0,085.
   En una empresa que fabrica cierto tipo de equipos, están considerando utilizar estas varillas en el proceso de producción. En esta empresa saben, por experiencia, que si la varilla presenta ambos

tipos de falla, el proceso de armado del equipo se detiene, pudiendo producirse importantes demoras en todo el proceso en general. Por ese motivo, solo comprarían este tipo de varillas si el riesgo de que el proceso se detenga es inferior a 0,02.
A partir de la información brindada por el laboratorio, ¿qué le recomendaría Ud. a la empresa en relación a la compra de este tipo de varillas? Justifique su respuesta. Incluya los cálculos que haya realizado, expresados formalmente.

6. Reconsidere la Actividad 3, referida a la evaluación de piezas metálicas especiales.
   a) Obtenga $P(A|B)$ e interprete su valor en el contexto de la situación planteada.
   b) Analice si los sucesos A y B pueden considerarse independientes. Justifique formalmente su respuesta y explique, en contexto, el significado de dicha independencia o dependencia.

7. En una empresa constructora que realiza obras de gran envergadura, se conoce, por experiencia, que el 70 % de las uniones de caños de PVC se realiza con cierta una resina tipo B. De estas uniones preocupa que la resistencia a la tensión sea inferior a la deseada.
   Se definen los sucesos: I: la unión tiene resistencia a la tensión inferior a la deseada y B: la unión se realizó con resina tipo B.
   a) ¿Qué significa en este contexto $P(I \cup B) = 0,72$? , ¿y $P(I \cap B) = 0,034$?
   b) Uno de los ingenieros a cargo propone seguir usando resina B si la proporción de uniones realizadas con dicha resina que tienen resistencia inferior a la deseada es como máximo 0,04. ¿Qué le recomendaría Ud.? Justifique, expresando formalmente las probabilidades correspondientes.
   c) ¿Son I y B sucesos independientes? Justifique su respuesta.

8. Considere nuevamente la Actividad 4. A partir de la información allí proporcionada, indique si las siguientes afirmaciones son verdaderas o falsas. En cada caso, justifique su respuesta mediante los cálculos correspondientes.
   a) La presencia de soldaduras defectuosas en un componente es independiente de la presencia de defectos superficiales.
   b) Entre los componentes que presentan soldaduras defectuosas, la proporción de aquellos que presentan defectos superficiales es 0,25.
   c) La proporción de componentes que no presentan ninguno de los problemas mencionados es 0,96.

9. Durante el año anterior, en un taller dedicado a la reparación de equipos especiales se realizaron 370 reparaciones, las cuales se clasificaron según la marca del equipo y el tiempo requerido. La siguiente tabla muestra la información obtenida. Entre paréntesis se indican las letras sugeridas para cada suceso de interés.

	Hasta 15 horas (L)	**Más de 15 horas (M)**
Marca A (A)	10	20
Marca B (B)	120	100
Marca C (C)	67	53

   Se elige al azar una reparación realizada durante el año anterior.
   a) Calcule la probabilidad de que ocurra cada uno de los siguientes sucesos:
      1) La reparación seleccionada demandó hasta 15 h.

2) La reparación seleccionada correspondió a un equipo de marca C.
3) La reparación seleccionada correspondió a un equipo de marca A o C.
4) La reparación seleccionada no correspondió a un equipo de marca A.
5) La reparación seleccionada correspondió a un equipo de marca C o demandó hasta 15 h.
6) La reparación seleccionada demandó hasta 15 h y el equipo era de marca C.
7) La reparación seleccionada demandó hasta 15 h cuando el equipo era de marca C.

b) ¿Son los sucesos L y C independientes? Justifique su respuesta.

c) Si en el año anterior, más del 40 % de las reparaciones de equipos de marca A duraron más de 15 h, se considerará contratar a una técnica más en el presente año.
1) Plantee el problema mencionando población, variable y parámetro de interés. Escriba las opciones que se plantean en el taller en función de los valores de dicho parámetro.
2) ¿Qué cree usted que harán en el taller? Justifique su respuesta formalizando el cálculo realizado.
3) ¿Qué otra información considera que tendrán en cuenta en el taller para tomar finalmente la decisión?

10. Se cuenta con un lote de 10 artículos del que se conoce que 3 están en mal estado. Se seleccionarán 2 artículos al azar del lote, en forma sucesiva y sin reposición (esto quiere decir que el artículo que se extrae no se vuelve a reponer al lote).
a) Calcule la probabilidad de que uno de los artículos extraídos esté en buen estado y el otro no.
b) Obtenga la misma probabilidad; pero suponiendo ahora que los artículos se seleccionaron con reposición.

11. Un fabricante de cierto tipo de piezas las envía a sus clientes en lotes de 150 unidades. Un lote en el que hay menos de 4 piezas defectuosas se considera de buena calidad; mientras que un lote con 4 o más piezas defectuosas se considera de mala calidad.
Un cliente utiliza las piezas para fabricar otros productos y compra habitualmente lotes a este fabricante. Cuando recibe un lote selecciona al azar 3 unidades, en forma sucesiva y sin reposición, y las examina. Si las tres unidades seleccionadas son "no defectuosas", el cliente acepta el lote y envía la totalidad de las piezas a su proceso de producción. En caso contrario, devuelve al proveedor el lote completo.
a) Suponga que el proveedor acaba de enviar a su cliente un lote que tiene 2 piezas defectuosas (lote de buena calidad). ¿Qué riesgo corre de que el cliente se lo devuelva?
b) Suponga que el cliente acaba de recibir otro lote del proveedor, que tiene 8 piezas defectuosas (lote de mala calidad). ¿Cuál es el riesgo de que el cliente envíe todas las piezas de ese lote a producción?

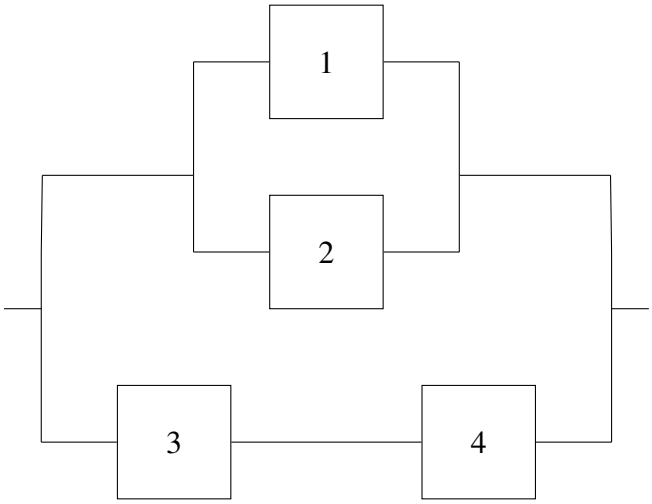
12. En una planta industrial tienen conectada una alarma que se activa ante un aumento sustancial de la temperatura, habilitando rociadores de líquido. Se conoce que esta alarma tiene un riesgo de fallar de 0,10.
a) Defina la población en estudio.
b) ¿Cuál es la probabilidad de que la alarma se active ante un aumento sustancial de la temperatura? Interprete dicho valor como una proporción.
c) Alguien sugiere conectar una segunda alarma, en un circuito independiente del circuito

al que está conectada la alarma actual, para aumentar la chance de que se activen los rociadores de líquido en caso necesario. Esta nueva alarma tiene probabilidad 0,94 de detectar el aumento de temperatura.

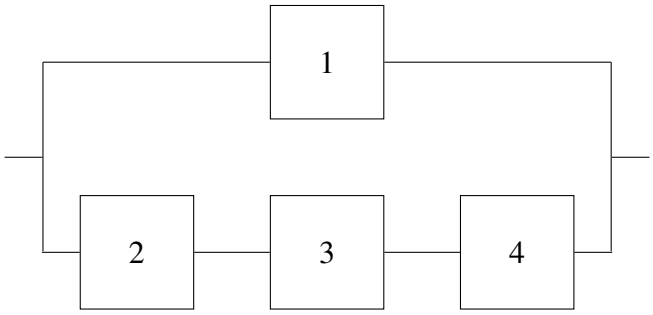
¿Cuál es la probabilidad de que se activen los rociadores de líquido por la acción de cualquiera de las dos alarmas?

d) Escriba un párrafo informando sobre la ventaja de conectar una segunda alarma con las características mencionadas.

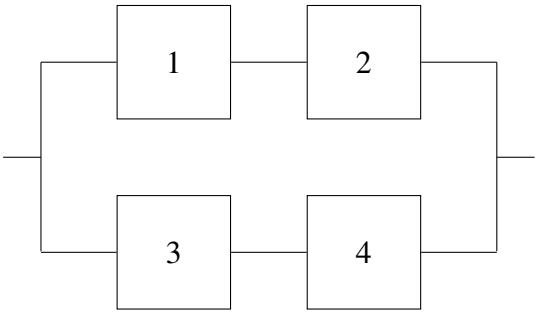
13. Se necesita armar sistemas conectando 4 componentes. Las opciones para conectarlas se muestran en las figuras. En cualquiera de los casos, las componentes se van a conectar de manera tal que sus funcionamientos sean independientes entre sí. Se conoce además que las probabilidades de que estas componentes fallen antes de las 100 horas son 0,10; 0,05; 0,10 y 0,20 para las componentes 1, 2, 3 y 4 respectivamente.



(a)



(b)



(c)

Un usuario de esos sistemas requiere que los mismos funcionen como mínimo 100 h. ¿Cuál configuración le resultará más conveniente? Justifique su decisión.

14. Se cuenta con dos componentes de las que se conoce que su probabilidad de fallar antes de cierto tiempo t vale p.

a) Se tiene la posibilidad de armar un sistema conectando ambas componentes en serie. ¿Cuál es la probabilidad de que dicho sistema funcione pasado el tiempo t? Suponga que las componentes funcionan de manera independiente.

b) Obtenga la probabilidad de que el sistema funcione pasado el tiempo t, si las componentes fueron conectadas en paralelo.

c) ¿En cuál de los casos la probabilidad de que el sistema funcione pasado el tiempo t es mayor? Justifique.

15. Reconsidere la información de la Tabla 4.5, a partir de la cual se construyó el diagrama de árbol que se presenta en la Sección 4.9.2. Si se arma un nuevo lote, con todas las placas de los Lotes 1 y 2, ¿qué proporción de placas buenas, regulares y malas se presentan en dicho lote combinado? Utilice la información del diagrama de árbol para dar su respuesta.

16. Reconsidere la situación propuesta en la Sección 4.6, relativa a las barras tipos A, B y C fabricadas en la empresa metalúrgica.

a) ¿Qué proporción de las barras del proceso son del tipo C y además son "no conformes"?

b) Suponga que se selecciona una barra al azar de la producción y se encuentra que no cumple con las especificaciones en relación a la longitud (es decir, es no conforme).

1) Obtenga la probabilidad de que dicha barra sea del tipo B.
2) Obtenga la probabilidad de que dicha barra sea del tipo C.

17. Por experiencia se conoce que el 1 % de las unidades producidas por una fábrica se encuentran defectuosas cuando el proceso de fabricación se encuentra bajo control; mientras que, si el proceso se encuentra fuera de control,se produce un 30 % de unidades defectuosas. Se cree que la probabilidad de que el proceso se encuentre bajo control es 0,92.

a) Vuelque toda la información del enunciado en un diagrama de árbol.

b) Si la proporción de unidades defectuosas del proceso es menor a 0,05, en la empresa no considerarían realizan acciones de mejora debido a los costos que estas representan. ¿Cree Ud. que deberán realizar mejoras en este caso?

c) ¿Cuál sería su respuesta si, en realidad, el 75 % de las veces el proceso se encuentra bajo control?

d) Volviendo a la situación original, suponga que se selecciona una unidad al azar del proceso y esta resulta defectuosa. ¿Cuál es la probabilidad de que haya sido producida con el proceso bajo control?

e) Analice si los sucesos "la unidad fue producida cuando el proceso se encontraba bajo control" y "la unidad es defectuosa" son independientes.

18. Para ejecutar un proceso se selecciona uno de tres periféricos: A, B o C. La probabilidad de elegir cada uno de ellos es $0,60$, $0,20$ y $0,20$ respectivamente. Como resultado de la elección se pueden producir perturbaciones que detienen la ejecución del proceso. Esto ocurre el 2 % de las veces si el periférico seleccionado fue el A, el 1 % si fue B y el $2,5$ % si fue C.

a) Presente la información en un diagrama de árbol. Exprese formalmente las probabilidades que mencione en dicho diagrama.

b) Calcule la probabilidad de que el proceso no se ejecute.

c) Si el proceso se ha ejecutado, ¿cuál es la probabilidad de que lo haya hecho desde el periférico C?

19. Para cada una de las siguientes proposiciones, indique si es verdadera o falsa. Justifique su respuesta.

a) $P(A \cup B) = P(A) + P(B)$ para todo par de sucesos A y B de un mismo espacio muestral.

b) $P(A \cap B) = P(A)\,P(B)$ para todo par de sucesos A y B de un mismo espacio muestral.

c) $P(\bar{A} \cup A) = 1 - P(\bar{A} \cap A)$.

d) En un proceso de producción se toma una muestra de 100 piezas y se observa que 3 son defectuosas. Entonces, la probabilidad de encontrar una pieza defectuosa en el proceso es exactamente 0,03.

e) Sean $S = \{s_1,\ s_2,\ s_3,\ s_4\}$ y un suceso $A = \{s_1, s_2\}$, entonces $P(A) = \frac{1}{2}$.

f) Sean dos sucesos A y B, ambos con probabilidades no nulas, asociados a una experiencia aleatoria. Si dichos sucesos son mutuamente excluyentes, entonces, resultan también independientes.

g) $P(A|B) + P(B|A) = 1$.

h) $P(A|B) = P(B|A)$.

4.9 Apéndice

4.9.1 Probabilidad. Definición axiomática. Consecuencias

Se define a la probabilidad sobre un espacio muestral S a través de las siguientes propiedades (comúnmente denominadas axiomas), inspiradas por las propiedades de las frecuencias relativas:

Axiomas de la probabilidad de un suceso

1. $P(A) \geq 0$ para cualquier suceso A,
2. $P(S) = 1$,
3. si A y B son sucesos mutuamente excluyentes y se considera la unión de ellos, llamando C al nuevo suceso (es decir $C = A \cup B$), resulta

$$P(C) = P(A \cup B) = P(A) + P(B)$$

A continuación se enuncian algunas propiedades que surgen como consecuencia de la definición de probabilidad introducida a través de los axiomas. Independientemente de cómo se haya obtenido la probabilidad de un suceso, las propiedades que se desprenden de la definición axiomática permiten obtener probabilidades de otros sucesos.

1. La probabilidad del suceso complemento:

$$P(\bar{A}) = 1 - P(A)$$

Se puede razonar de la siguiente manera: siendo $S = A \cup \bar{A}$, se puede utilizar el hecho que A y $\bar{A}$ son mutuamente excluyentes para afirmar que $P(S) = P(A) + P(\bar{A})$ y, a continuación a partir de $P(S) = 1$, concluir que $1 = P(A) + P(\bar{A})$.

2. La probabilidad de la unión de dos sucesos cualesquiera (es decir, no necesariamente excluyentes):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Para demostrar esta igualdad, se puede razonar de acuerdo a los siguientes pasos:

a) siendo $A \cup B = A \cup (B \cap \bar{A})$, se puede utilizar el hecho que A y $B \cap \bar{A}$ son mutuamente excluyentes para afirmar que $P(A \cup B) = P(A) + P(B \cap \bar{A})$;

b) siendo $B = (A \cap B) \cup (B \cap \bar{A})$, se puede afirmar de manera similar que $P(B) = P(A \cap B) + P(B \cap \bar{A})$;

c) a partir de a. y b. surge que $P(A \cup B) = P(A) + P(B \cap \bar{A}) = P(A) + P(B) - P(A \cap B)$, y la propiedad queda demostrada.

3. La probabilidad de la unión de dos sucesos cualesquiera mostrada en el ítem 2 puede generalizarse a tres o más sucesos cualesquiera de un mismo espacio muestral. Por ejemplo, para tres sucesos A, B y C,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Para demostrar esta igualdad, se puede aplicar la propiedad asociativa de la unión y utilizar a continuación la propiedad 2.

4.9.2 Construcción de un diagrama de árbol

Un diagrama de árbol es una representación gráfica de los posibles resultados de un experimento (aleatorio en nuestro caso), los cuales pueden clasificarse según diferentes criterios. En general, un diagrama de este tipo se utiliza en los problemas de conteo y probabilidad. Su construcción es sencilla; aunque difícil de explicar en palabras.

A continuación, se exponen algunas indicaciones para su elaboración y, finalmente, se presenta un ejemplo. Para su construcción, es importante tener en cuenta lo siguiente:

1. Los sucesos que forman el primer conjunto de ramas corresponden al primer criterio.
   Estos sucesos deben ser mutuamente excluyentes entre sí y sus probabilidades deben sumar uno; es decir deben conformar una partición del espacio muestral (ver definición en la Sección 4.6 de este Apéndice).
2. A partir de cada rama del primer conjunto se construye un segundo conjunto de ramas asociado al segundo criterio.
   Los sucesos de estas nuevas ramas también conforman una partición, pero del espacio muestral condicionado a la rama del primer conjunto de la cual se desprenden. Es importante notar que las probabilidades asignadas a cada rama del segundo conjunto son probabilidades condicionadas al suceso de la rama anterior.
3. Se continúa con este proceso en el caso en que hubiera nuevos conjuntos de ramas asociados a otros criterios.



Problema 2 (pág. 7): *Reconsiderando este ejemplo, la información de la Tabla 4.5, se puede volcar en un diagrama de árbol siguiendo las reglas mencionadas.*

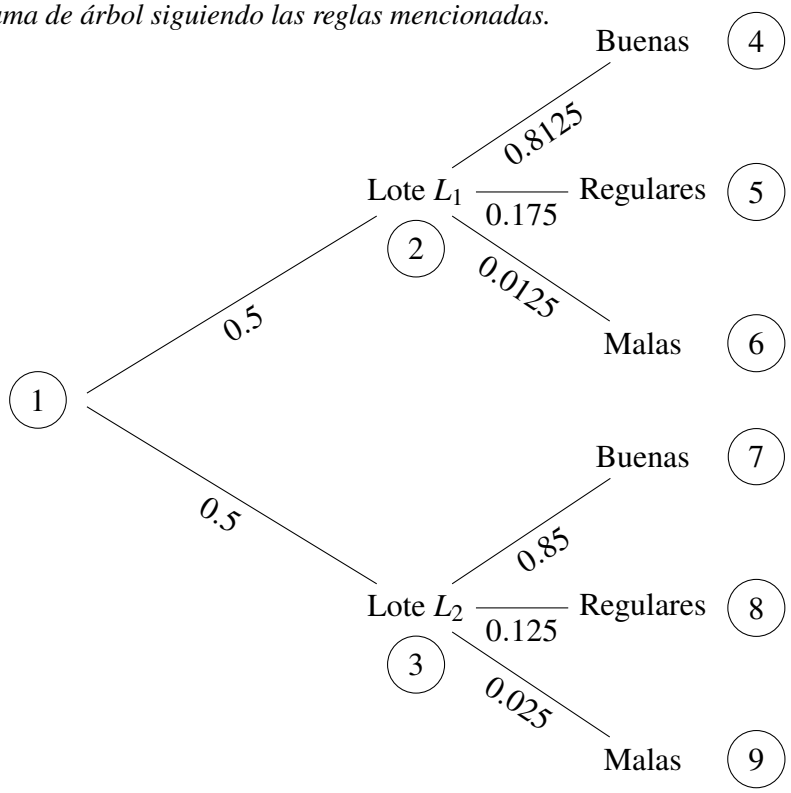


Figura 4.5 *Diagrama de árbol para la calidad de las placas por lote*

El primer conjunto de ramas se refiere al lote de donde se extrae la placa: en cada rama se representa uno de los lotes (suceso L_i) y su probabilidad ($P(L_i)$), para $i = 1,2$.

La probabilidad de que la placa seleccionada provenga del Lote 1 vale 0,50 ($P(L_1) = \frac{80}{160} = 0,50$).

La probabilidad de que la placa seleccionada provenga del Lote 2 vale 0,50 ($P(L_2) = \frac{80}{160} = 0,50$).

Observe que los sucesos L_1 y L_2 son mutuamente excluyentes y sus probabilidades suman 1.

El segundo conjunto de ramas se refiere a la calidad de las placas y en el diagrama se representa uno de estos segundos conjuntos para cada lote, es decir, para cada rama del primer conjunto.

En cada rama del segundo conjunto se representa una de las calidades (suceso B, R o M) y su probabilidad condicionada, según el lote correspondiente (es decir, según en qué rama del primer conjunto se encuentre).

Para el segundo conjunto de ramas asociado al Lote 1 se tienen las siguientes probabilidades:

- *La probabilidad de que la placa seleccionada sea de calidad buena si proviene del Lote 1 vale 0,8125 ($P(B|L_1) = \frac{65}{80} = 0,8125$).*
- *La probabilidad de que la placa seleccionada sea de calidad regular si proviene del Lote 1 vale 0,175 ($P(R|L_1) = \frac{14}{80} = 0,175$).*
- *La probabilidad de que la placa seleccionada sea de calidad mala si proviene del Lote 1 vale 0,0125 ($P(M|L_1) = \frac{1}{80} = 0,0125$).*

Observe que los sucesos B, R y M son mutuamente excluyentes y que sus tres probabilidades (condicionadas al Lote 1) suman 1.

Para el segundo conjunto de ramas asociado al Lote 2 se tienen las siguientes probabilidades:

- *La probabilidad de que la placa seleccionada sea de calidad buena si proviene del Lote 2 vale 0,85 ($P(B|L_2) = \frac{68}{80} = 0,85$).*
- *La probabilidad de que la placa seleccionada sea de calidad regular si proviene del Lote 2 vale 0,125 ($P(R|L_2) = \frac{10}{80} = 0,125$).*
- *La probabilidad de que la placa seleccionada sea de calidad mala si proviene del Lote 2 vale 0,025 ($P(M|L_2) = \frac{2}{80} = 0,025$).*

Observe que los sucesos B, R y M son mutuamente excluyentes y que sus tres probabilidades (condicionadas al Lote 2) suman 1.

Toda esta información es la que se presenta en el diagrama de árbol de la Figura 4.5.

Si, por ejemplo, se pretende calcular la probabilidad del suceso $L_1 \cap R$, es decir de que una placa extraída al azar provenga del Lote 1 y sea regular, esta se puede obtener recorriendo el árbol y multiplicando las probabilidades 0,5 (reflejada en la rama con extremos en los nodos 1 y 2) con 0,175 (reflejada en la rama con extremos en los nodos 2 y 5).
Formalmente, $P(L_1 \cap R) = P(L_1)\,P(R|L_1) = 0,50 \cdot 0,175 = 0,0875$.

4.9.3 Teorema de la Probabilidad Total

El Teorema de la Probabilidad Total (ver esquema en Figura 4.6) permite calcular la probabilidad de un suceso B de S cuando se conocen $P(B|A_i)$ y $P(A_i)$ para todo $i = 1,2,\cdots r$, siendo A_1, $A_2,\cdots A_r$ una partición de S.

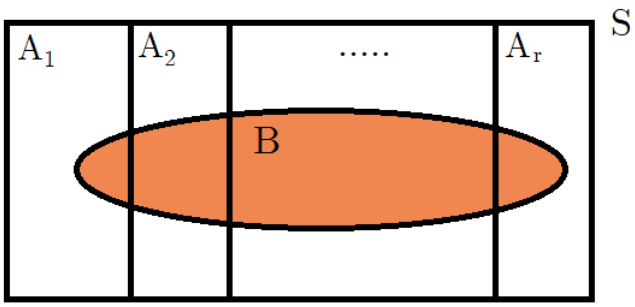


Figura 4.6 *Esquema de partición del espacio muestral y suceso B*

Los sucesos A_1, $A_2, \cdots, A_r$ de un espacio muestral S constituyen una **partición** de S cuando:

1. $A_1 \cup A_2 \cdots A_r = S$
2. $A_i \cap A_j = \emptyset$ para todo $i \neq j$
3. $P(A_i) > 0$ para todo $i = 1, 2, \cdots r$

En efecto, en la Figura 4.6 se puede observar que el suceso B se puede expresar de la siguiente manera:

$$B = B \cap S = B \cap (A_1 \cup A_2 \cup \cdots \cup A_n) = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_r)$$

Entonces:

$$P(B) = P[(B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_r)]$$

Siendo $B \cap A_1, B \cap A_2, \cdots B \cap A_r$ sucesos mutuamente excluyentes , resulta:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_r)$$

y por lo tanto:

$$P(B) = P(A_1)\,P(B|A_1) + P(A_2)\,P(B|A_2) + \cdots + P(A_r)\,P(B|A_r).$$

El resultado obtenido se conoce como **Teorema de la Probabilidad Total**.



5. Distribuciones de probabilidades de uso frecuente

5.1 Introducción

En el Capítulo 3 se presenta el concepto de distribución de probabilidades como el modelo que describe la variabilidad de una variable aleatoria en la población. Además, se definen de manera general las funciones de probabilidad puntual y de densidad de probabilidad, así como las distribuciones de probabilidades acumuladas, y se explica cómo obtener algunos parámetros como el promedio y la desviación estándar, entre otros.

En este capítulo se presentan algunas distribuciones de probabilidades particulares, de uso frecuente en Ingeniería. En la Sección 5.2 se estudian las distribuciones Normal, Uniforme, Triangular y Exponencial, correspondientes a variables aleatorias continuas; mientras que en la Sección 5.3 se analizan las distribuciones Bernoulli, Binomial, Hipergeométrica, Geométrica y Poisson, asociadas a variables aleatorias discretas.

Para cada una de las distribuciones consideradas, los objetivos de este capítulo son:

- Presentar la expresión de la función de densidad de probabilidad o de la función de probabilidad puntual, según corresponda, y, en el caso de variables continuas, la expresión de la función de distribución acumulada.
- Describir sus principales características y propiedades.
- Obtener e interpretar el promedio, la desviación estándar y otros parámetros de interés.

5.2 Variables aleatorias continuas

5.2.1 Distribución Normal

La distribución Normal (o Gaussiana), permite describir una amplia variedad de fenómenos en distintos campos. En Ingeniería, por ejemplo, se utiliza para modelar el comportamiento del contenido de líquido en envases estándar, la dureza o resistencia de materiales, las mediciones de magnitudes de interés en productos manufacturados, los errores aleatorios en estas mediciones, la humedad de granos o el diámetro interior de anillos de pistón, entre otros.

Además, esta distribución es fundamental en la inferencia estadística clásica, ya que la distribución de muchos de los estadísticos empleados para construir intervalos de confianza o realizar pruebas de hipótesis se aproxima a ella (Capítulos 7 a 9).

Una variable aleatoria continua Y tiene una distribución Normal de parámetros matemáticos μ y σ, y se simboliza $Y \sim N(\mu, \sigma)$, si su función de densidad de probabilidad es:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \; con \; y \in \mathbb{R}$$

donde $\mu \in \mathbb{R}$ y $\sigma \in \mathbb{R}^+$.

Los parámetros matemáticos de una distribución son aquellos valores que, si se conocen, hacen que la función de densidad de probabilidad (si la variable aleatoria es continua) o de probabilidad puntual (si la variable aleatoria es discreta) quede completamente definida.

La función de densidad de probabilidad Normal se nombra como curva Normal y se grafica en la Figura 5.1.

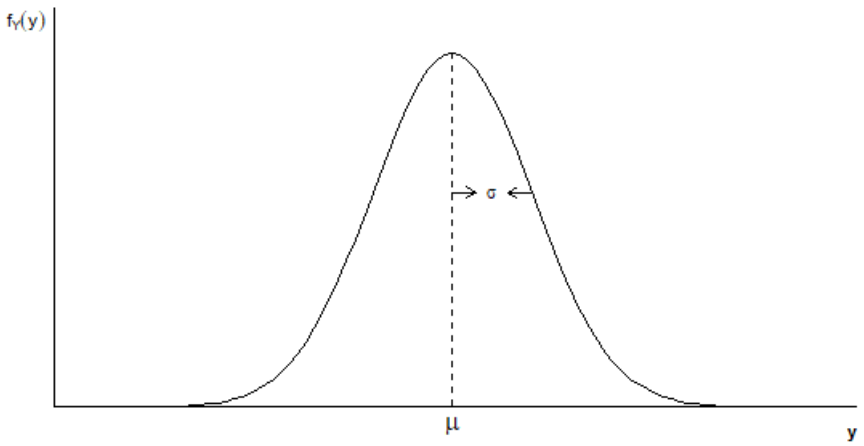


Figura 5.1 *Función de densidad de probabilidad Normal, $f_Y(y)$, de parámetros μ y σ*

Observando la Figura 5.1 se pueden destacar algunas particularidades de la curva Normal:

- tiene forma de "campana";
- es simétrica respecto a $y = \mu$;
- presenta un máximo relativo en el punto $\left(\mu, \frac{1}{\sqrt{2\pi}\sigma}\right)$;
- en $y = \mu$ presenta el valor máximo. Es decir, la moda es μ;
- la media, la mediana y la moda coinciden.

Si solo varía el valor de μ, la campana se traslada horizontalmente, sin variar su forma, como se observa en la Figura 5.2.a. Si solo cambia el valor de σ, la campana se dilata o se contrae, sin variar su posición, como se observa en la Figura 5.2.b.

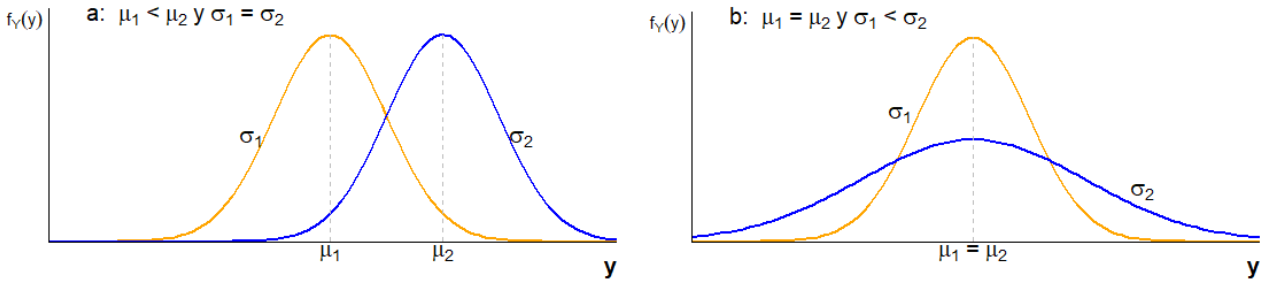


Figura 5.2 *Cambios en la curva Normal al variar μ o σ*

La función de distribución acumulada Normal está definida como:

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(s-\mu)^2}{2\sigma^2}} ds, \; con \; y \in \mathbb{R}.$$

En la Figura 5.3 se muestra la gráfica de la función de distribución acumulada Normal.

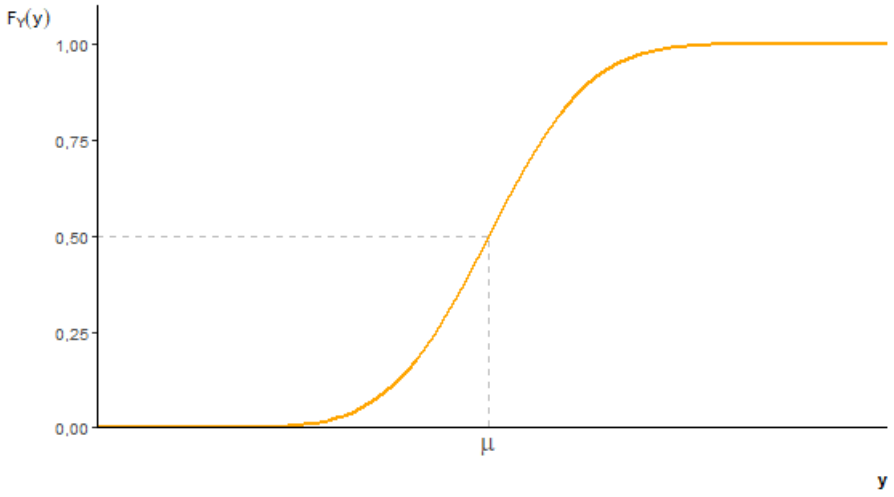


Figura 5.3 *Función de distribución acumulada Normal, $F_Y(y)$*

Si $Y \sim N(\mu, \sigma)$, se demuestra que su media es $E(Y) = \mu$ y su desviación estándar es $D(Y) = \sigma$.

En esta distribución los parámetros matemáticos, μ y σ, son justamente la media y el desvío estándar de la variable Y, respectivamente.

Si a la variable aleatoria Y se le aplica una transformación lineal tal que surge una nueva variable aleatoria $X = a + bY$, con a y b pertenecientes a los números reales, se puede afirmar que:

- Si Y se distribuye según el modelo Normal, X también se distribuye según ese modelo.
- $E(X) = a + bE(Y) = a + b\mu$ (Capítulo 3, Sección 3.3.3).
- $D(X) = |b| D(Y) = |b| \sigma$ (Capítulo 3, Sección 3.3.3).

Utilizando las expresiones anteriores se puede afirmar que si $Y \sim N(\mu, \sigma)$, entonces $X \sim N(a + b\mu, |b| \sigma)$.

En particular, si se define una variable aleatoria $Z = \dfrac{Y - \mu}{\sigma} = \dfrac{1}{\sigma} Y - \dfrac{\mu}{\sigma}$, la cual se denomina **variable estandarizada**, se puede afirmar que si Y se distribuye según el modelo normal, Z también se distribuye según ese modelo, con $E(Z) = 0$ y $D(Z) = 1$ (Capítulo 3, Ejemplo 3.1).
Es decir, si $Y \sim N(\mu, \sigma)$, entonces $Z \sim N(0, 1)$.

La distribución $N(0, 1)$ se denomina **distribución Normal estándar**.

Como se explica en el Capítulo 3, para obtener probabilidades se requiere integrar la función de densidad de probabilidad, lo cual resulta una tarea compleja para el caso de una variable aleatoria con distribución Normal. Si se dispone de un software estadístico (como puede ser R o inclusive Excel), las probabilidades se pueden obtener ingresando los valores de los parámetros matemáticos de la distribución y el/los valor/es de la variable Y. También se dispone de aplicaciones de uso en diferentes dispositivos móviles, como `Probability Distributions` [12], entre otras. En la Sección 5.6 se muestra cómo obtener probabilidades con `R` y en el Apéndice (Sección 5.7) se ejemplifica el uso de la aplicación mencionada.



El uso de software estadístico o de aplicaciones para dispositivos móviles para el cálculo de probabilidades se puede aplicar no solo para el caso de la distribución Normal sino para cualquiera de las distribuciones asociadas a variables continuas o discretas que se presentan en este texto.



Problema 1 (pág. 7): *Una empresa produce barras de acero para la industria automotriz. Una característica de calidad importante para este tipo de barras es su longitud, en mm (Y).*
Suponga que la distribución de esta variable es Normal con parámetros matemáticos $\mu = 250$ ***y*** $\sigma = 0,2$**.**
Interesa responder las siguientes preguntas:

1. *¿Cuál es la longitud media de las barras? ¿y el desvío estándar?*
2. *¿Qué proporción de las barras fabricadas tienen una longitud...*
   a) *inferior a 250,142 mm?*
   b) *superior a 250,276 mm?*
   c) *entre 249,668 mm y 250,08 mm?*
   d) *entre 249,8 mm y 250,2 mm?*
   e) *entre 249,6 mm y 250,4 mm?*
   f) *entre 249,4 mm y 250,6 mm (dentro de las especificaciones requeridas por la automotriz)?*
3. *¿Qué valor de la longitud es superado por el 20% de las barras?*

4. *Se conoce ahora que la longitud de las barras se comporta según el modelo Normal, con $\mu = 250$ mm y que el 25% de las barras mide menos de 249 mm. ¿Podría obtener el valor de la desviación estándar σ?*

Resolución:

En símbolos, el supuesto del enunciado se expresa como $Y \sim N(250; 0,2)$.

1. *Como se dijo anteriormente, la media y el desvío coinciden con los parámetros matemáticos de la distribución Normal. Por lo tanto, la longitud media de las barras es 250 mm y el desvío estándar es 0,2 mm.*

   > ! *Los valores siguientes se obtuvieron con R (Sección 5.6) y con la aplicación `Probability-Distributions` (Sección 5.7).*

2. a) $P(Y < 250,142) = F_Y(250,142) = 0,7611$ *(Figura 5.4.a).*

   ¿Cómo se interpreta el valor 0,7611?

   - *Si se considera una gran cantidad de barras producidas por la empresa, aproximadamente el 76% de las mismas tiene longitud menor a 250,142 mm.*
   - *Si se selecciona una barra al azar, la chance (o probabilidad) de que su longitud sea menor a 250,142 mm es aproximadamente 0,76.*

   Análogamente se obtienen las restantes probabilidades:

   b) $P(Y > 250,276) = 1 - F_Y(250,276) = 0,0838$ *(Figura 5.4.b).*

   c) $P(249,668 < Y < 250,08) = F_Y(250,08) - F_Y(249,668) = 0,6554 - 0,0485 = 0,6069$ *(Figura 5.4.c).*

   d) $P(249,8 < Y < 250,2) = F_Y(250,2) - F_Y(249,8) = 0,8413 - 0,1587 = 0,6826.$

   e) $P(249,6 < Y < 250,4) = F_Y(250,4) - F_Y(249,6) = 0,9772 - 0,0228 = 0,9544.$

   f) $P(249,4 < Y < 250,6) = F_Y(250,6) - F_Y(249,4) = 0,9986 - 0,0013 = 0,9973.$
   Se concluye que aproximadamente el 99,73% de las barras producidas por la empresa cumple con las especificaciones requeridas por la automotriz cuando se supone que la distribución de las longitudes de las barras (en mm) es Normal con $\mu = 250$ y $\sigma = 0,2$.

3. *Se desea determinar la longitud superada por el 20% de las barras. En este caso, a partir de una probabilidad dada, se busca un valor de la variable aleatoria Y, denotado y^*, que la satisfaga.*

   $P(Y > y^*) = 0,20.$

   Se encuentra que $y^* = 250,168$. Es decir que el 20% de las barras tiene longitud superior a 250,168 mm.

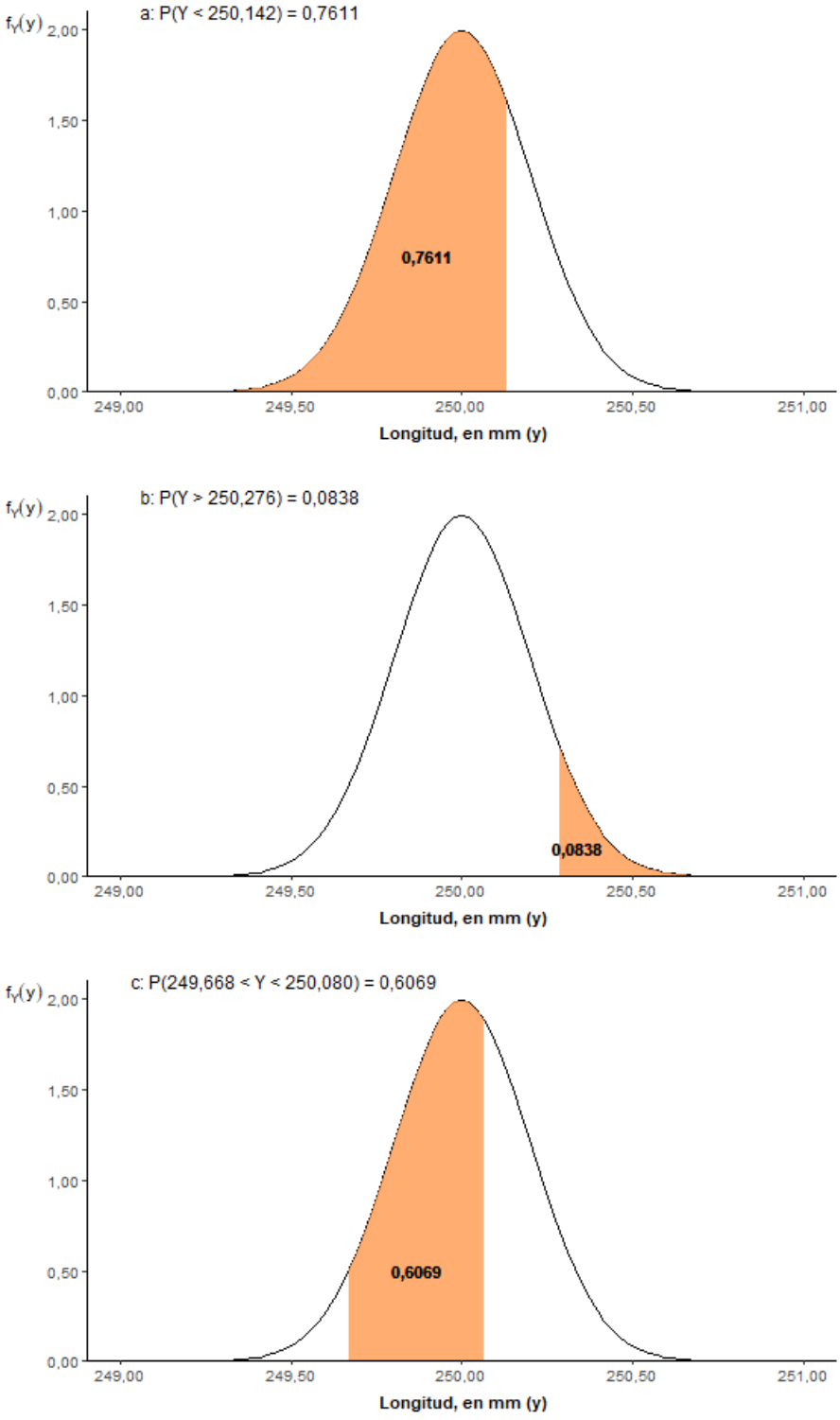


Figura 5.4 *Probabilidades correspondientes a los ítems 2.a, 2.b y 2.c, suponiendo* $Y \sim N(250; 0,2)$

4. *Suponga ahora que respecto de la longitud de las barras se conoce que se comporta según el modelo Normal, con media 250 mm; pero no se conoce el valor del desvío estándar. Se sabe además que el 25 % de las barras mide menos de 249 mm. ¿Podría determinar el valor de* σ *con esa información?*

   En este caso, no está completamente determinado el modelo de comportamiento de la variable Y, por cuanto se desconoce el valor de uno de sus parámetros. Se debe utilizar la variable Z para finalmente

obtener el valor de σ, como se muestra a continuación:
Se conoce que $P(Y < 249) = 0,25$. En esta expresión no aparece el valor de σ; pero sí lo hace si se reescribe en función de la variable estandarizada Z:
$P(Y < 249) = P(\frac{Y-250}{\sigma} < \frac{249-250}{\sigma}) = P(Z < \frac{-1}{\sigma}) = 0,25.$

Si se denomina z^* al valor de la variable estandarizada que verifica la probabilidad anterior, queda que $z^* = \frac{-1}{\sigma}$
Para una distribución $N(0, 1)$ el valor de la variable Z que acumula una probabilidad igual a $0,25$ es $z^* = -0,67$.
Igualando ambas expresiones se puede despejar el valor de σ que resulta $\sigma = \frac{-1}{-0,67} = 1,49$ mm.

En la Sección 5.6.1 se muestra cómo obtener las probabilidades y los valores de la variable correspondientes a los ítems 2, 3 y 4, con R.

Del ítem 4 se desprende la utilidad de obtener probabilidades y percentiles a partir de la variable estandarizada, en aquellos casos en los que se desconoce el valor de alguno de los parámetros de la distribución Normal asociada a la variable en estudio.

Las probabilidades obtenidas en los ítems 2.d, 2.e y 2.f ponen de manifiesto una regla práctica que surge de esta distribución y sirve para recordar ciertas probabilidades. A esta regla se la conoce como **Regla Empírica** y contempla que:

- $P(|Y - \mu_Y| < \sigma_Y) = P(\mu_Y - \sigma_Y < Y < \mu_Y + \sigma_Y) \approx 0,68$, es decir, aproximadamente el 68 % de los valores de la variable aleatoria Y se encuentran en el intervalo $\mu \pm \sigma$;
- $P(|Y - \mu_Y| < 2\,\sigma_Y) = P(\mu_Y - 2\,\sigma_Y < Y < \mu_Y + 2\,\sigma_Y) \approx 0,95$, es decir, aproximadamente el 95 % de los valores de la variable aleatoria Y se encuentran en el intervalo $\mu \pm 2\,\sigma$;
- $P(|Y - \mu_Y| < 3\,\sigma_Y) = P(\mu_Y - 3\,\sigma_Y < Y < \mu_Y + 3\,\sigma_Y) \approx 0,9973$, es decir, aproximadamente el 99,73 % de los valores de la variable aleatoria Y se encuentran en el intervalo $\mu \pm 3\,\sigma$.

En la Figura 5.5 se ilustran estas probabilidades.

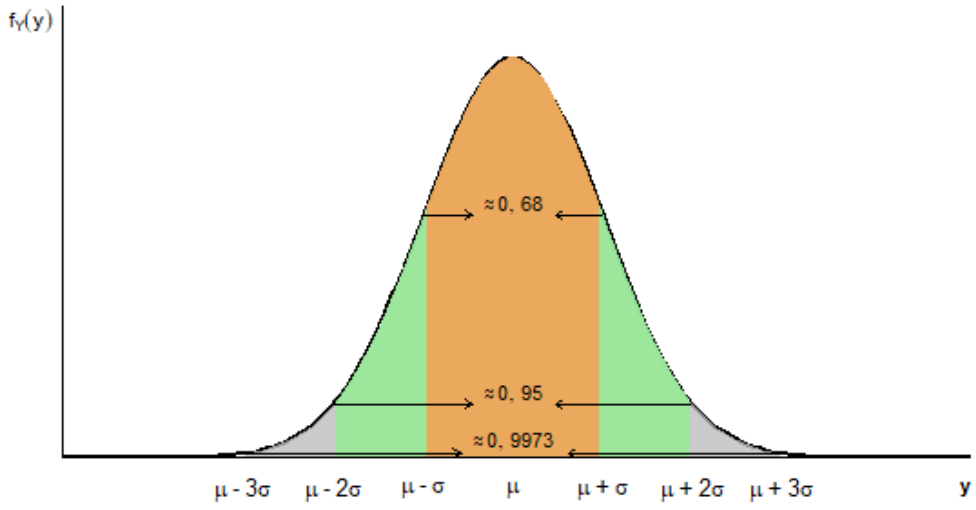


Figura 5.5 *Regla empírica de la distribución Normal*

La regla mencionada es válida para cualquier variable aleatoria con distribución Normal, independientemente de los valores de la media y el desvío estándar.

En la Sección 2.3.6 se abordan aspectos que involucran a las medidas de localización y a las medidas de dispersión. Se plantea que, si se obtienen los valores de la media aritmética y del desvío estándar muestrales, se pueden definir intervalos de la forma $(\bar{y} \pm k\,s)$ y analizar la proporción de unidades de la muestra con valores de la variable dentro de dichos intervalos.

Análogamente, dada una variable aleatoria Y, y conocidos el modelo y los valores de los parámetros, se pueden definir intervalos de la forma $(\mu \pm k\,\sigma)$ y analizar la proporción de unidades de la población con valores de la variable dentro de dichos intervalos.

En el caso particular del modelo Normal, la regla empírica define intervalos para valores de $k = 1, 2$ o 3, que se asocian a una proporción de unidades de la población con valores dentro de ellos, aproximadamente igual a 0,68, 0,95 y 0,9973 respectivamente. También pueden definirse intervalos para otros valores de k, asociados a otros valores de dicha proporción.

Si se conoce que alguna variable de interés relativa a un producto o servicio se comporta según el modelo Normal, con ciertos valores de μ y σ, se puede definir un intervalo de la forma $(\mu \pm k\,\sigma)$ e informar a los usuarios la proporción aproximada de unidades de la población que tendrán valores de la variable en dicho intervalo. En particular, si se usa el valor $k = 3$, se puede informar que aproximadamente el 99,73 % de la población tiene valores en dicho intervalo.



Si se define un valor alto para la proporción de unidades de la población con valores de la variable dentro del intervalo de la forma $(\mu \pm k\,\sigma)$, se puede considerar que el intervalo resultante brinda los **límites naturales del proceso**, los cuales reflejan la variabilidad real del mismo, en condiciones estables. Cuando se informan los límites naturales, se indican además las proporciones de unidades con valores dentro y fuera de ellos.



Problema 1 (pág. 7): *Una empresa produce barras de acero para la industria automotriz. Una característica de calidad importante para este tipo de barras es su longitud, en mm (Y). Se supone que la distribución de esta variable es Normal con media 250 y desviación estándar 0,2.*

Dado que $Y \sim N(250; 0,2)$, los fabricantes de las barras pueden informar que aproximadamente el 99,73 % de las mismas tiene longitud en el intervalo $(250 \pm 0,6)$ mm.

Por otra parte, si respecto de algún producto o servicio un cliente define límites de especificación para alguna característica de interés y el productor o vendedor conoce que la misma se comporta según el modelo Normal, puede ocurrir lo siguiente:

- se pueden determinar los valores ideales para μ y σ de tal forma de satisfacer los requerimientos del cliente. En particular, si las especificaciones se expresan de la forma $(T \pm h)$, donde T es el valor nominal u objetivo y h es la tolerancia (ver página 98), el valor ideal de μ es T; mientras que el valor ideal de σ depende de la mínima proporción de unidades que se pretendan dentro de especificación. Si se pretende como mínimo 0,9973 para la proporción de unidades dentro de

especificaciones (o equivalentemente, se admite como máximo 0,0027 para la proporción de unidades fuera), σ debe ser como máximo $h/3$.

- conocidos los valores de μ y de σ, se puede analizar si es factible cumplir con lo que el cliente pretende. Con esta información se pueden obtener las proporciones de unidades dentro y fuera de especificación y analizar si se deben hacer cambios en el proceso para modificar su centrado y/o su variabilidad.



Problema 1 (pág. 7): *Una empresa metalúrgica de la provincia de Santa Fe fabrica diferentes tipos de barras de acero que luego se utilizan en la construcción y en otras industrias. Una empresa automotriz de la zona requiere uno de los tipos de barras y está dispuesta a comprarlas siempre y cuando sus longitudes se encuentren dentro del intervalo 250 +/- 0,6 mm. En la gerencia de la empresa metalúrgica están interesados en convertirse en proveedores de esta empresa.*

Si la longitud de las barras se comporta según el modelo Normal, ¿cuáles serían los valores ideales de los parámetros de esta distribución para satisfacer los requerimientos de la automotriz?
El cliente plantea, en este caso, un intervalo de especificación para la longitud de las barras en el cual el valor nominal u objetivo, T, vale 250 mm y la tolerancia, h, vale 0,6 mm.
Falta definir la mínima proporción de barras que requiere dentro de esos valores (o la máxima proporción de barras fuera que admite).

- *Suponga que el cliente requiere como mínimo una proporción de barras dentro de especificaciones de 0,9973 (o un máximo fuera de especificaciones de 0,0027).*
  El valor ideal de μ debería coincidir con T, es decir, μ debería ser igual a 250 mm y el valor de σ debería ser como máximo $\frac{0,6}{3} = 0,2$ mm.
- *Suponga que el cliente requiere como mínimo una proporción de barras dentro de especificaciones de 0,95 (o un máximo fuera de especificaciones de 0,05).*
  El valor ideal de μ debería coincidir con T, es decir, μ debería ser igual a 250 mm y el valor de σ debería ser como máximo $\frac{0,6}{2} = 0,3$ mm.



¿Cuánto debería valer el desvío estándar del proceso si el cliente pretendiera como máximo un 1 % de barras fuera de especificaciones? ¿y si pretendiera un máximo de 0,5 %? Queda propuesto determinar esos valores.



Problema 1 (pág. 7): *Una empresa metalúrgica de la provincia de Santa Fe fabrica diferentes tipos de barras de acero que luego se utilizan en la construcción y en otras industrias. Una empresa automotriz de la zona requiere uno de los tipos de barras y está dispuesta a comprarlas siempre y cuando sus longitudes se encuentren dentro del intervalo 250 +/- 0,6 mm. En la gerencia de la empresa metalúrgica están interesados en convertirse en proveedores de esta empresa para lo cual necesitan estudiar si la producción cumple los requerimientos impuestos en relación a las longitudes.*

El cliente plantea, en este caso, un intervalo de especificación para la longitud de las barras en el cual el valor nominal u objetivo, T, vale 250 mm y la tolerancia, h, vale 0,6 mm. Suponga que requiere como mínimo una proporción de barras dentro de especificaciones de 0,9973 (o un máximo fuera de especificaciones de 0,0027).

Ya se monstró anteriormente que el valor ideal de μ debería coincidir con T, es decir, $\mu = 250$ mm y el valor de σ debería ser como máximo $\frac{0,6}{3} = 0,2$ mm.
En la empresa productora conocen que la longitud (Y) de sus barras se comporta normalmente. En la tabla siguiente se presentan diferentes casos, según los valores de los parámetros correspondientes:

Caso	***Modelo asociado a la variable Y: longitud de una barra***	$P(Y < 249,4)$ *(*)*	$P(Y > 250,6)$ *(**)*	$P(Y < 249,4) + P(Y > 250,6)$ *(***)*	$P(249,4 < Y < 250,6)$ *(****)*
a	$Y \sim N(250; 0,17)$	0,0002	0,0002	0,0004	0,9996
b	$Y \sim N(250; 0,4)$	0,0668	0,0668	0,1336	0,8664
c	$Y \sim N(250,4; 0,15)$	0	0,0912	0,0912	0,9088
d	$Y \sim N(250,4; 0,3)$	0,0004	0,2525	0,2529	0,7471

Observaciones:
(*) Se trata de la proporción de barras con longitud menor que 249,4 (mínimo admitido por el cliente)
(**) Se trata de la proporción de barras con longitud mayor que 250,6 (máximo admitido por el cliente)
(***) Se trata de la proporción de barras con longitud fuera de especificaciones.
(****) Se trata de la proporción de barras con longitud dentro de especificaciones.

- *En el caso "a", tanto la longitud promedio como el desvío estándar son ideales ($\mu = 250$ y $\sigma = 0,17 < 0,2$). En consecuencia, la proporción de barras que no satisfacen las especificaciones es menor que la máxima admitida $(0,0004 < 0,0027)$, y la proporción de barras que sí lo hacen resulta mayor que la mínima pretendida $(0,9996 > 0,9973)$.*
- *En el caso "b", si bien la longitud promedio es ideal ($\mu = 250$), el desvío estándar no lo es ($\sigma = 0,4 > 0,2$). Esto genera una proporción de barras que no satisfacen las especificaciones mayor que la máxima admitida $(0,1336 > 0,0027)$ y, por lo tanto, una proporción de barras que sí las satisfacen menor que la mínima pretendida $(0,8664 < 0,9973)$.*
- *En el caso "c", la longitud promedio no es ideal ($\mu = 250,4 \neq 250$) pero el desvío estándar sí lo es ($\sigma = 0,15 < 0,2$). En consecuencia, la proporción de barras que no satisfacen las especificaciones es mayor que la máxima admitida $(0,0912 > 0,0027)$, y la proporción de barras que sí las satisfacen resulta menor que la mínima pretendida $(0,9088 < 0,9973)$.*
- *En el caso "d", ni la longitud promedio ni el desvío estándar son ideales ($\mu = 250,4 \neq 250$ y $\sigma = 0,3 > 0,2$). Por lo tanto, la proporción de barras que no satisfacen las especificaciones es mayor que la máxima admitida $(0,2529 > 0,0027)$, y la proporción de barras que sí las satisfacen resulta menor que la mínima pretendida $(0,7471 < 0,9973)$.*

En la Figura 5.6 se ilustran estos casos.

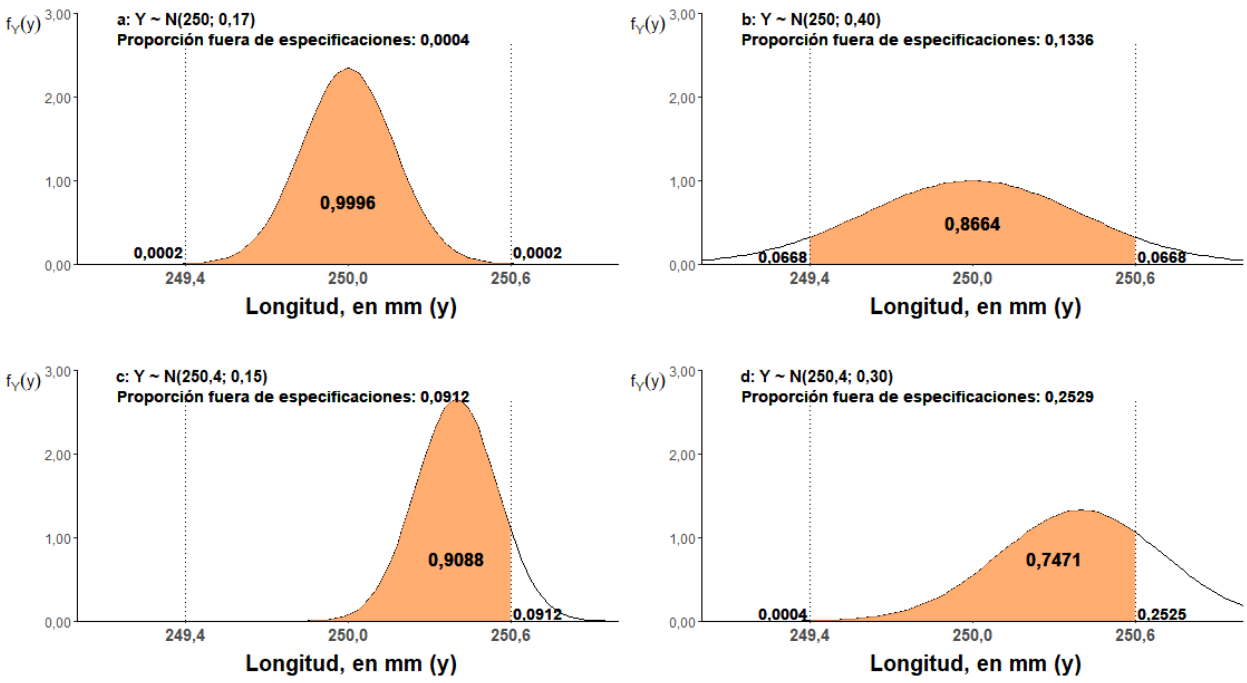


Figura 5.6 *Proporción de barras que satisfacen y que no satisfacen las especificaciones en relación a su longitud, según el caso*

Observe que, cuando el proceso está centrado en el valor objetivo, se presenta una misma proporción de barras con valores menores y mayores que los límites respectivos; mientras que, cuando el proceso no está centrado en dicho valor objetivo, la proporción de valores menores que el mínimo admitido difiere de la proporción de valores mayores que el máximo admitido.

Con lo visto hasta ahora, se puede resolver las Actividades 1 a 6, Sección 5.5.

5.2.2 Distribución Uniforme

La distribución Uniforme es la más simple de las funciones de densidad de probabilidad debido a que se representa con una línea horizontal en todo el recorrido de la variable. Si bien su aplicación no es tan usual, sirve para afianzar muchos de los conceptos vistos en el Capítulo 3.

Una variable aleatoria continua Y tiene una distribución Uniforme de parámetros matemáticos a y b, y se simboliza $Y \sim U(a,b)$, si su función de densidad de probabilidad es:

$$f_Y(y) = \begin{cases} \dfrac{1}{b-a} \;\; si\; a \leq y \leq b \\ \\ 0 \; si \;\; y < a \; o \; y > b \end{cases}$$

donde a y $b \in \mathbb{R}$.

La función de densidad de probabilidad Uniforme se grafica en la Figura 5.7.

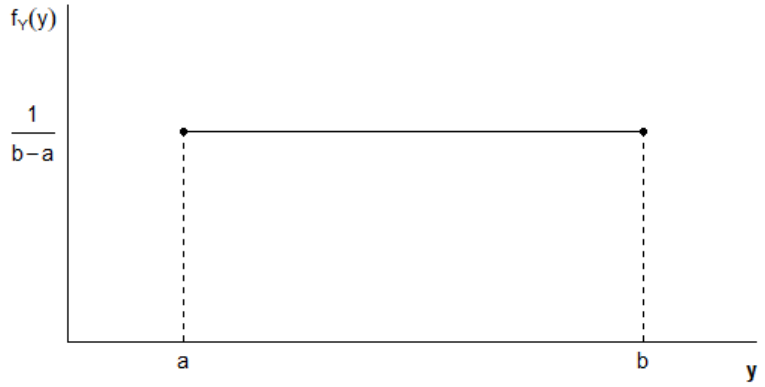


Figura 5.7 *Función de densidad de probabilidad Uniforme, $f_Y(y)$, de parámetros a y b*

Observando la Figura 5.7 se pueden destacar algunas características de la distribución Uniforme:

- para valores en el intervalo $[a,b]$, su función de densidad de probabilidad toma siempre el mismo valor;
- es simétrica respecto a $y = \dfrac{a+b}{2}$;
- la media y la mediana coinciden;
- no presenta un valor máximo. Es decir, ningún valor es moda.

Note que dependiendo de los valores de a y b, la gráfica adopta diferentes alturas y amplitudes. En la Figura 5.8 se ilustran dos distribuciones en particular.

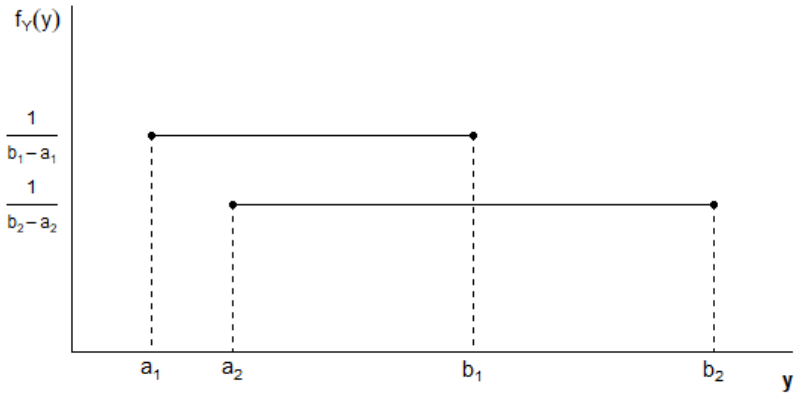


Figura 5.8 *Cambios en la distribución Uniforme al variar a y b*

La función de distribución acumulada Uniforme está definida como:

$$F_Y(y) = \begin{cases} 0 & si \quad y < a \\ \dfrac{y-a}{b-a} & si \quad a \leq y \leq b \\ 1 & si \quad y > b. \end{cases}$$

En la Figura 5.9 se muestra la gráfica de la función de distribución acumulada Uniforme.

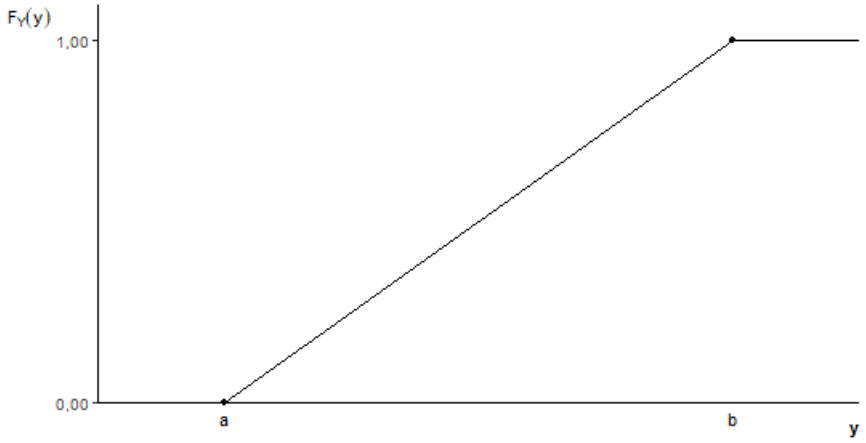


Figura 5.9 *Función de distribución acumulada Uniforme, $F_Y(y)$*

Si $Y \sim U(a,b)$, se demuestra que su media es $E(Y) = \dfrac{a+b}{2}$ y su desvío estándar es $D(Y) = \sqrt{\dfrac{(b-a)^2}{12}}$.

Problema 1 (pág. 7): *Se retoma la situación del Problema 1.*
Suponga que las longitudes de las barras varían entre 249,655 mm y 250,345 mm, de manera uniforme.
Interesa determinar:

1. *¿Cuál es la longitud media de las barras?¿y el desvío estándar?*
2. *¿Qué proporción de las barras fabricadas tienen una longitud ...*
   a) *inferior a 250,142 mm?*
   b) *superior a 250,276 mm?*
   c) *entre 250 mm y 250,138 mm?*
   d) *entre 250,1 mm y 250,238 mm?*
   e) *entre 249,4 mm y 250,6 mm (dentro de las especificaciones requeridas por la automotriz)?*
3. *¿Qué valor de la longitud es superado por el 20% de las barras?*

Resolución:

En símbolos, el supuesto del enunciado se expresa como $Y \sim U(249,655; 250,345)$*, por lo que la función de densidad de probabilidad correspondiente resulta:*

$$f_Y(y) = \frac{1}{250,345 - 249,655} = \frac{1}{0,69}\ ,\ \textit{para}\ 249,655 \leq y \leq 250,345.$$

1. *Aplicando las fórmulas para la media y el desvío estándar:*

   $E(Y) = \frac{249,655+250,345}{2} = 250$ *y* $D(Y) = \sqrt{\frac{(250,345-249,655)^2}{12}} = 0,2.$

   La longitud media de las barras es 250 mm y el desvío estándar es igual a 0,2 mm.

2. a) $P(Y < 250,142) = F_Y(250,142) = \frac{250,142-249,655}{250,345-249,655} = 0,71$ *(Figura 5.10.a).*

   ¿Cómo se interpreta el valor 0,71?

   - *Si se considera una gran cantidad de barras producidas por la empresa, aproximadamente el 71 % de las mismas tiene longitud menor a 250,142 mm.*
   - *Si se selecciona una barra al azar, la chance (o probabilidad) de que su longitud sea menor a 250,142 mm es aproximadamente 0,71.*

   Análogamente se obtienen las restantes probabilidades:

   b) $P(Y > 250,276) = 1 - F_Y(250,276) = 1 - 0,9 = 0,1$ *(Figura 5.10.b).*

   c) $P(250 < Y < 250,138) = F_Y(250,138) - F_Y(250) = 0,70 - 0,50 = 0,20$ *(Figura 5.10.c).*

   d) $P(250,1 < Y < 250,238) = F_Y(250,238) - F_Y(250,1) = 0,845 - 0,645 = 0,20$ *(Figura 5.10.d).*

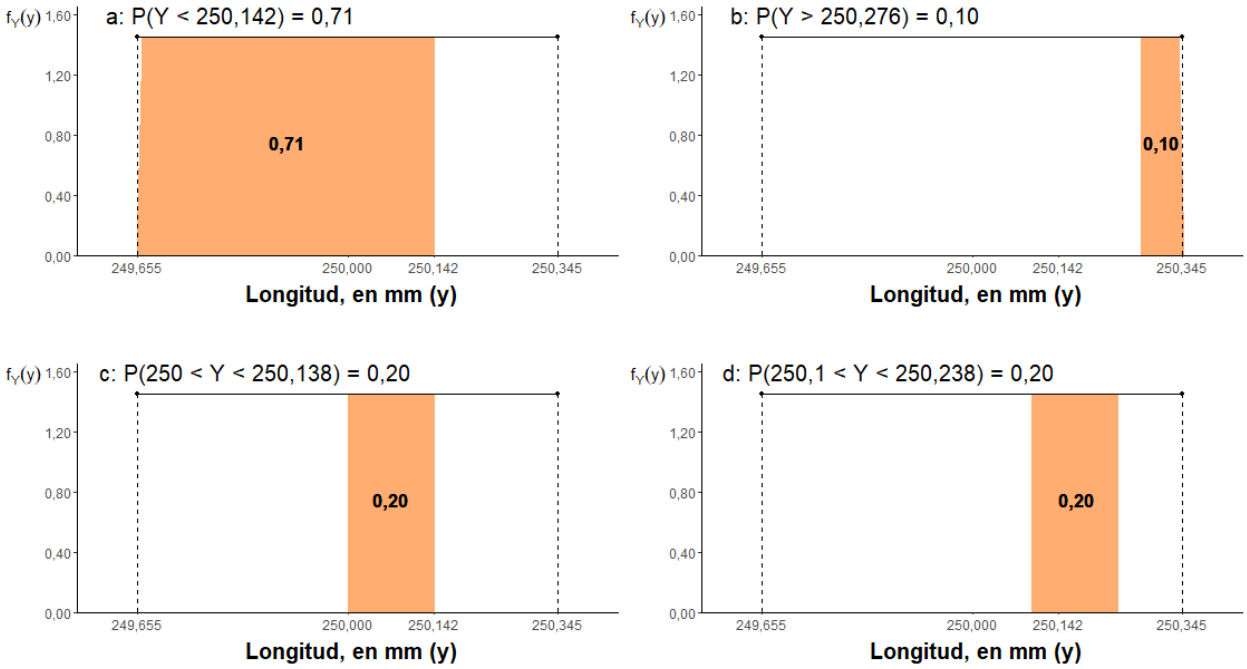


Figura 5.10 *Probabilidades correspondientes a los ítems 2.a, 2.b, 2.c y 2.d, suponiendo* $Y \sim U(249,655; 250,345)$

e) $P(249,4 < Y < 250,6) = F_Y(250,6) - F_Y(249,4) = 1 - 0 = 1$.
Se concluye que la totalidad de las barras producidas por la empresa cumple con las especificaciones requeridas por la automotriz cuando se supone que la distribución de las longitudes de las barras es Uniforme con a = 149,655 mm y b = 250,345 mm.

3. *Se desea determinar la longitud superada por el 20% de las barras. En este caso, a partir de una probabilidad dada, se busca un valor de la variable aleatoria Y, denotado y^*, que la satisfaga.*

   $P(Y > y^*) = 0,20$*; entonces,* $P(Y \leq y*) = F_Y(y^*) = 0,80$.

   En este caso, $F_Y(y^*) = \frac{y^* - 249,655}{250,345 - 249,655} = 0,80$.
   Entonces, $y^* = 0,80\,(250,345 - 249,655) + 249,655 = 250,207$ *mm.*
   Es decir que el 20% de las barras tienen longitudes superiores a 250,207 mm.

En la Sección 5.6.2 se muestra cómo obtener las probabilidades y los valores de la variable correspondientes a los ítems 2 y 3, con R.



¿Por qué las probabilidades calculadas en los ítems 2.c y 2.d son iguales?

Una propiedad de la distribución Uniforme $U(a, b)$ es que, para intervalos de valores de la variable de igual amplitud contenidos en el intervalo (a, b), las probabilidades son iguales, independientemente de la localización de dichos intervalos. Es decir, en esta distribución, la probabilidad de que la variable asuma valores en un determinado intervalo dentro de (a, b) no depende de la ubicación del intervalo, sino únicamente de su longitud.



Con lo visto hasta ahora, se puede resolver las Actividades 7 a 10, Sección 5.5.

5.2.3 Distribución Triangular

La distribución Triangular debe su nombre a que su función de densidad de probabilidad junto con el eje de abscisas forman un triángulo.

Se utiliza como una aproximación cuando no se dispone de información suficiente sobre el comportamiento de una variable, pero se conocen el valor mínimo, el valor máximo y la moda. Por ejemplo, si a partir de la experiencia se conocen los tiempos mínimo y máximo que tarda una computadora en responder a una orden, así como el tiempo que ocurre con mayor frecuencia, es posible emplear la distribución triangular para modelizar el tiempo de respuesta.

A veces se utiliza para modelar la variabilidad de errores aleatorios en la medición de un valor predeterminado o conocido.

Una variable aleatoria continua Y tiene una distribución Triangular de parámetros matemáticos a, b y c, y se simboliza $Y \sim Tri(a, b, c)$, si su función de densidad de probabilidad es:

$$f_Y(y) = \begin{cases} \dfrac{2(y-a)}{(b-a)(c-a)} & si \quad a \leq y \leq c \\ \dfrac{2(b-y)}{(b-a)(b-c)} & si \quad c < y \leq b \\ 0 & si \quad y \notin [a,b] \end{cases}$$

donde a, b, $c \in \mathbb{R}$

La función de densidad de probabilidad Triangular se muestra en la Figura 5.11.

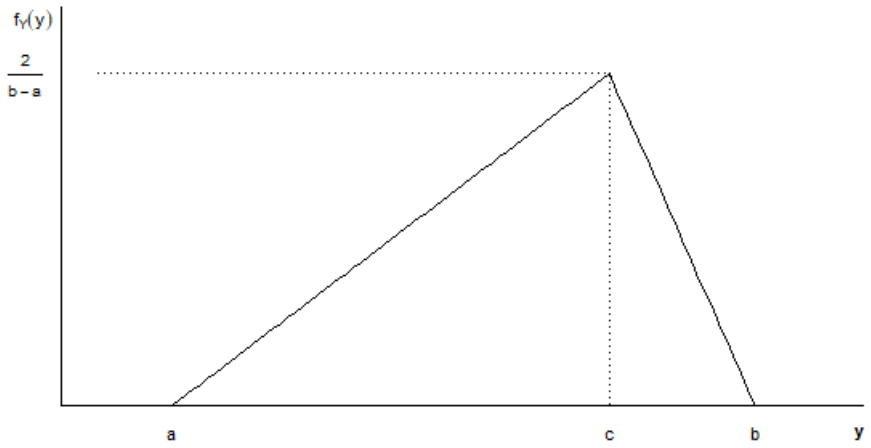


Figura 5.11 *Función de densidad de probabilidad Triangular, $f_Y(y)$, de parámetros a, b y c*

Observando la Figura 5.11 se pueden destacar algunas características de la distribución Triangular:

- junto al eje de absisas tiene forma de triángulo;
- es simétrica respecto de c cuando este valor es el punto medio del intervalo $[a, b]$;
- en $y = c$ presenta el valor máximo. Es decir, la moda es c.

Note que dependiendo de los valores de a, b y c, la gráfica forma diferentes triángulos. En la Figura 5.12 se ilustran dos distribuciones en particular.

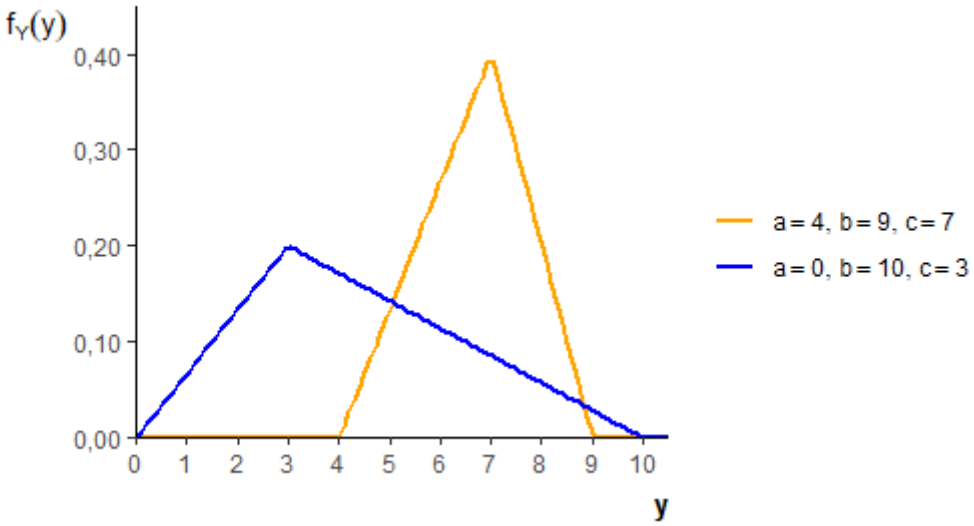


Figura 5.12 *Cambios en la distribución Triangular al variar a b y c*

La función de distribución acumulada Triangular está definida como:

$$F_Y(y) = \begin{cases} 0 & si \quad y \leq a \\ \dfrac{(y-a)^2}{(b-a)(b-c)} & si \quad a < y \leq c \\ 1 - \dfrac{(b-y)^2}{(b-a)(b-c)} & si \quad c < y < b \\ 1 & si \quad y \geq b. \end{cases}$$

En la Figura 5.13 se muestra la gráfica de la función de distribución acumulada Triangular.

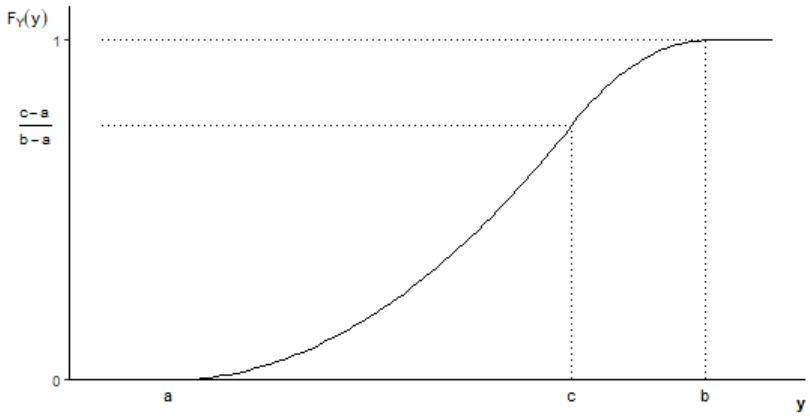


Figura 5.13 *Función de distribución acumulada Triangular,* $F_Y(y)$

Si $Y \sim Tri(a, b, c)$, se demuestra que su media es $E(Y) = \dfrac{a+b+c}{3}$ y su desvío estándar es $D(Y) = \sqrt{\dfrac{a^2+b^2+c^2-ab-ac-bc}{18}}$.

Un caso particular de este modelo es la distribución Triangular simétrica, donde c es el punto medio del intervalo $[a, b]$, es decir $c = \dfrac{a+b}{2}$. De este modo, la función de densidad de probabilidad de Y queda en función de los parámetros matemáticos a y b y resulta:

$$f_Y(y) = \begin{cases} \dfrac{2(y-a)}{(b-a)\left(\frac{a+b}{2}-a\right)} & si \quad a \leq y \leq \frac{a+b}{2} \\ \dfrac{2(b-y)}{(b-a)\left(b-\frac{a+b}{2}\right)} & si \quad \frac{a+b}{2} < y \leq b \\ 0 & si \quad y \notin [a,b]. \end{cases}$$

Por lo que la función de distribución acumulada Triangular simétrica es:

$$F_Y(y) = \begin{cases} 0 & si \quad y \leq a \\ \dfrac{(y-a)^2}{(b-a)\left(b-\frac{a+b}{2}\right)} & si \quad a < y \leq \frac{a+b}{2} \\ 1 - \dfrac{(b-y)^2}{(b-a)\left(b-\frac{a+b}{2}\right)} & si \quad \frac{a+b}{2} < y < b \\ 1 & si \quad y \geq b. \end{cases}$$

Reemplazando c por $\dfrac{a+b}{2}$ en las fórmulas de $E(Y)$ y $D(Y)$ mediante paso algebraico se obtiene que la media de la distribución Triangular simétrica es $E(Y) = \dfrac{a+b}{2}$ y el desvío estándar es $D(Y) = \sqrt{\dfrac{(b-a)^2}{24}}$.

! Observe que las distribuciones Uniforme y Triangular simétrica tienen la misma media. Esta última distribución tiene desvío estándar menor debido a que más valores de Y están concentrados alrededor de su media.



Problema 1 (pág. 7): *Se retoma la situación del Problema 1.*

Suponga que las longitudes de las barras varían entre 249,655 mm y 250,345 mm, y que no se distribuyen de manera uniforme, sino que pueden modelarse mediante una distribución Triangular simétrica.

Interesa determinar:

1. *¿Cuál es la longitud media de las barras?¿y el desvío estándar?*
2. *¿Qué proporción de las barras fabricadas tienen una longitud dots*
   a) *inferior a 250,142 mm?*
   b) *superior a 250,276 mm?*
   c) *entre un 250 mm y 250,138 mm?*
   d) *entre 249,4 mm y 250,6 mm (dentro de las especificaciones requeridas por la automotriz)?*
3. *¿Qué valor de la longitud es superado por el 20% de las barras?*

Resolución:

En símbolos, el supuesto del enunciado se expresa como $Y \sim Tri(249,655;\ 250,345)$*, por lo que la función de densidad de probabilidad correspondiente resulta:*

$$f_Y(y) = \begin{cases} \frac{2(y-249,655)}{0,24} & si \quad 249,655 \leq y \leq 250 \\ \\ \frac{2(250,345-y)}{0,24} & si \quad 250 < y < 250,345. \end{cases}$$

1. *Aplicando las fórmulas para la media y el desvío estándar:*

   $E(Y) = \frac{249,655+250,345}{2} = 250$ *y* $D(Y) = \sqrt{\frac{(250,345-249,655)^2}{24}} = \sqrt{0,02} = 0,141.$

   La longitud media de las barras es 250 mm y el desvío estándar es 0,141 mm.

2. a) $P(Y < 250,142) = F(250,142) = 1 - \frac{(250,345-250,142)^2}{(250,345-249,655).(250,345-\frac{249,655+250,345}{2})} = 0,827$ *(Figura 5.14.a).*

   ¿Cómo se interpreta el valor 0,827?

   - *Si se considera una gran cantidad de barras producidas por la empresa, aproximadamente el 83% de las mismas tiene longitud menor a 250,142 mm.*
   - *Si se selecciona una barra al azar, la chance (probabilidad) de que su longitud sea menor a 250,142 mm es aproximadamente 0,83.*

   Análogamente se obtienen las restantes probabilidades:

   b) $P(Y > 250,276) = 1 - F(250,276) = 1 - 0,98 = 0,02$ *(Figura 5.14.b).*

   c) $P(250 < Y < 250,138) = F_Y(250,138) - F_Y(250) = 0,82 - 0,50 = 0,18$ *(Figura 5.14.c).*

   d) $P(249,4 < Y < 250,6) = F_Y(250,6) - F_Y(249,4) = 1 - 0 = 1.$
   Se concluye que la totalidad de las barras producidas por la empresa cumple con las especificaciones requeridas por la automotriz cuando se supone que la distribución de las longitudes de las barras es Triangular Simétrica con a = 149,655 mm y b = 250,345 mm.

3. *Se desea conocer la longitud superada por el 20% de las barras. En este caso, a partir de una probabilidad dada, se busca un valor de la variable Y, denotado* y^**, que la satisfaga.*

   $P(Y > y^*) = 0,20$*; entonces,* $P(Y \leq y*) = F_Y(y^*) = 0,80.$

   En este caso:

$$F_Y(y^*) = 1 - \frac{(250,345 - y^*)^2}{(250,345 - 249,655)(250,345 - \frac{249,655+250,345}{2})} = 0,80.$$

Entonces, $y^* = 0,80\,(250,345 - 249,655) + 249,655 = 250,127$.
Es decir que el 20% de las barras tienen longitudes superiores a 250,127 mm.

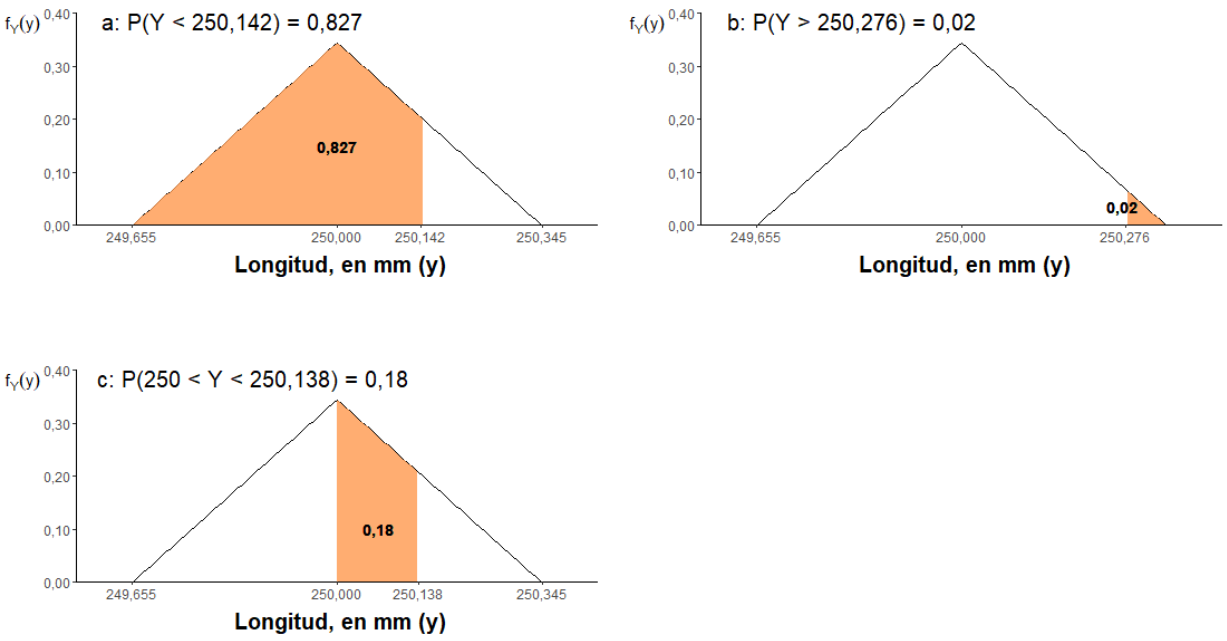


Figura 5.14 *Probabilidades correspondientes a los ítems 2.a, 2.b y 2.c, suponiendo* $Y \sim Tri(249,655; 250,345)$



Con lo visto hasta ahora, se puede resolver las Actividades 11 a 14, Sección 5.5.

5.2.4 Distribución Exponencial

La distribución Exponencial se utiliza para modelar tiempos o distancias entre eventos aleatorios sucesivos, como llamadas que entran a un conmutador, accidentes que ocurren en una esquina, quejas de clientes sobre cierto producto, llegada de pedidos a un centro de distribución o fallas superficiales en alambres, entre otros.

En particular, desempeña un papel importante en el área de la teoría de la confiabilidad, ya que permite modelar la vida útil (duración o tiempo hasta la falla) de componentes, sistemas, máquinas o piezas. Se aplica especialmente cuando la falla ocurre únicamente por causas aleatorias y no por desgaste; en estos casos, es razonable suponer que el buen funcionamiento en períodos previos no modifica la probabilidad de falla en el siguiente período.

Una variable aleatoria continua Y tiene una distribución Exponencial de parámetro matemático α, y se simboliza $Y \sim Exp(\alpha)$, si su función de densidad de probabilidad es:

$$f_Y(y) = \begin{cases} \alpha e^{-\alpha y} & si \quad y \geq 0 \\ 0 & si \quad y < 0. \end{cases}$$

donde $\alpha > 0$

La función de densidad de probabilidad Exponencial se muestra en la Figura 5.15.

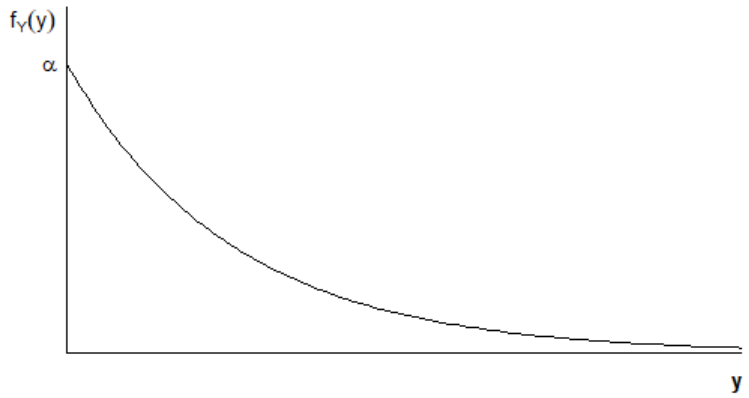


Figura 5.15 *Función de densidad de probabilidad Exponencial, $f_Y(y)$, de parámetro α*

Observando la Figura 5.15 se pueden destacar algunas características de la distribución Exponencial:

- es asimétrica a la derecha;
- la media resulta mayor a la mediana;
- en $y = 0$ presenta el valor máximo. Es decir, la moda es cero.

En la Figura 5.16 se ilustran tres distribuciones en particular. Note que dependiendo del valor de α, la gráfica adopta diferentes formas.

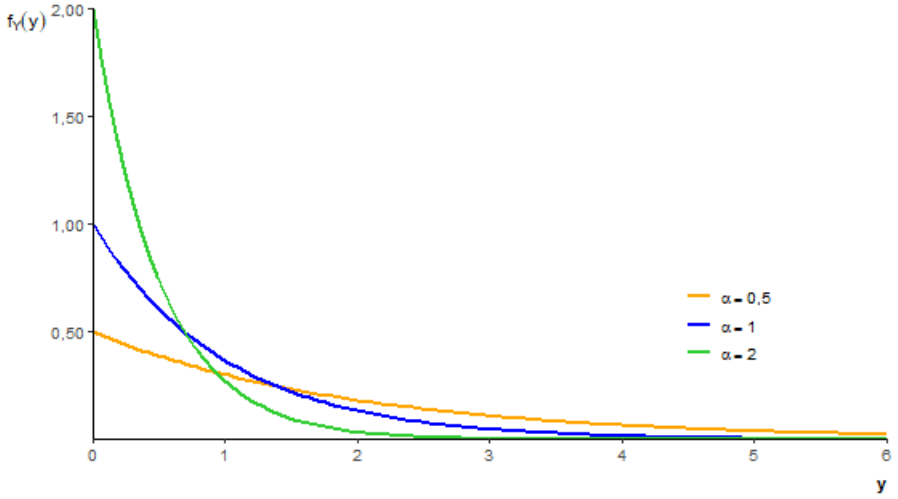


Figura 5.16 *Cambios en la distribución Exponencial al variar α*

La función de distribución acumulada Exponencial está definida como:

$$F_Y(y) = \begin{cases} 0 & si \quad y < 0 \\ 1 - e^{-\alpha y} & si \quad y \geq 0 \end{cases}$$

Por lo tanto, $P(Y > y) = e^{-\alpha y}$.

En la Figura 5.17 se muestra la gráfica de la función de distribución acumulada Exponencial.

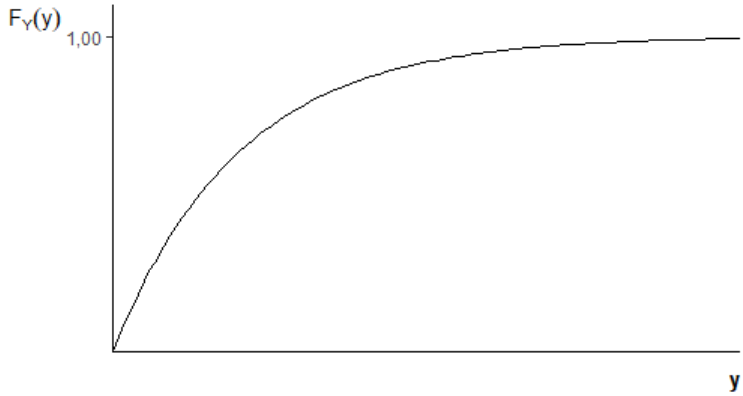


Figura 5.17 *Función de distribución acumulada Exponencial,* $F_Y(y)$

Si $Y \sim Exp(\alpha)$, se demuestra que su media es $E(Y) = \frac{1}{\alpha}$ y su desvío estándar es $D(Y) = \frac{1}{\alpha}$.

Note que el valor esperado de Y y el desvío estándar coinciden y son iguales al recíproco del parámetro α.

Como se menciona al principio del apartado, se puede describir el comportamiento de la duración de una componente o un sistema que no falla como consecuencia del desgaste mediante la distribución Exponencial. Es decir que, su falla no sea consecuencia de su uso sino de algún evento aleatorio. Por ejemplo, para la duración de un neumático o batería de un automovil esta distribución no sería oportuna.

Esta propiedad se conoce como **la propiedad de la falta de memoria.** Se refiere a que, si el tiempo transcurrido hasta la ocurrencia de un evento es mayor que s (con lo cual $Y > s$), entonces la probabilidad de que transcurra un tiempo adicional t sin que ocurra un evento (y, por tanto, haya pasado un tiempo total $Y > t + s$), es igual a la probabilidad de que transcurra un tiempo mayor que t, contabilizado desde el inicio del tiempo de observación. Por lo tanto, transcurrido un tiempo s sin haber observado la ocurrencia del evento, se puede empezar a contabilizar el tiempo de nuevo, olvidando lo ocurrido (falta de memoria) hasta ese momento.

Formalizando, para cualquier $s, t > 0$, $P(Y > s+t \mid Y > s)$, se tiene:

$$P(Y > s+t \mid Y > s) = \frac{P(Y > s+t)}{P(Y > s)} = \frac{e^{-\alpha(s+t)}}{e^{-\alpha s}} = e^{-\alpha t}$$

.

Por tanto, $P(Y > s+t \mid Y > s) = P(Y > t)$.

Problema 8 (pág. 8): *Una empresa se especializa en realizar ensayos de vida en dispositivos electrónicos. Un usuario está interesado en utilizar estos dispositivos para construir circuitos especiales y por seguridad los reemplazaría ante la ocurrencia de una falla. Este usuario pretende que la mayoría de ellos fallen después de las 150 h y realiza una consulta en la empresa especializada para ver si esto es razonable.* ***Suponga que la distribución de la variable duración del dispositivo (o tiempo hasta la falla), Y, es Exponencial con $\alpha = 0,002$ fallas/h.***

Interesa determinar:

1. *¿Cuál es la duración media de estos dispositivos? ¿y el desvío estándar?*
2. *¿Qué proporción de dispositivos electrónicos fallan*
   a) *antes de las 100 h?*
   b) *después de las 150 h (pretensión del usuario)?*
   c) *entre las 150 h y 160 h?*
3. *¿Qué proporción de los dispositivos con duración mayor a 150 h tienen una duración mayor a 160 h?*
4. *¿Qué valor de la duración es superado por el 20% de los dispositivos?*

Resolución:

En símbolos, el supuesto del enunciado se expresa como $Y \sim Exp(0,002)$ por lo que la función de densidad de probabilidad correspondiente resulta:

$$f_Y(y) = \begin{cases} 0,002\, e^{-0,002 \cdot y} & si \quad y \geq 0 \\ \\ 0 & si \quad y < 0. \end{cases}$$

1. *Aplicando las fórmulas para la media y el desvío estándar:*
   $E(Y) = \frac{1}{0,002} = 500$ *y* $D(Y) = \frac{1}{0,002} = 500$.
   La duración media de los dispositivos es 500 h y el desvío estándar de las duraciones es 500 h.
   El valor del desvío estándar es muy grande, en un proceso de fabricación no es deseable que las duraciones de los dispositivos se desvíen tanto de su valor medio.
2. a) $P(Y < 100) = F(100) = 1 - e^{-0,002 \cdot 100} = 0,181$ *(Figura 5.18.a).*

   ¿Cómo se interpreta el valor 0,1813?

   - *Si se ensayan una gran cantidad de dispositivos, aproximadamente el 18% de los mismos tiene duración menor a 100 h.*
   - *Si se selecciona un dispositivo al azar, la chance (o probabilidad) de que su duración sea menor a 100 h es aproximadamente 0,18.*

   Análogamente se obtienen las restantes probabilidades:

   b) $P(Y > 150) = 1 - F(150) = 1 - [1 - e^{-0,002 \cdot 150}] = 0,7408$ *(Figura 5.18.b).*
   Se concluye que aproximadamente el 74% de los dispositivos que se ensayan cumplen con lo requerido por el usuario cuando se supone que la distribución de las duraciones de los dispositivos es Exponencial con $\alpha = 0,002$.

   c) $P(150 < Y < 160) = F_Y(160) - F_Y(150) = [1 - e^{-0,002 \cdot 160}] - [1 - e^{-0,002 \cdot 150}] = 0,2739 - 0,2592 = 0,0147$ *(Figura 5.18.c).*

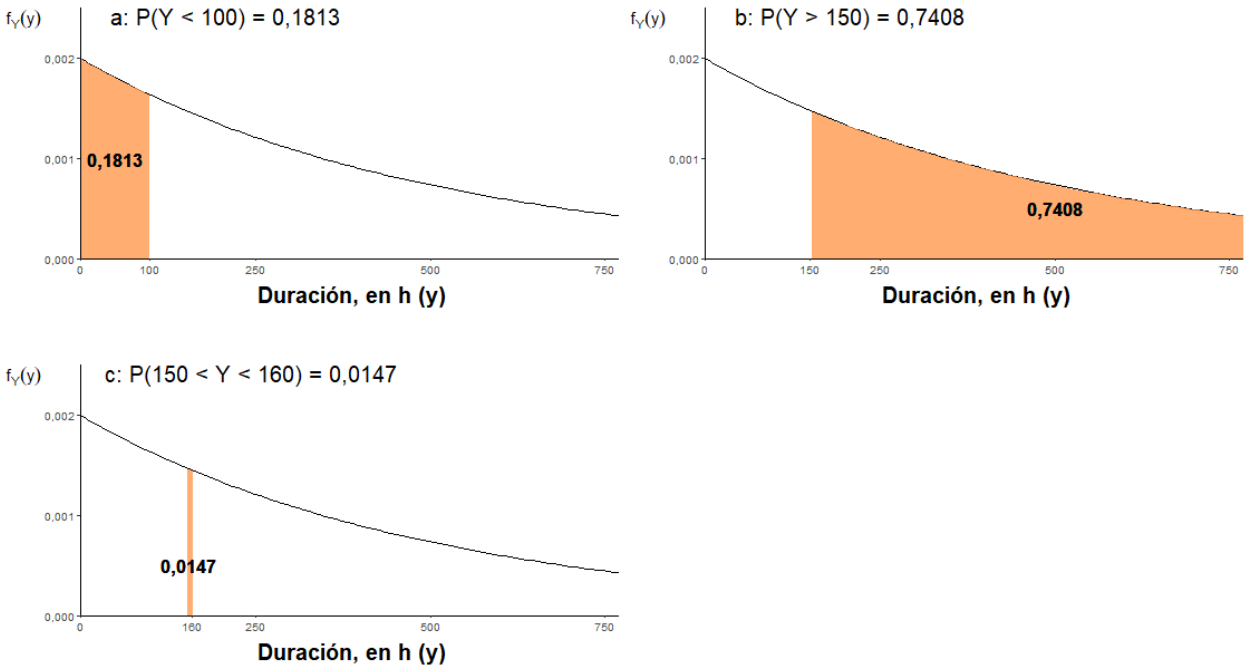


Figura 5.18 *Probabilidades correspondientes a los ítems 2.a, 2.b y 2.c, suponiendo* $Y \sim Exp(0,002)$

3. *En este caso la probabilidad solicitada es una probabilidad condicional en la cual se puede aplicar la propiedad de la falta de memoria que tiene la distribución Exponencial. Por lo tanto,*

$$P(Y > 160|Y > 150) = P(Y > 10) = e^{-0,002 \cdot 10} = 0,9802.$$

Esto se interpreta como: de los dispositivos que tienen duración mayor a 150 h, aproximadamente el 98% tienen una duración mayor a 160 h. Es decir, el 98% van a durar como mínimo 10 h más.

4. *Se desea conocer la duración superada por el 20% de los dispositivos. En este caso, a partir de una probabilidad dada, se busca un valor de la variable Y, denotado* y^**, que la satisfaga.*
$P(Y > y^*) = 0,20$; entonces, $P(Y \leq y*) = F_Y(y^*) = 0,80$.
En este caso, $F_Y(y^*) = 1 - e^{-0,002y^*} = 0,80$.
Entonces, $y^* = \frac{ln(0,20)}{-0,002} = 804,72$.
Es decir, el 20% de los dispositivos tienen duraciones superiores a 804,72 h.



Con lo visto hasta ahora, se puede resolver las Actividades 15 a 18, Sección 5.5.

5.3 Variables aleatorias discretas

5.3.1 Distribución Bernoulli

Esta distribución es importante porque permite estudiar el comportamiento poblacional de una variable originalmente cualitativa, a través de una nueva variable que resulta cuantitativa discreta. Además, constituye la base para otras distribuciones de probabilidades para variables aleatorias discretas, como la Binomial y la Geométrica, que se presentan más adelante.

Suponga que en las unidades de una población se observa una variable cualitativa de la cual interesa analizar una categoría particular. Por ejemplo, en una empresa metalúrgica una de las variables en

estudio es la calidad de las barras que produce, cuyas categorías son “mala”, “regular” y “buena”, siendo la categoría de interés “mala”; o bien, en una distribuidora con un gran número de empleados, se estudia el máximo nivel educativo alcanzado, siendo la categoría de interés “estudios secundarios completos”.

En ambos casos, la variable cualitativa puede describirse en términos de dos categorías: la de interés y el conjunto formado por todas las demás categorías posibles. En la empresa metalúrgica, esto corresponde a distinguir entre barras de mala calidad y barras que no presentan dicha característica; mientras que en la distribuidora se distingue entre empleados con estudios secundarios completos y aquellos con un nivel educativo diferente.

Se define π como la proporción poblacional de unidades que presentan la categoría de interés. En consecuencia, la proporción de unidades que presentan alguna de las restantes categorías es $(1-\pi)$.

A partir de la variable cualitativa original se define una variable aleatoria discreta Y, asociada a la observación de cada unidad, que toma el valor 1 cuando la unidad presenta la categoría de interés y el valor 0 cuando presenta alguna de las restantes categorías. Por lo tanto, el recorrido de Y es $R_Y = \{0\ ,\ 1\}$ y se cumple que $P(Y = 1) = \pi$.

Una variable aleatoria discreta Y tiene una distribución Bernoulli con parámetro matemático π, y se simboliza $Y \sim Be(\pi)$, si su función de probabilidad puntual es:

$$p_Y(y) = \pi^y (1-\pi)^{(1-y)} \; con \; y = 0 \, o \, 1$$

donde $\pi \in [0,1]$.

Note que $P(Y = 1) = p_Y(1) = \pi$. Es decir, la probabilidad de que Y sea igual a 1 se puede obtener valorizando la función de probabilidad puntual en 1, cuyo resultado es la proporción de unidades de la población que presentan la categoría de interés de la variable cualitativa.

La función de probabilidad puntual Bernoulli se muestra en la Figura 5.19.

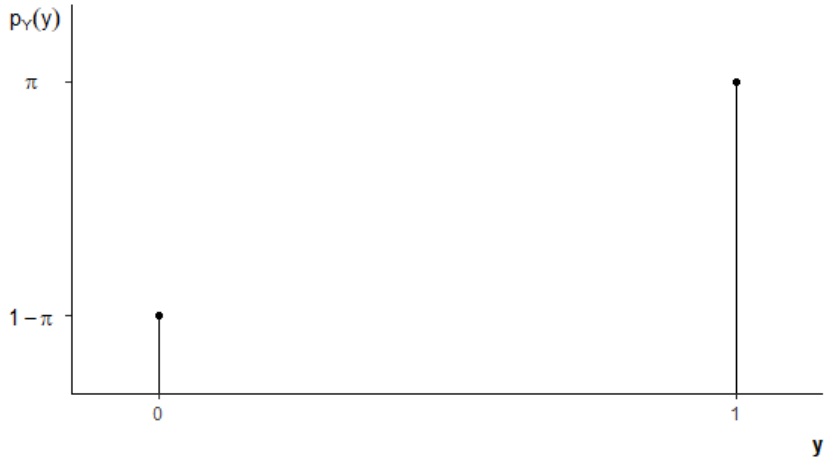


Figura 5.19 *Función de probabilidad puntual Bernoulli, $p_Y(y)$, de parámetro π*

Observando la Figura 5.19 se pueden destacar algunas particularidades de la distribución de Bernoulli:

- la mediana y la moda coinciden;
- el rango siempre es 1.

Si $Y \sim Be(\pi)$, se demuestra que su media es $E(Y) = \pi$ y su desvío estándar es $D(Y) = \sqrt{\pi(1-\pi)}$.

El parámetro matemático, π, es la proporción de unidades de la población que presentan la categoría de interés. Este valor coincide con la media de la variable aleatoria Y.



Note que la variable original, a partir de la cual se construye la variable Bernoulli, puede ser cuantitativa. En ese caso, mediante un criterio de clasificación, se define una categoría de interés y su categoría complementaria, dando lugar a una variable aleatoria con dos únicos valores posibles, como se observa en el ejemplo que se presenta a continuación.



Problema 1 (pág. 7): *Se retoma la situación del Problema 1, en la que una empresa automotriz requiere que barras de cierto tipo cumplan con las siguientes especificaciones: su longitud debe estar comprendida entre 249,4 mm y 250,6 mm.*
Interesa observar si las barras fabricadas por la empresa metalúrgica cumplen (categoría de interés) o no con esas especificaciones.
Se define la variable dicotómica Y, que puede denominarse "cumplimiento de especificaciones", la cual asume el valor 1 si la barra cumple con las especificaciones, es decir, si presenta longitud entre 249,4 mm y 250,6 mm y 0 en caso contrario.
***Para esta producción, se supone que* la proporción de barras que cumplen con las especificaciones es 0,9**.

A partir de esta información, interesa determinar:

1. *¿Cuáles son la media y el desvío estándar de la variable aleatoria dicotómica Y que se define a partir de considerar como categoría de interés que la barra cumpla con las especificaciones?*
2. *¿Qué proporción de las barras fabricadas no cumple con las especificaciones?*

Resolución:

En símbolos, y considerando el supuesto del enunciado, se puede expresar $Y \sim Be(0,90)$ *por lo que su función de probabilidad puntual resulta:*

$$p_Y(y) = 0,9^y\,(1-0,9)^{(1-y)} \text{ con } y = 0 \text{ o } 1$$

1. *Aplicando las fórmulas para la media y el desvío estándar:*

   $E(Y) = 0,9$ y $D(Y) = \sqrt{0,9\,(1-0,9)} = 0,3$.

   La media de Y es 0,9 y el desvío estándar es igual a 0,3.

2. $P(Y=0) = 1-0,9 = 0,1$ *(Figura 5.20).*

   ¿Cómo se interpreta el valor 0,1?

- *Si se considera una gran cantidad de barras producidas por la empresa, aproximadamente el 10% de las mismas no cumplen con las especificaciones.*
- *Si se selecciona una barra al azar, la chance (o probabilidad) de que no cumpla con las especificaciones es aproximadamente 0,1.*

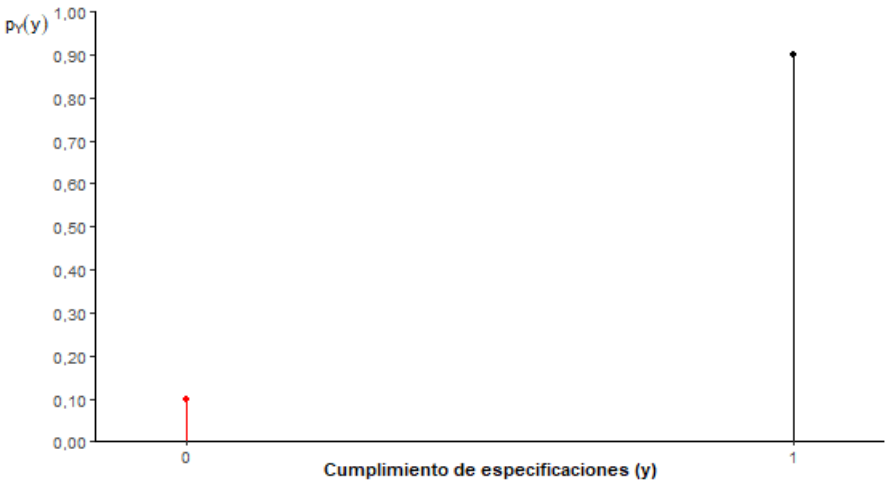


Figura 5.20 *Probabilidad correspondiente al ítem 2, suponiendo* $Y \sim Be(0,90)$



Con lo visto hasta ahora, se puede resolver las Actividades 19 a 21, Sección 5.5.

5.3.2 Distribución Binomial

Considere que se selecciona una muestra de n unidades de una población y que, para cada una de ellas, se registra si presenta cierta categoría de interés o no. Suponga que el muestreo se realiza de manera que los resultados sean independientes entre sí[1], garantizando que la probabilidad π de presentar la categoría de interés sea constante en cada extracción.

Se define la variable aleatoria Y como el número de unidades que presentan la categoría de interés en una muestra de tamaño n. El recorrido de Y es $R_Y = \{0, 1, 2, \ldots, n\}$, es decir, esta variable asume valores enteros entre 0 y n.

Note que, por la forma en que se define Y, la medición se realiza sobre grupos de n unidades. Esto implica que **la unidad de análisis asociada a una variable aleatoria Binomial es una muestra de tamaño** n.

¿Cómo se puede obtener la función de probabilidad puntual para una variable Binomial?

Considere a modo de ejemplo que se selecciona una muestra aleatoria de $n = 3$ unidades de una población. La Figura 5.21 muestra el árbol de probabilidad (Ver Sección 4.9.2) que va a permitir obtener la distribución de probabilidades de Y. Cada ramificación representa las opciones que tiene una unidad de presentar la categoría de interés o no.

En el árbol, los sucesos se definen de la siguiente manera:

[1] La independencia se garantiza muestreando a partir de una población infinita o de una población finita con reposición.

- A: la unidad presenta la categoría de interés,
- $\bar{A}$: la unidad no presenta la categoría de interés.

Note que $P(A) = \pi$ y $P(\bar{A}) = 1 - \pi$.

A modo de ejemplo, observe que la segunda trayectoria corresponde al caso donde la primera unidad seleccionada presenta la categoría de interés, la segunda unidad seleccionada también y la tercera no $(A, A, \bar{A})$. Por lo tanto, la variable aleatoria Y vale 2. Observe que Y también vale 2 para la tercera y la quinta trayectorias donde se presentan $(A, \bar{A}, A)$ y $(\bar{A}, A, A)$, respectivamente.

Sumando las probabilidades de estas tres trayectorias, se puede encontrar la probabilidad que Y sea igual a 2. Es decir, $P(Y = 2) = p_Y(2) = \pi^2(1-\pi) + \pi^2(1-\pi) + \pi^2(1-\pi) = 3\pi^2(1-\pi)$. El valor 3 indica la cantidad de trayectorias donde Y vale 2 y representa la cantidad de posibilidades de que entre las 3 unidades seleccionadas, 2 de ellas presenten la categoría de interés. Esto se puede escribir como el número combinatorio de 3 en 2, $\binom{3}{2}$.

De la misma forma se puede calcular $p_Y(0)$, $p_Y(1)$ y $p_Y(3)$.

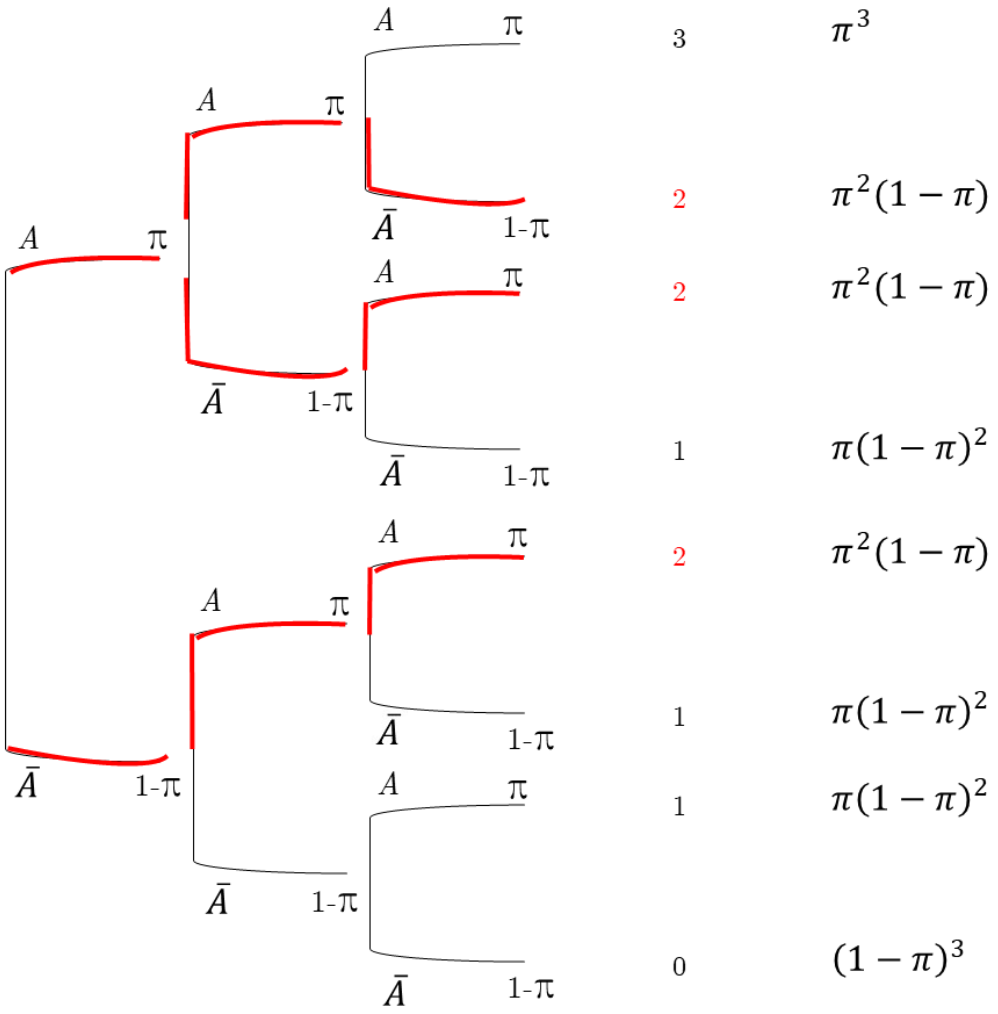


Figura 5.21 *Árbol de probabilidades para una distribución Binomial de parámetros n=3 y π*

Una variable aleatoria discreta Y tiene una distribución Binomial con parámetros matemáticos n y π, y se simboliza $Y \sim Bi(n, \pi)$, si su función de probabilidad puntual es:

$$p_Y(y) = \binom{n}{y} \pi^y (1-\pi)^{(n-y)} \; con \; y = 0, 1, ..., n \;\; ^2$$

donde $n \in \mathbb{N}$ y $\pi \in [0, 1]$.

La función de probabilidad puntual Binomial se muestra en la Figura 5.22. Se puede destacar de la distribución Binomial que el rango es n.

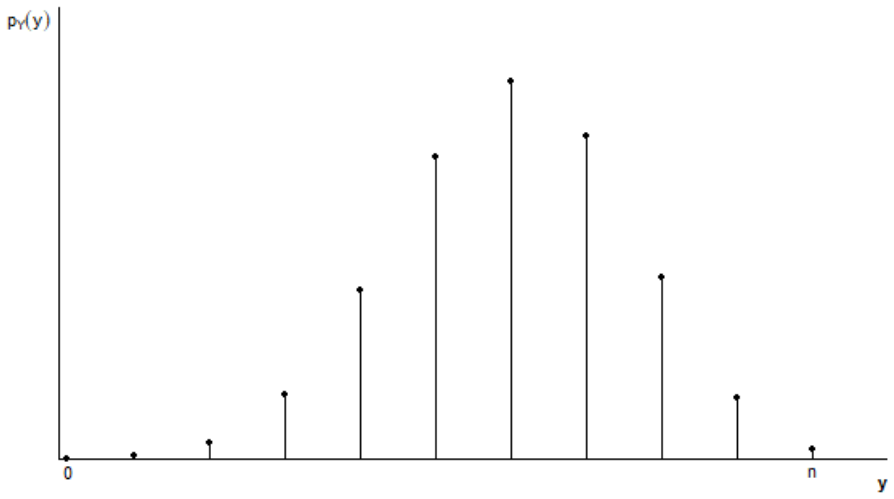


Figura 5.22 *Función de probabilidad puntual Binomial, $p_Y(y)$, de parámetros n y π*

El desplazamiento de la función de probabilidad puntual Binomial con respecto al eje de abscisas y su forma van a depender de los valores de n y π. En la Figura 5.23 se presentan diferentes situaciones.

En la primera columna se presentan gráficas de distribuciones binomiales con $n = 5$ y diferentes valores de π. Observe que:

- cuando π es pequeña, los valores más probables de la variable Y son los más bajos (0 y 1 en este caso) mientras que los valores altos son los menos probables. La distribución es asimétrica a la derecha.
- cuando $\pi = 0,50$, los valores más probables de la variable Y son los intermedios (2 y 3 en este caso). La distribución es simétrica.
- cuando π es alta, los valores más probables de la variable Y son los más altos (4 y 5 en este caso) mientras que los menos probables son los bajos. La distribución es asimétrica a la izquierda.

En la segunda columna se presentan gráficas de distribuciones binomiales con $\pi = 0,20$ y diferentes valores de n. Observe que a medida que n aumenta, la forma de la distribución se vuelve cada vez más simétrica, independientemente del valor de π.

[2] La expresión $\binom{n}{y}$ hace referencia a las "combinaciones de n elementos tomadas de y" es decir, al número de grupos distintos (si difieren de un elemento sin importar el orden) de tamaño y que se pueden formar a partir de un total de n elementos. Se obtiene de la siguiente manera: $\binom{n}{y} = \frac{n!}{y!(n-y)!}$

Si $Y \sim Bi(n, \pi)$, se demuestra que su media es $E(Y) = n\pi$ y su desvío estándar es $D(Y) = \sqrt{n\pi(1-\pi)}$.

Analizando la Figura 5.23, reflexione sobre cuál es el valor más probable en cada distribución ¿Siempre coincide con $E(Y)$?

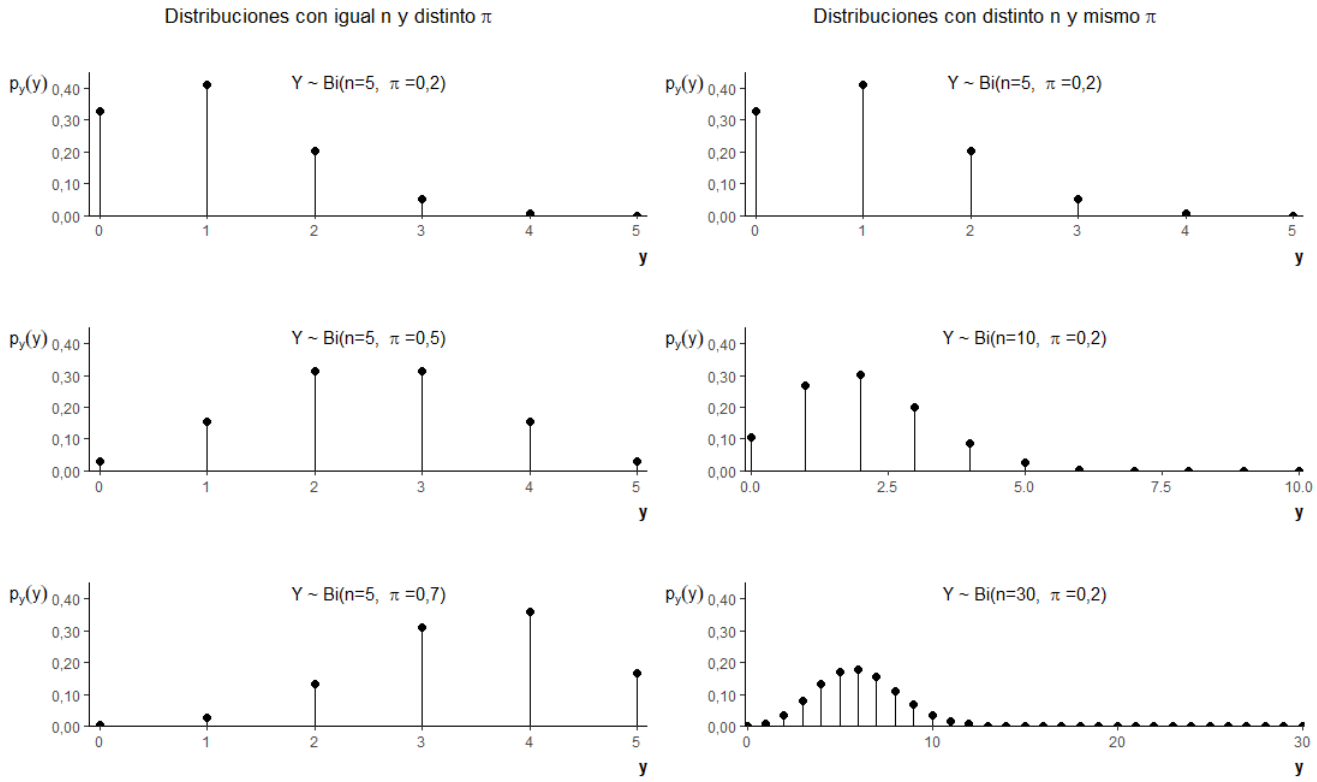


Figura 5.23 *Cambios en la distribución Binomial al variar π o n*

Problema 1 (pág. 7): *Suponga que la empresa metalúrgica que fabrica las barras de acero las comercializa en cajas de 100 y que, en cada caja, interesa observar la cantidad de barras que cumplen las especificaciones (longitud en el intervalo 250 mm $\pm$ 0,6 mm). Para esta producción, se supone que* la proporción de barras que cumplen con las especificaciones es 0,9.
A partir de esta información, interesa determinar:

1. *¿Cuál es la cantidad media de barras de una caja que cumplen con las especificaciones? ¿y el desvío estándar?*
2. *¿Qué proporción de cajas…*
   a) *tienen todas las barras que cumplen con las especificaciones?*
   b) *tienen menos de 95 barras que cumplen con las especificaciones?*
   c) *tienen al menos 90 barras que cumplen con las especificaciones?*
   d) *tienen entre 95 y 100 barras que cumplen con las especificaciones, ambos valores incluidos?*
3. *¿Qué valor de la cantidad de barras que cumplen con las especificaciones es superado por el 20% de las cajas?*

Resolución:

Se define la variable aleatoria Y: cantidad de barras que cumplen con las especificaciones en una caja de 100. La unidad asociada a esta variable es cada caja. Para el armado de cajas, se supone que las barras se seleccionan de forma independiente y que la chance de seleccionar una barra al azar de la producción y que esta cumpla con las especificaciones es 0,9. Al cumplirse las hipótesis de una distribución Binomial, en símbolos, se puede expresar $Y \sim Bi(100; 0,90)$ *por lo que su función de probabilidad puntual resulta:*

$$p_Y(y) = \binom{100}{y} 0,9^y (1-0,9)^{(100-y)} \text{ con } y = 0,1,...,100$$

1. *Aplicando las fórmulas para la media y el desvío estándar:*

   $E(Y) = 100 \cdot 0,9 = 90$ *y* $D(Y) = \sqrt{100 \cdot 0,9 \cdot (1-0,9)} = 3.$

   La cantidad media de barras que cumplen con las especificaciones es 90 por caja y el desvío estándar es igual a 3 barras.

2. a) $P(Y = 100) = \binom{100}{100} 0,9^{100} (1-0,9)^{100-100} \approx 0$ *(Figura 5.24.a). ¿Cómo se interpreta el valor 0?*

   - *Si se considera una gran cantidad de cajas armadas por la empresa, prácticamente ninguna de ellas va a tener todas las barras dentro de especificaciones.*
   - *Si se selecciona una caja al azar, la chance (o probabilidad) de que todas las barras cumplan con las especificaciones es prácticamente nula.*

   Análogamente se obtienen las restantes probabilidades.

   b) $P(Y < 95) = P(Y \leq 94) = \sum_{y=0}^{94} p_Y(y) = 0,942$ *(Figura 5.24.b).*

   c) $P(Y \geq 90) = 1 - P(Y < 90) = 1 - P(Y \leq 89) = 1 - \sum_{y=0}^{89} p_Y(y) = 1 - 0,417 = 0,583$ *(Figura 5.24.c).*

   d) $P(95 \leq Y \leq 100) = P(Y \leq 100) - P(Y < 95) = P(Y \leq 100) - P(Y \leq 94) = 1 - 0,942 = 0,058$ *(Figura 5.24.d).*

   En la Figura 5.24 se representa la distribución $Bi(100; 0,90)$ *solo para valores de la variable mayores o iguales que 75, ya que los valores menores tienen probabilidad nula o prácticamente nula.*

3. *Se desea conocer la cantidad de barras que cumplen con las especificaciones superada por el 20% de las cajas. En este caso, a partir de una probabilidad dada, se busca un valor de la variable aleatoria Y,* y^**, que la satisfaga.*

   $P(Y > y^*) = 0,20$*; entonces,* $P(Y \leq y*) = 0,80.$

   $P(Y \leq y*) = \sum_{y=0}^{y*} p_Y(y) = 0,80$

   Observe que, $P(Y \leq 92) = 0,793$ *y* $P(Y \leq 93) = 0,882$*. Entonces,* $P(Y > 92) = 0,207$ *y* $P(Y > 93) = 0,118.$

   Es decir, para ningún valor de Y se cumple exactamente con lo que se pretende, lo cual es frecuente para el caso de variables discretas.

   Estas probabilidades se interpretan como:

   - *aproximadamente en el 20,7% de las cajas, la cantidad de barras que cumplen con las especificaciones es mayor a 92.*
   - *aproximadamente en el 11,8% de las cajas, la cantidad de barras que cumplen con las especificaciones es mayor a 93.*

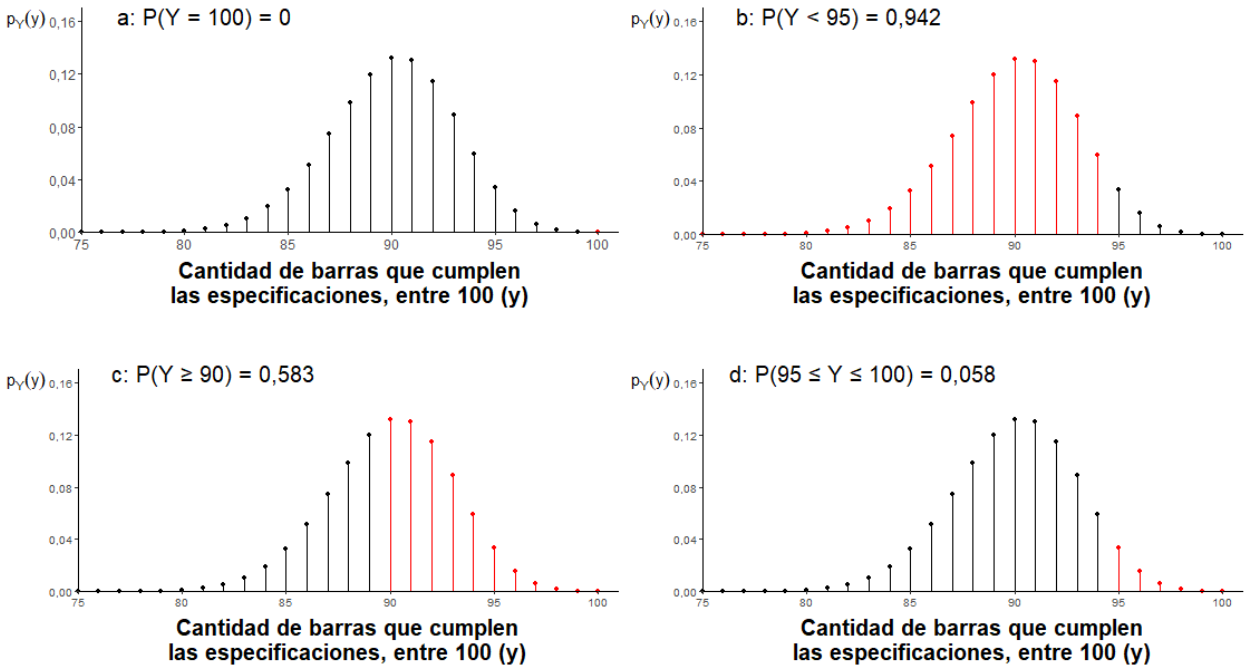


Figura 5.24 *Probabilidades correspondientes a los ítems 2.a, 2.b, 2.c y 2.d, suponiendo* $Y \sim Bi(100; 0,90)$

En la Sección 5.6.3 se muestra cómo obtener las probabilidades y los valores de la variable correspondientes a los ítems 2 y 3, con R.



Con lo visto hasta ahora, se puede resolver las Actividades 22 a 25, Sección 5.5.

5.3.3 Distribución Hipergeométrica

Considere un población finita de N unidades, cada una de las cuales se clasifica según presente o no presente cierta categoría de interés. Se sabe además que, en ella, N_E unidades presentan la categoría de interés.

Se seleccionan al azar y sin reposición, n unidades de dicha población. Interesa, al igual que en la distribución Binomial, registrar la variable aleatoria Y: cantidad de unidades que presentan la categoría de interés en una muestra de tamaño n. El recorrido de Y es $R_Y = \{\text{máx}(0,\ n + N_E - N),\ \ldots,\ \text{mín}(n,\ N_E)\}$.

En este modelo, a partir de la población original de N unidades surge una nueva población, también finita, formada por todos los conjuntos de n elementos que se pueden extraer de un total de N elementos. Por la naturaleza de la experiencia se puede definir una expresión para obtener la frecuencia relativa poblacional o probabilidad asociada a cada uno de los valores posibles de la variable Y, que en este caso se comporta según la distribución Hipergeométrica.

Una variable aleatoria discreta Y tiene una distribución Hipergeométrica con parámetros matemáticos N, N_E y n, y se simboliza $Y \sim Hip(N, N_E, n)$, si su función de probabilidad puntual es:

$$p_Y(y) = \frac{\binom{N_E}{y}\binom{N-N_E}{n-y}}{\binom{N}{n}} \; con \; y = \text{máx}(0,\, n+N_E-N),\, \ldots,\, \text{mín}(n,\, N_E)$$

donde N, N_E y $n \in \mathbb{N}$.

La función de probabilidad puntual Hipergeométrica se grafica en la Figura 5.25.

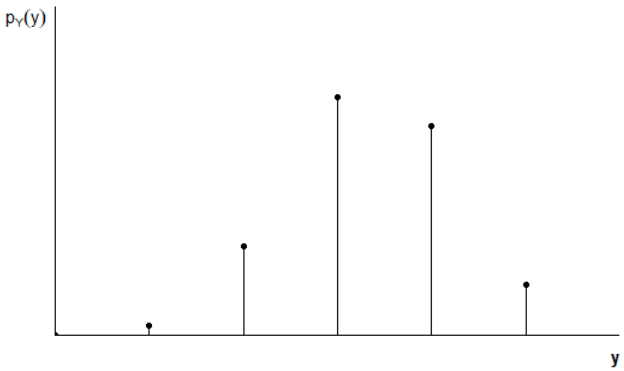


Figura 5.25 *Función de probabilidad puntual Hipergeométrica, $p_Y(y)$, de parámetros N, N_E y n*

El desplazamiento con respecto al eje de abscisas y la forma de la distribución dependen de los valores de N, N_E y n. En la Figura 5.26 se ilustran distintas situaciones.

Las filas representan las posibles relaciones entre N, N_E y n: $N_E > n$, $N_E < n$, $N_E = n$, respectivamente. Por su parte, las columnas muestran distintos valores de N (100, 1000 y 5000, de izquierda a derecha).

Si $Y \sim Hip(N, N_E, n)$, se demuestra que su media es $E(Y) = n\dfrac{N_E}{N}$ y su desvío estándar es

$D(Y) = \sqrt{n\dfrac{N_E}{N}\,(1-\dfrac{N_E}{N})\,(\dfrac{N-n}{N-1})}$.

Analizando la Figura 5.26, reflexione sobre cuál es el valor más probable en cada distribución ¿Siempre coincide con $E(Y)$?

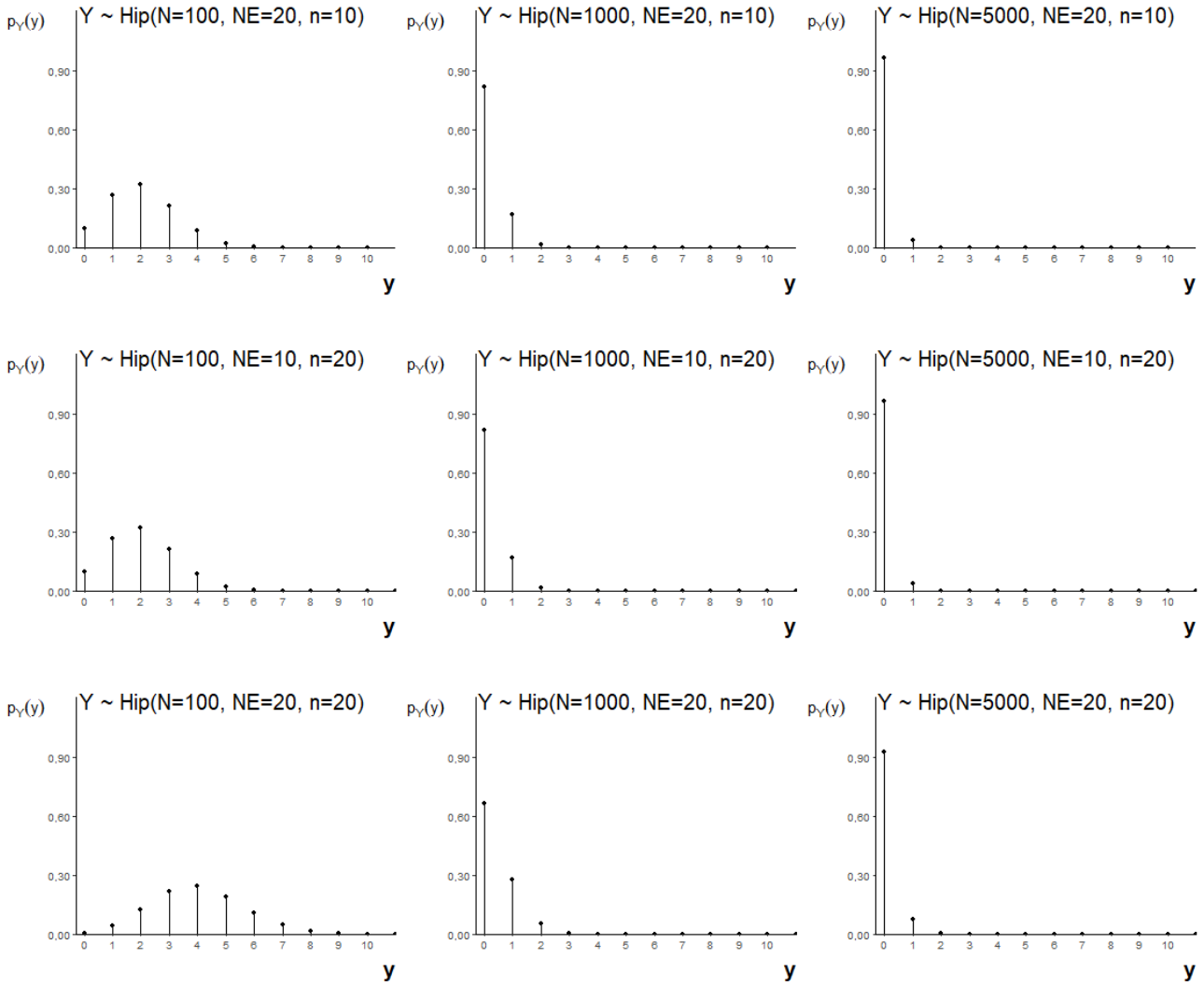


Figura 5.26 *Cambios en la distribución Hipergeométrica según la relación de N_E y n, variando N*



Problema 1 (pág. 7): *Suponga que la empresa metalúrgica que fabrica las barras de acero, las comercializa en cajas de 100 unidades. En cada caja, hay una cierta cantidad de barras que cumplen las especificaciones (longitud en el intervalo 250 mm $\pm$ 0,6 mm).*
Un comprador acuerda con la empresa el siguiente plan de aceptación por muestreo: de cada caja selecciona aleatoriamente 10 barras (sin reposición). Si al menos 9 de las barras seleccionadas cumplen las especificaciones, acepta la caja; en caso contrario, la devuelve.
Suponga que una caja contiene exactamente 95 barras que cumplen las especificaciones y 5 que no las cumplen.
A partir de esta información, interesa determinar:

1. *¿Cuál es la cantidad media de barras que cumplen con las especificaciones en una muestra de tamaño $n = 10$? ¿y el desvío estándar?*
2. *¿Qué proporción de muestras…*
   a) *tienen todas las barras que cumplen con las especificaciones?*
   b) *tienen menos de 9 barras que cumplen con las especificaciones?*

c) *tienen al menos 9 barras que cumplen con las especificaciones?*

d) *tienen entre 8 y 10 barras que cumplen con las especificaciones, ambos valores incluidos?*

3. *¿Cuál es la probabilidad de que el comprador acepte una caja?*
4. *¿Cuál sería la probabilidad de que el comprador acepte una caja si se supone que en las cajas hay 98 barras que cumplen con las especificaciones y 2 que no?*
5. *¿Qué valor de la cantidad de barras que cumple con las especificaciones es superado por el 20% de las muestras de una caja con 95 barras que cumplen con las especificaciones?*

Resolución:

Se define la variable aleatoria Y: cantidad de barras que cumplen con las especificaciones en una muestra de 10 barras (tomada de una caja que se supone que tiene 95 que cumplen las especificaciones y 5 que no). La unidad asociada a esta variable es cada muestra. Como en cada caja hay una población finita de 100 unidades, de las cuales 10 se extraen sin reposición, se puede pensar que Y tiene una distribución Hipergeométrica. En símbolos, se puede expresar $Y \sim Hip(100, 95, 10)$ *por lo que su función de probabilidad puntual resulta:*

$$p_Y(y) = \frac{\binom{95}{y}\binom{100-95}{10-y}}{\binom{100}{10}} \text{ con } y = 5, \ldots, 10$$

1. *Aplicando las fórmulas para la media y el desvío estándar:*

   $E(Y) = 10\,\frac{95}{100} = 9{,}5$ *y* $D(Y) = \sqrt{10\,\frac{95}{100}\,(1-\frac{95}{100})\,(\frac{100-10}{100-1})} = 0{,}657.$

   La cantidad media de barras que cumple con las especificaciones en una muestra es 9,5 barras y el desvío estándar es igual a 0,657 barras.

2. a) $P(Y = 10) = \frac{\binom{95}{10}\binom{100-95}{10-10}}{\binom{100}{10}} = 0{,}584$ *(Figura 5.27.a).*

   ¿Cómo se interpreta el valor 0,584?

   - *Si se consideran todas las muestras de tamaño 10 que se pueden sacar de esa caja, aproximadamente el 58% de las muestras va a tener todas las barras dentro de las especificaciones.*
   - *Si se selecciona una muestra al azar, la chance (o probabilidad) de que todas las barras de dicha muestra cumplan con las especificaciones es aproximadamente 0,58.*

   Análogamente se obtienen las restantes probabilidades:

   b) $P(Y < 9) = P(Y \leq 8) = \sum_{y=0}^{8} p_Y(y) = 0{,}077$ *(Figura 5.27.b).*

   c) $P(Y \geq 9) = 1 - P(Y < 9) = 1 - P(Y \leq 8) = 1 - 0{,}077 = 0{,}923$ *(Figura 5.27.c).*

   d) $P(8 \leq Y \leq 10) = P(Y \leq 10) - P(Y < 8) = P(Y \leq 10) - P(Y \leq 7) = 1 - \sum_{y=0}^{7} p_Y(y) = 1 - 0{,}007 = 0{,}993$ *(Figura 5.27.d).*

3. *La probabilidad de que el comprador acepte una caja se puede calcular como la probabilidad de que el número de barras que cumplen con las especificaciones de la muestra sea al menos 9. Este cálculo se hizo en el ítem 2.c. y la probabilidad resultante es igual a 0,923.*

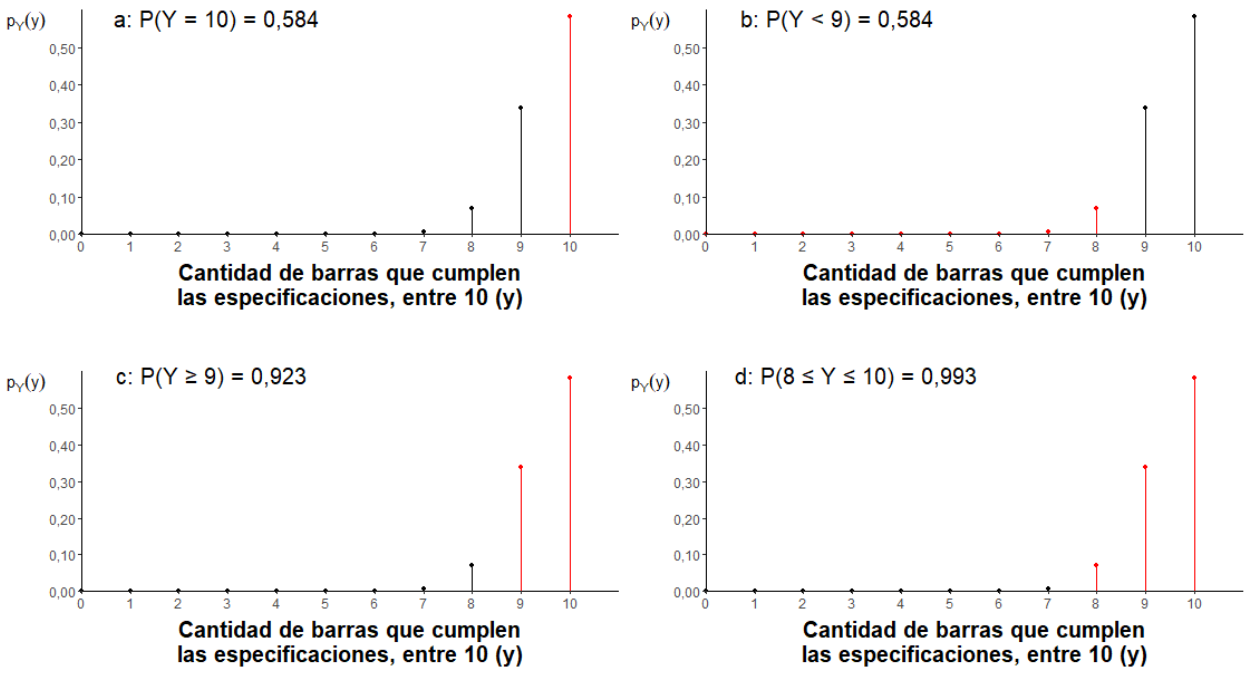


Figura 5.27 *Probabilidades correspondientes a los ítems 2.a, 2.b, 2.c y 2.d, suponiendo* $Y \sim Hip(100, 95, 10)$

4. *Si se supone que en la caja hay 98 barras que cumplen con las especificaciones y 2 que no lo hacen, cambia la distribución de probabilidades de la variable de interés. En símbolos, se puede expresar* $Y \sim Hip(100, 98, 10)$*, por lo que su función de probabilidad puntual resulta:*

$$p_Y(y) = \frac{\binom{98}{y}\binom{100-98}{10-y}}{\binom{100}{10}} \text{ con } y = 2, \ldots, 10$$

$P(Y \geq 9) = 1 - P(Y < 9) = 1 - P(Y \leq 8) = 1 - 0,009 = 0,991$

Cuando el total de barras dentro de especificaciones de las cajas aumenta a 98, la probabilidad de que el comprador acepte una caja es 0,991.

5. *Se desea conocer la cantidad de barras que cumplen con las especificaciones superada por el 20% de las muestras de una caja en la cual 95 barras cumplen con las especificaciones. En este caso, a partir de una probabilidad dada, se busca un valor de la variable aleatoria* Y*,* y^**, que la satisfaga.*

$P(Y > y^*) = 0,20$*; entonces,* $P(Y \leq y^*) = 0,80$.

$P(Y \leq y^*) = \sum_{y=0}^{y^*} p_Y(y) = 0,80$

Observe que, $P(Y \leq 9) = 0,416$ *y* $P(Y \leq 10) = 1$*. Entonces,* $P(Y > 9) = 0,584$ *y* $P(Y > 10) = 0$*. Es decir, para ningún valor de Y se cumple con lo que se pretende. Estas probabilidades se interpretan como:*

- *aproximadamente en el 42% de las muestras, la cantidad de barras que cumplen con las especificaciones es mayor a 9.*
- *en ninguna muestra, la cantidad de barras que cumplen con las especificaciones es mayor a 10.*



Con lo visto hasta ahora, se puede resolver las Actividades 26 a 29, Sección 5.5.

5.3.4 Distribución Geométrica

Considere que se van seleccionando unidades de una población y para cada una se registra si presenta cierta categoría de interés o no. La secuencia de unidades se interrumpe cuando aparece una unidad que presenta la categoría de interés. Suponga que las unidades se seleccionan de tal manera que los resultados obtenidos en las mismas resultan independientes entre sí y que la probabilidad (π) de que una unidad presente la categoría de interés es igual para todas las unidades de la población (es decir, se mantiene constante a medida que se seleccionan las unidades que conforman la secuencia).

Dado que se seleccionan unidades hasta que aparezca la primera que presenta la categoría de interés, se define una variable aleatoria que informa el tamaño de esta secuencia: Y es la cantidad de unidades que deben seleccionarse hasta que aparezca una unidad con la categoría de interés. El recorrido de Y es $R_Y = \{1,\ 2,\ \ldots\}$, es decir, esta variable asume valores enteros mayores o iguales que 1.

Una variable aleatoria discreta Y tiene una distribución Geométrica con parámetro matemático π, y se simboliza $Y \sim Geom(\pi)$, si su función de probabilidad puntual es:

$$p_Y(y) = \pi(1-\pi)^{y-1} \; con \; y = 1,\ 2,\ \ldots$$

donde $\pi \in [0,1]$.

La función de probabilidad puntual Geométrica se presenta en la Figura 5.28.

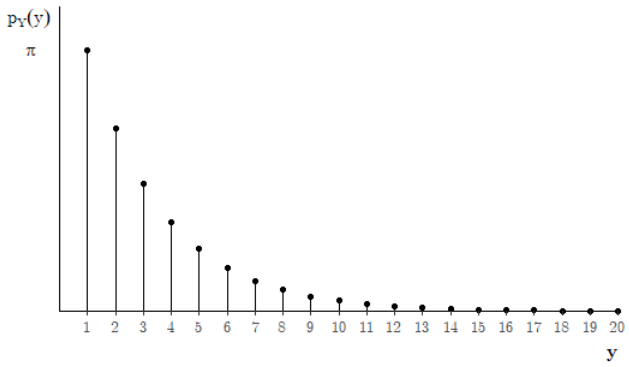


Figura 5.28 *Función de probabilidad puntual Geométrica, $p_Y(y)$, de parámetro π*

Observando la Figura 5.28 se pueden destacar algunas particularidades de la distribución Geométrica:

- es asimétrica por derecha;
- la media resulta mayor que la mediana;
- toma su valor máximo en $y = 1$, igual a π, por lo que la moda es 1.

Su asimetría depende del valor de π. En la Figura 5.29 se presentan tres situaciones (π igual a 0,2; 0,5 y 0,7). Observe que, a medida que π aumenta, la asimetría es más pronunciada.

Si $Y \sim Geom(\pi)$, se demuestra que su media es $E(Y) = \dfrac{1}{\pi}$ y su desvío estándar es $D(Y) = \sqrt{\dfrac{1-\pi}{\pi^2}}$.

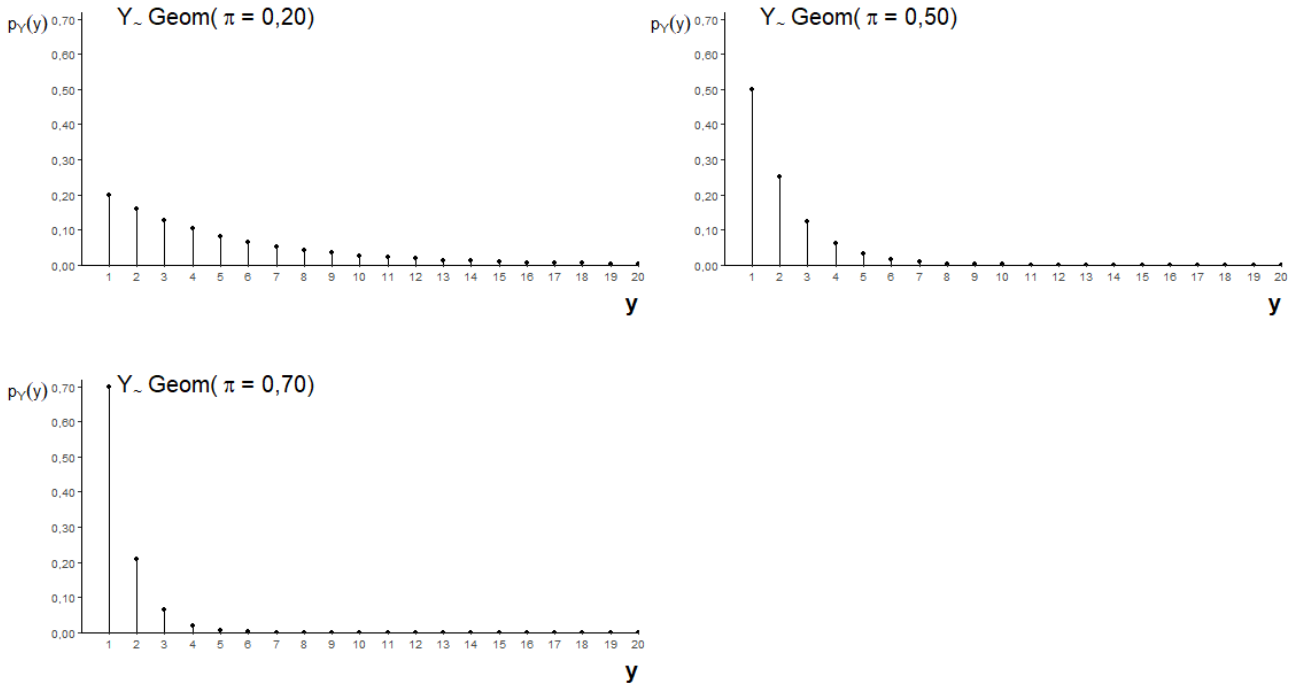


Figura 5.29 *Cambios en la distribución Geométrica al variar* π

Una propiedad importante, también de la distribución Geométrica, es **la propiedad de la falta de memoria**.

Se considera que la ocurrencia de las unidades que presentan la categoría de interés es por causa del azar y no sigue un patrón. Si el número de unidades que se seleccionan hasta que aparece la primera que presenta la categoría de interés es mayor que s (con lo cual $Y > s$), entonces la probabilidad de que se seleccionen t unidades adicionales hasta que aparezca la que presenta dicha categoría (y, por tanto, $Y > t+s$), es igual a la probabilidad de que se seleccione más de t unidades. Por lo tanto, seleccionadas s unidades que no presentan la categoría, se puede comenzar a contar de nuevo, olvidando lo ocurrido (falta de memoria) hasta ese momento.

Formalizando, para cualquier $s > 0$ y $t > 0$, $P(Y > s+t \mid Y > s) = \dfrac{P(Y > s+t)}{P(Y > s)} = \dfrac{1 - \sum_{y=1}^{s+t} \pi(1-\pi)^{y-1}}{1 - \sum_{y=1}^{s} \pi(1-\pi)^{y-1}} =$

$1 - \sum_{y=1}^{t} \pi(1-\pi)^{y-1}$.

Por tanto, $P(Y > s+t \mid Y > s) = P(Y > t)$.

Problema 1 (pág. 7): *Se retoma la situación del Problema 1 donde se considera que una barra cumple con las especificaciones si su longitud se encuentra entre 249,4 mm y 250,6 mm. Para esta producción, se supone que la proporción de barras que cumplen con las especificaciones es 0,9. En el laboratorio se revisan las barras hasta encontrar las que no cumplen con las especificaciones para hacerles a estas últimas un estudio particular. Por ese motivo, se estudian las secuencias de barras que hay que seleccionar hasta que se elige una barra que no cumple con las especificaciones. A partir de esta información, interesa determinar:*

1. *¿Cuál es la cantidad media de barras que hay que seleccionar hasta que se elija una que no cumpla con las especificaciones? ¿y el desvío estándar?*
2. *¿Qué proporción de secuencias ...*
   a) *están formadas por 10 barras?*
   b) *están formadas por menos de 7 barras ?*
   c) *están formadas por al menos 9 barras?*
   d) *están formadas entre 7 y 10 barras, ambos valores incluidos?*
3. *¿Qué valor de la cantidad de barras seleccionadas hasta que se elige una barra que no cumple con las especificaciones es superado por el 20% de las secuencias?*

Resolución:

Se define la variable aleatoria Y: cantidad de barras que se seleccionan hasta que aparece la primera que no cumple con las especificaciones. La unidad asociada a esta variable es cada secuencia de selección. En símbolos, se puede expresar $Y \sim Geom(0,10)$ *por lo que su función de probabilidad resulta*

$$p_Y(y) = 0,1\,(1-0,1)^{y-1} \text{ con } y = 1,2,...$$

.

1. *Aplicando las fórmulas para la media y el desvío estándar:*

   $E(Y) = \frac{1}{0,1} = 10$ *y* $D(Y) = \sqrt{\frac{1-0,1}{0,1^2}} = 9,49$.

   La cantidad media de barras en una secuencia es 10 y el desvío estándar es igual a 9,49.

2. a) $P(Y = 10) = 0,1\,(1-0,1)^{10-1} = 0,0387$ *(Figura 5.30.a).*

   ¿Cómo se interpreta el valor 0,0387?

   - *Si se considera una gran cantidad de secuencias, aproximadamente el* $3,9\%$ *de las mismas va a estar formada por 10 barras.*
   - *Si se selecciona una secuencia al azar, la chance (o probabilidad) de que dicha secuencia esté formada por 10 barras es 0,039.*

   Análogamente se obtienen las restantes probabilidades:

   b) $P(Y < 7) = P(Y \leq 6) = \sum_{y=0}^{6} 0,1\,0,9^{y-1} = 0,4686$ *(Figura 5.30.b).*

   c) $P(Y \geq 9) = 1 - P(Y \leq 8) = 1 - 0,5695 = 0,4305$ *(Figura 5.30.c).*

   d) $P(7 \leq Y \leq 10) = P(Y \leq 10) - P(Y < 7) = P(Y \leq 10) - P(Y \leq 6) = 0,6513 - 0,4686 = 0,1827$ *(Figura 5.30.d).*

3. *Se desea conocer la cantidad de barras seleccionadas hasta que aparece la primera que no cumple con las especificaciones, superada por el 20% de las secuencias. En este caso, a partir de una probabilidad dada, se busca un valor de la variable aleatoria Y, denotado* y^**, que la satisfaga.*

   $P(Y > y^*) = 0,20$*; entonces,* $P(Y \leq y*) = 0,80$.
   $P(Y \leq y*) = \sum_{y=0}^{y*} p_Y(y) = 0,80$

   Observe que, $P(Y \leq 15) = 0,794$ *y* $P(Y \leq 16) = 0,815$*. Entonces,* $P(Y > 15) = 0,206$ *y* $P(Y > 16) = 0,185$.

   Es decir, para ningún valor de Y se cumple exactamente con lo que se pretende. Estas probabilidades se interpretan como:

- *aproximadamente en el 20,6 % de las secuencias, la cantidad de barras seleccionadas hasta que aparece la primera que no cumple con las especificaciones es mayor a 15.*
- *aproximadamente en el 18,5 % de las secuencias, la cantidad de barras seleccionadas hasta que aparece la primera que no cumple con las especificaciones es mayor a 16.*

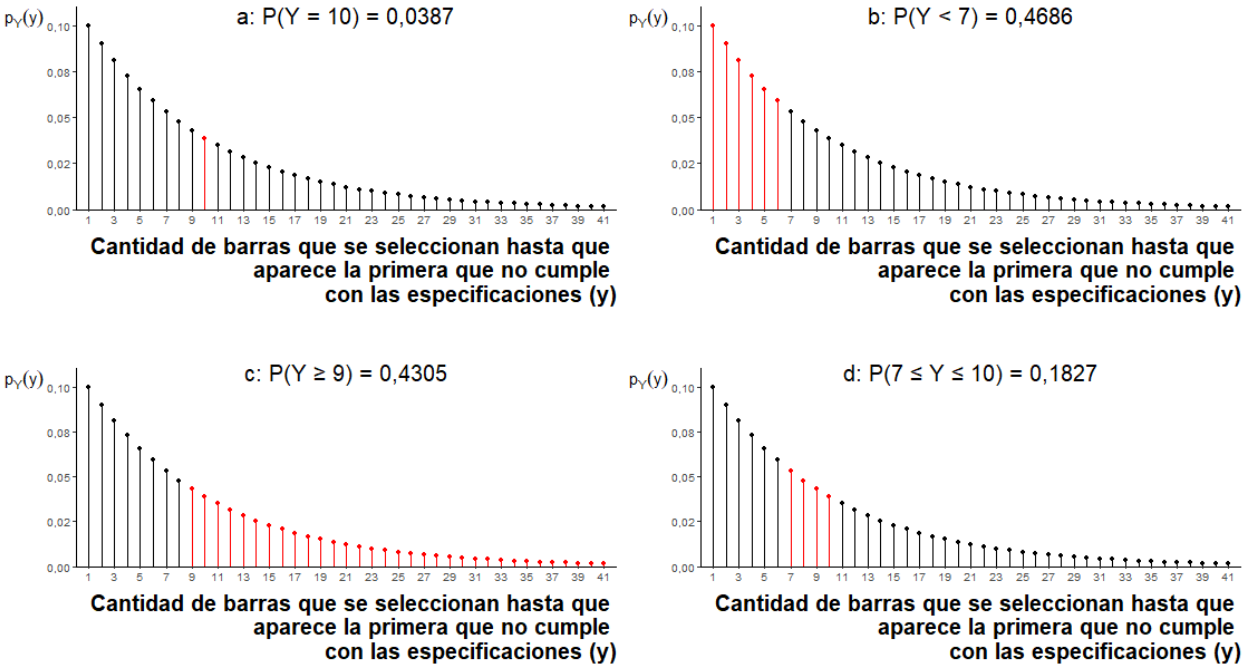


Figura 5.30 *Probabilidades correspondientes a los ítems 2.a, 2.b, 2.c y 2.d, suponiendo* $Y \sim Geom(0, 10)$

Con lo visto hasta ahora, se puede resolver las Actividades 30 a 33, Sección 5.5.

5.3.5 Distribución Poisson

La distribución Poisson se emplea para describir el comportamiento de eventos discretos que ocurren a lo largo de intervalos continuos de tiempo, espacio, longitud u otras magnitudes. En particular, permite modelar la cantidad de eventos que se observan en intervalos con una amplitud dada.

Por ejemplo, puede resultar de interés estudiar cuántos accidentes ocurren en una esquina determinada durante un día, o cuántos poros se observan al pintar un metro cuadrado de chapa. En el primer caso, el evento es el accidente y el intervalo corresponde a un período de tiempo de amplitud un día. En el segundo, el evento es el poro y el intervalo es una región de un metro cuadrado de superficie.

Una variable aleatoria discreta Y tiene una distribución Poisson con parámetro matemático α (con $\alpha > 0$), y se simboliza $Y \sim Po(\alpha)$, si su función de probabilidad puntual es:

$$p_Y(y) = \frac{e^{-\alpha}\alpha^y}{y!} \; con \; y = 0,\ 1,\ 2,\ \ldots$$

La función de probabilidad puntual Poisson se muestra en la Figura 5.31.

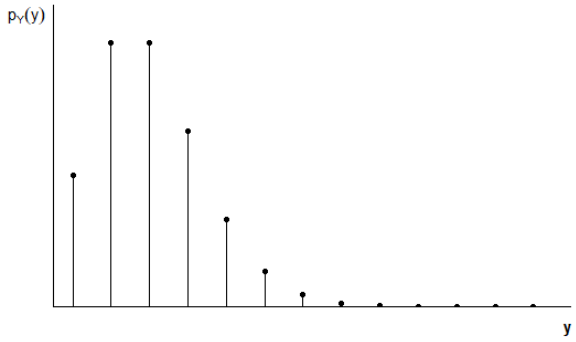


Figura 5.31 *Función de probabilidad puntual Poisson, $p_Y(y)$, de parámetro α*

Observando la Figura 5.31 se pueden destacar algunas particularidades de la distribución Poisson:

- es asimétrica por derecha, lo que ocurre para valores de α pequeños;
- en ese caso, la media resulta mayor que la mediana;
- puede ser que presente su valor máximo en dos valores diferentes de la variable Y.

Si el parámetro matemático α aumenta, la distribución tiende a ser simétrica. En la Figura 5.32 se presentan distribuciones Poisson con distintos valores de α.

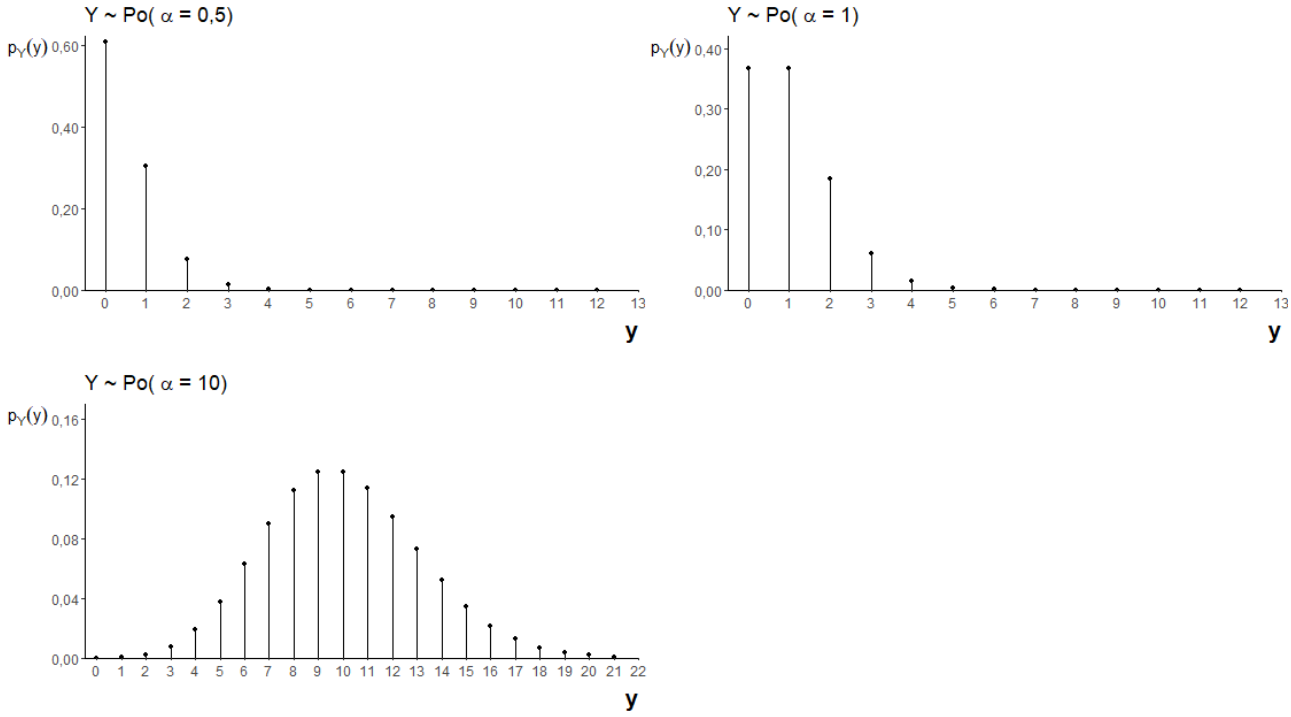


Figura 5.32 *Cambios en la distribución Poisson al variar α*

Si $Y \sim Po(\alpha)$, se demuestra que su media es $E(Y) = \alpha$ y su desvío estándar es $D(Y) = \sqrt{\alpha}$.

Note que en esta distribución, la variancia tiene el mismo valor que la media: ambos valen α. Este parámetro matemático representa el número promedio de eventos por unidad de tiempo, longitud, etc.

1. Proceso Poisson

La distribución Poisson permite modelar la cantidad de eventos observados en intervalos de una amplitud dada. Para extender este modelo a distintos intervalos de tiempo o espacio, se introduce el concepto de proceso Poisson.

Un proceso Poisson es una familia de variables aleatorias discretas $\{Y_t\}$, donde cada variable Y_t representa la cantidad de eventos ocurridos en un intervalo de amplitud t. De esta manera, al variar t, se obtiene una variable aleatoria distinta asociada al intervalo considerado.

Una familia de variables aleatorias discretas $\{Y_t : t \geq 0\}$ sigue un proceso Poisson con parámetro αt, y se simboliza $Y_t \sim Po(\alpha t)$, si, para cada t, la función de probabilidad de Y_t es:

$$p_{Y_t}(y_t) = \frac{e^{-\alpha t}(\alpha t)^{y_t}}{y_t!} \qquad \text{con } y_t = 0, 1, 2, \ldots$$

donde $\alpha > 0$.

Si $Y_t \sim Po(\alpha t)$, se demuestra que su media es $E(Y_t) = \alpha t$ y su desvío estándar es $D(Y_t) = \sqrt{\alpha t}$.

Existen ciertas hipótesis que un proceso debe cumplir para ser considerado Poisson:

1. La probabilidad de que no ocurra ningún evento en el instante inicial vale 1 (condición inicial).
2. La probabilidad de que ocurran dos o más eventos en un intervalo lo suficientemente pequeño es despreciable.
3. Si el intervalo de ocurrencia es lo suficientemente pequeño, la probabilidad de que ocurra exactamente un evento durante ese intervalo es directamente proporcional a la longitud del intervalo.
4. Las cantidades de eventos que ocurren en intervalos no superpuestos son variables aleatorias independientes.
5. Si Y_t es el número de eventos que ocurrren en el intervalo $[0, t)$ y X_t es el número de eventos que ocurren durante $[t_1, t_1 + t)$ para cualquier $t_1 > 0$, las variables aleatorias Y_t y X_t tienen la misma distribución de probabilidades.



Problema 8 (pág. 8): *Una empresa se especializa en realizar ensayos de vida en dispositivos electrónicos. Para un dispositivo en particular llevó adelante un estudio por más de 5000 horas y registró la ocurrencia de fallas así como el tiempo transcurrido entre cada una.*
Un usuario está interesado en utilizar dispositivos de este tipo para construir circuitos especiales y por este motivo quiere saber el número medio de fallas por semana. Es por ello que realiza una consulta en la empresa especializada en ensayos de vida.
Suponga que la variable cantidad semanal de fallas (1 semana = 7 días = 168 h), Y_{168}***, se comporta según la distribución Poisson con*** $\alpha \cdot t = 0,002 \cdot 168 = 0,336$ ***fallas cada 168 h.***
Interesa determinar:

1. *¿Cuál es la cantidad media de fallas por semana? ¿y el desvío estándar?*
2. *¿Qué proporción de semanas ...*

a) *no se produce ninguna falla?*
b) *se producen menos de 2 fallas?*
c) *se producen al menos 2 fallas?*
d) *entre 1 y 3 fallas, ambas incluidas?*

3. *¿Qué valor de la cantidad semanal de fallas es superado por el 20% de las semanas?*

Resolución:

Se define la variable aleatoria Y_{168}: cantidad semanal de fallas (una semana = 168 h). La unidad asociada a esta variable es una semana. En símbolos, se puede expresar $Y_{168} \sim Po(0,336)$ por lo que su función de probabilidad puntual resulta:

$$p_{Y\,168}(y_{168}) = \frac{e^{-0,336}\,0,336^{y_{168}}}{y_{168}!} \text{ con } y_{168} = 0, 1, \ldots$$

1. *Aplicando las fórmulas para la media y el desvío estándar:*

   $E(Y_{168}) = 0,336$ *y* $D(Y_{168}) = \sqrt{0},336 = 0,580$.

   La cantidad media de fallas por semana es 0,336 y el desvío estándar es igual a 0,58 fallas semanales.

2. a) $P(Y_{168} = 0) = \frac{e^{-0,336}\,0,336^0}{0!} = 0,7146$ *(Figura 5.33.a).*
   ¿Cómo se interpreta el valor 0,7146?
   - *Si el dispositivo se ensaya una gran cantidad de semanas, aproximadamente en el 71% de las mismas no va a tener ninguna falla.*
   - *Si se selecciona una semana al azar, la chance (o probabilidad) de que el dispositivo no presente fallas es aproximadamente 0,71.*

   Análogamente se obtienen las restantes probabilidades:

   b) $P(Y_{168} < 2) = P(Y_{168} \leq 1) = P(Y_{168} = 0) + P(Y_{168} = 1) = 0,7146 + 0,2401 = 0,9547$ *(Figura 5.33.b).*

   c) $P(Y_{168} \geq 2) = 1 - P(Y_{168} < 2) = 1 - P(Y_{168} \leq 1) = 1 - 0,9547 = 0,0453$ *(Figura 5.33.c).*

   d) $P(1 \leq Y_{168} \leq 3) = P(Y_{168} \leq 3) - P(Y_{168} < 1) = P(Y_{168} \leq 3) - P(Y_{168} \leq 0) = 0,9996 - 0,7146 = 0,285$ *(Figura 5.33.d).*

3. *Se desea conocer la cantidad semanal de fallas del dispositivo superadas por el 20% de las semanas. En este caso, a partir de una probabilidad dada, se busca un valor de la variable aleatoria Y_{168}, y^*, que la satisfaga.*

   $P(Y_{168} > y^*) = 0,20$*; entonces,* $P(Y_{168} \leq y^*) = 0,80$.

   $P(Y_{168} \leq y^*) = \sum_{y=0}^{y^*} p_Y(y) = 0,80$.

   Observe que, $P(Y_{168} \leq 0) = 0,7146$ *y* $P(Y_{168} \leq 1) = 0,9547$. *Entonces,* $P(Y_{168} > 0) = 0,2854$ *y* $P(Y_{168} > 1) = 0,0453$.
   Es decir, para ningún valor de Y_{168} se cumple con lo que se pretende. Estas probabilidades se interpretan como:
   - *aproximadamente en el 28,5% de las semanas, la cantidad de fallas del dispositivo es mayor a 0.*
   - *aproximadamente en el 4,5% de las semanas, la cantidad de fallas del dispositivo es mayor a 1.*

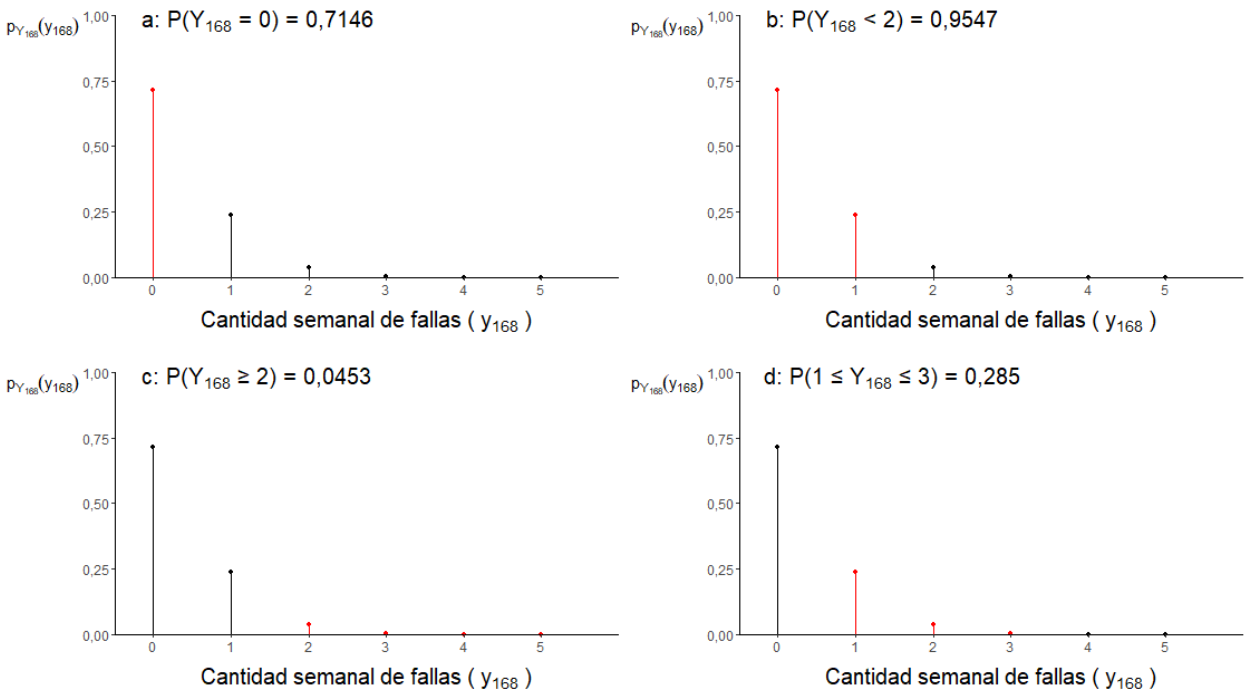


Figura 5.33 *Probabilidades correspondientes a los ítems 2.a, 2.b, 2.c y 2.d, suponiendo* $Y_{168} \sim Po(168 \cdot 0,002 = 0,336)$

2. Relación entre la distribución Poisson y la distribución Exponencial

Las distribuciones Poisson y Exponencial están íntimamente relacionadas. Considere que se define la variable Y_t como el número de ocurrencia de eventos en intervalos de amplitud t cuyo comportamiento se describe mediante un proceso Poisson, $Y_t \sim Po(\alpha t)$. Entonces, W, el tiempo (o también la longitud, la distancia, etc) que transcurre entre dos eventos sucesivos sigue una ley Exponencial, $W \sim Exp(\alpha)$.

Tenga en cuenta que, para ambas variables los intervalos deben estar medidos en la misma unidad de medida (horas, minutos, metros, segundos, etc.) y que el número promedio de eventos por unidad, α es el mismo.

Si en un intervalo de amplitud t no aparecen eventos $(Y_t = 0)$ significa que el tiempo (o longitud, distancia, etc) hasta que aparezca un evento es mayor que t unidades $(W > t)$. Como consecuencia, se puede demostrar que $P(Y_t = 0) = P(W > t)$.

Problema 8 (pág. 8): *Se supone que Y: cantidad de fallas por h, se comporta según la distribución Poisson con* $\alpha = 0,002$. *Interesa determinar:*

1. *¿Cuál es el tiempo medio entre dos fallas consecutivas (en h)? ¿y el desvío estándar?*
2. *¿Qué proporción de pares de fallas consecutivas tienen un tiempo entre las mismas*
   a) *de menos de 100 h?*
   b) *de más de 150 h?*
   c) *de entre 150 h y 160 h?*

Resolución:

Se define la variable aleatoria $Y = Y_1$*: número de fallas por h. La unidad asociada a esta variable es*

un período o intervalo de 1 h. En símbolos, se puede expresar $Y \sim Po(0,002)$ *por lo que su función de probabilidad puntual resulta:*

$$p_Y(y) = \frac{e^{-0,002}\,0,002^y}{y!} \; con \; y = 0,\, 1,\, ...$$

Relacionada a la variable Y se define la variable aleatoria W: tiempo entre dos fallas consecutivas (en h). La unidad asociada a esta variable es un par de fallas consecutivas. En símbolos, se puede expresar $W \sim Exp(0,002)$ *por lo que su función de densidad de probabilidad resulta:*

$$f_W(w) = \begin{cases} 0,002\, e^{-0,002\,w} & si \quad w \geq 0 \\ \\ 0 & si \quad w < 0. \end{cases}$$

1. *Aplicando las fórmulas para la media y el desvío estándar de una distribución Exponencial:*

   $E(W) = \frac{1}{0,002} = 500$ *y* $D(W) = \frac{1}{0,002} = 500$

   El tiempo medio entre dos fallas consecutivas es 500 h y el desvío estándar de los tiempos es 500 h.

2. a) $P(W < 100) = F_W(100) = 1 - e^{-0,002 \cdot 100} = 0,1813$ *(Figura 5.34.a).*

   ¿Cómo se interpreta el valor 0,1813?

   - *Si se observa una gran cantidad de pares de fallas consecutivas, aproximadamente en el 18,13% de los mismos, el tiempo entre ambas resulta menor a 100 h.*
   - *Si se selecciona un par de fallas consecutivas al azar, la chance (o probabilidad) de que el tiempo entre ellas sea menor a 100 h es aproximadamente 0,18.*

   Análogamente se obtienen las restantes probabilidades:

   b) $P(W > 150) = 1 - F_W(150) = 1 - [1 - e^{-0,002 \cdot 150}] = 0,7408$ *(Figura 5.34.b).*

   c) $P(150 < W < 160) = F_W(160) - F_W(150) = [1 - e^{-0,002 \cdot 160}] - [1 - e^{-0,002 \cdot 150}] = 0,2739 - 0,2592 = 0,0147$ *(Figura 5.34.c).*

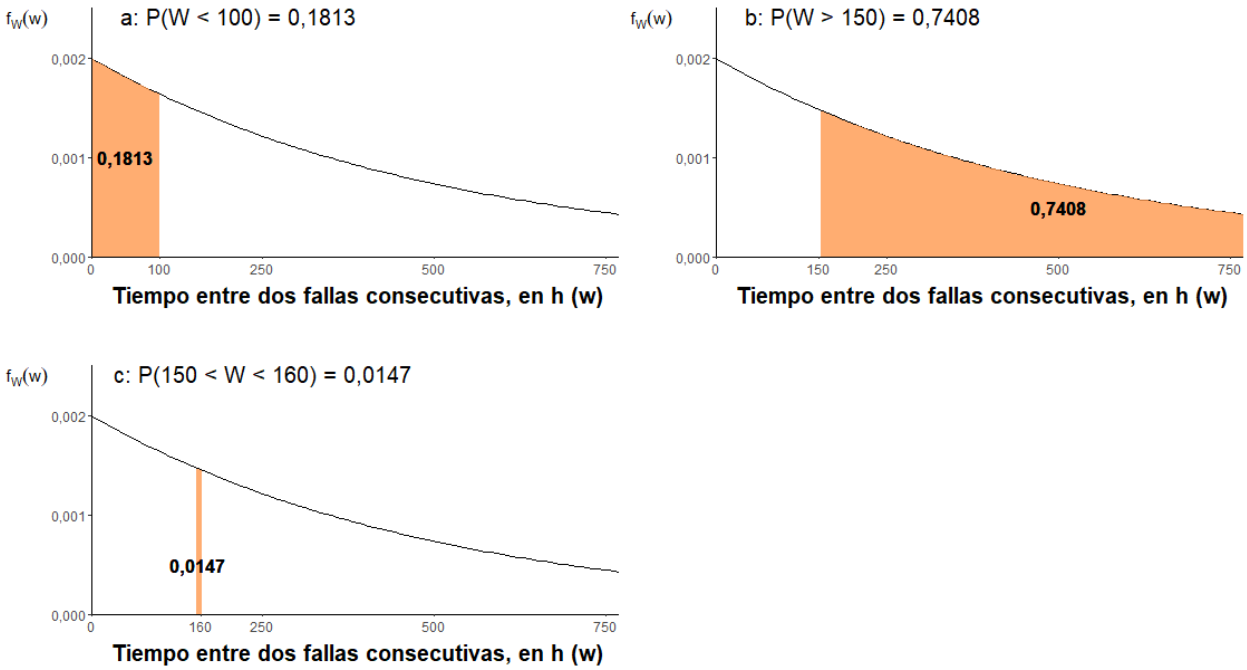


Figura 5.34 *Probabilidades correspondientes a los ítems 2.a, 2.b y 2.c, suponiendo* $W \sim Exp(0,002)$



Con lo visto hasta ahora, se puede resolver las Actividades 34 a 38, Sección 5.5.
Con todas las distribuciones de probabilidades vistas, se pueden resolver las Actividades 39 a 44, Sección 5.5.

5.4 Síntesis

Describir el comportamiento poblacional de una variable aleatoria utilizando alguno de los modelos de distribución de probabilidades definidos en la bibliografía simplifica mucho el estudio de las particularidades de dicho comportamiento, ya que este está especificado y detallado de antemano según los parámetros matemáticos del modelo. El uso de estos modelos particulares de distribución de probabilidades permite describir de manera aproximada el comportamiento de variables aleatorias, facilitando el análisis, la predicción y la toma de decisiones. Entre los modelos más importantes de aplicación en Ingeniería, se pueden mencionar el Normal y el Exponencial, asociados a variables aleatorias continuas. El primero permite describir numerosos fenómenos en el campo de la Ingeniería y juega un papel muy importante en el análisis inferencial; mientras que el segundo sirve para el estudio de la confiabilidad de productos y sistemas. Los modelos Binomial, Hipergeométrico y Poisson están asociados a variables aleatorias discretas. Los dos primeros resultan fundamentales para el análisis de procesos donde se contabilizan eventos en muestras de n unidades; mientras que el último analiza la ocurrencia de eventos en intervalos de tiempo o espacio, siendo especialmente útil en el control de calidad y el análisis de fallas.

Las tablas que se presentan a continuación resumen las distribuciones de probabilidades continuas y discretas abordadas en este capítulo, indicando sus funciones de densidad de probabilidad o de probabilidad puntual, parámetros matemáticos, promedio y desvío estándar.

Tabla 5.1 *Distribuciones para variables aleatorias continuas presentadas en este capítulo*

Modelo y notación	**Función de densidad de probabilidad,** $f_Y(y)$	**Parámetros matemáticos**	$E(Y)$	$D(Y)$
Normal $Y \sim N(\mu,\sigma)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, y \in \mathbb{R}$	$\mu \in \mathbb{R}$ y $\sigma \in \mathbb{R}^+$	μ	σ
Uniforme $Y \sim U(a,b)$	$\begin{cases} \frac{1}{b-a} & si\ a \le y \le b \\ 0\ si\ & y < a\ o\ y > b \end{cases}$	a y $b \in \mathbb{R}$	$\frac{a+b}{2}$	$\sqrt{\frac{(b-a)^2}{12}}$
Triangular $Y \sim Tri(a;b;c)$	$\begin{cases} \frac{2(y-a)}{(b-a)(c-a)} & si \quad a \le y \le c \\ \frac{2(b-y)}{(b-a)(b-c)} & si \quad c < y \le b \\ 0 & si \quad y \notin [a,b] \end{cases}$	a, b y $c \in \mathbb{R}$	$\frac{a+b+c}{3}$	$\sqrt{\frac{a^2+b^2+c^2-ab-ac-bc}{18}}$
Triangular simétrica $Y \sim Tri(a;b)$	$\begin{cases} \frac{2(y-a)}{(b-a)\left(\frac{a+b}{2}-a\right)} & si \quad a \le y \le \frac{a+b}{2} \\ \frac{2(b-y)}{(b-a)\left(b-\frac{a+b}{2}\right)} & si \quad \frac{a+b}{2} < y \le b \\ 0 & si \quad y \notin [a,b] \end{cases}$	a y $b \in \mathbb{R}$	$\frac{a+b}{2}$	$\sqrt{\frac{(b-a)^2}{24}}$
Exponencial $Y \sim Exp(\alpha)$	$\begin{cases} \alpha e^{-\alpha y} & si \quad y \ge 0 \\ 0 & si \quad y < 0 \end{cases}$	$\alpha > 0$	$\frac{1}{\alpha}$	$\frac{1}{\alpha}$

Tabla 5.2 *Distribuciones para variables aleatorias discretas presentadas en este capítulo*

Modelo y notación	**Función de probabilidad puntual,** $p_Y(y)$	**Parámetros matemáticos**	$E(Y)$	$D(Y)$
Bernoulli $Y \sim Be(\pi)$	$\pi^y (1-\pi)^{(1-y)}$ con y= 0 o 1	$\pi \in [0,1]$	π	$\sqrt{\pi(1-\pi)}$
Binomial $Y \sim Bi(n,\pi)$	$\binom{n}{y}\pi^y (1-\pi)^{(n-y)}$ con y= 0, 1 ,...,n	$n \in \mathbb{N}$ y $\pi \in [0,1]$	$n\pi$	$\sqrt{n\pi(1-\pi)}$
Hipergeométrica $Y \sim Hip(N, N_E, n)$	$\frac{\binom{N_E}{y}\binom{N-N_E}{n-y}}{\binom{N}{n}}$ con $y = \text{máx}(0, n+N_E-N)$, $\ldots, \text{mín}(n, N_E)$	N, N_E y $n \in \mathbb{N}$	$n\frac{N_E}{N}$	$\sqrt{n\frac{N_E}{N}(1-\frac{N_E}{N})(\frac{N-n}{N-1})}$
Geométrica $Y \sim Geom(\pi)$	$\pi(1-\pi)^{y-1}$ con y=1, 2, ...	$\pi \in [0,1]$	$\frac{1}{\pi}$	$\sqrt{\frac{1-\pi}{\pi^2}}$
Poisson $Y \sim Po(\alpha)$	$\frac{e^{-\alpha}\alpha^y}{y!}$ con $y = 0,1,2,\ldots$	$\alpha > 0$	α	$\sqrt{\alpha}$
Proceso Poisson $\{Y_t : t \geq 0\}$ $Y_t \sim Po(\alpha t)$	$\frac{e^{-\alpha t}(\alpha t)^{y_t}}{y_t!}$ con $y_t = 0,1,2,\ldots$	$\alpha > 0$	αt	$\sqrt{\alpha t}$

5.5 Actividades propuestas

1. La tensión eléctrica de salida (en voltios) de una fuente de energía eléctrica en diferentes instantes se distribuye Normal, con media 12 V y desvío estándar de $0,10$ V. Por norma establecida, la tensión de salida en cualquier instante debe ser un valor perteneciente al intervalo $12 \pm 0,15$.
   a) Indique la población en estudio, la variable que se mide (clasifíquela) y la población estadística.
   b) Bosqueje la función de densidad de probabilidad de la variable y comente qué información brinda sobre esta población estadística.
   c) Para un instante seleccionado al azar, ¿cuál es la chance de que la tensión de salida de la fuente sea de a lo sumo $11,9$ V?
   d) Calcule la probabilidad de que en un instante cualquiera la fuente de energía tenga una tensión de salida de al menos $12,2$ V.
   e) ¿En qué proporción de instantes, la fuente de energía cumple con la norma establecida para la tensión de salida?
   f) ¿Qué valor de la tensión de salida es superado en el 40 % de los instantes?
   g) Represente gráficamente los valores obtenidos en los ítems *c*) al *f*).

2. En una fábrica producen tapas de corchos para diferentes bodegas. Sobre una de las líneas, aseguran que el diámetro (en milímetros) de las tapas se comporta según un modelo Normal. Se cuenta, además, con la siguiente información, obtenida con `R`:

```
pnorm(21, 20, 0.7)
0.923

pnorm(20.8, 20, 0.7, lower.tail = FALSE)
0.127

qnorm(0.25, 20, 0.7)
19.528

qnorm(0.10, 20, 0.7, lower.tail = FALSE)
20.897
```

   a) Mencione la población en estudio, la variable de interés y la población estadística.
   b) Indique cuánto valen los parámetros matemáticos del modelo Normal en este caso. Interprete ambos valores en términos del problema.
   c) Exprese formalmente las probabilidades incluidas en la salida. Represéntelas gráficamente usando `R`.
   d) ¿Algun/os de los valores de la salida se corresponde con un percentil? Expréselo/s formalmente y represéntelo/s gráficamente.

3. La longitud (en milímetros) de cierto tipo de pieza de acero (Y), es una variable aleatoria con distribución Normal con mediana 10 mm. Se conoce, además, que aproximadamente el 95 % de las piezas tiene longitud en el intervalo (9,6; 10,4).
   a) ¿Cuánto valen la longitud media y el desvío estándar para ese tipo de piezas de acero?

b) ¿Cuánto vale el rango intercuartílico? Interprete ese valor en términos del problema, así como el de los dos cuartiles que intervienen en su cálculo. ¿Qué relación hay entre los valores de ambos cuartiles

c) Responda las siguientes preguntas aplicando la regla empírica. Justifique su respuesta en cada una.

1) ¿Cuánto vale $P(9,8 \leq Y \leq 10,2)$?¿Qué significa en términos del problema?
2) ¿Cuál es la proporción de piezas con longitud mayor que 10,4 mm?
3) ¿Qué proporción de las piezas tiene longitud menor que 9,4 mm?

4. En la producción de vidrios para ventanas, una de las características principales es su espesor (en milímetros). Una empresa de la zona produce diferentes tipos de vidrios y respecto de uno de ellos se puede afirmar que **el comportamiento del espesor (Y) puede describirse con una distribución Normal con $\mu = 6$ mm y $\sigma = 0,01$ mm**.

a) ¿Qué interpretación debe darse a la expresión resaltada?

b) Una empresa constructora (Empresa A) requiere vidrios cuyo espesor sea un valor perteneciente al intervalo $6,007 \pm 0,012$. Otra empresa (Empresa B) requiere vidrios cuyo espesor sea un valor perteneciente al intervalo $6 \pm 0,02$.

Por la gran cantidad de vidrios que necesitan estas empresas, la productora solo puede abastecer a una de ellas y acuerda reponer los vidrios con espesor fuera de las especificaciones definidas. ¿A cuál empresa le aconsejaría Ud. abastecer? Justifique su respuesta.

c) Si se pudieran disponer acciones en el proceso de producción de los vidrios tendientes a modificar los parámetros de la distribución del espesor, ¿qué parámetro/s procuraría Ud. cambiar si el cliente elegido admite como máximo un 1 % de vidrios fuera de especificación? Justifique.

d) Si se habilitara una nueva línea de producción para abastecer al otro cliente, ¿cuánto deberían valer el espesor promedio y la desviación estándar de los vidrios producidos por esa nueva línea? ¿Necesita información adicional para responder? Explique.

5. Considere la situación del Problema 7 (Capítulo 1, Sección 1.4, página 7), en la que se desea determinar la tensión (en voltios) utilizando un multímetro digital. Se dispone de dos equipos: uno que se ha utilizado históricamente y otro, más moderno, adquirido recientemente. Para evaluar su desempeño, se realizan múltiples mediciones sobre una tensión patrón conocida e igual a 12 V.

Suponga que las mediciones repetidas de la tensión patrón, con el multímetro actual, se distribuyen según un modelo Normal con media 12,002 V y desviación estándar 0,005 V; mientras que, con el nuevo multímetro, se distribuyen según un modelo Normal con media 12 V y desviación estándar 0,008 V.

El desempeño de los equipos se evaluará considerando simultáneamente su exactitud, su precisión y la capacidad de generar la mayor proporción posible de mediciones dentro del intervalo $(11,995; 12,005)$ V.

a) Identifique las variables de interés con sus respectivas poblaciones y poblaciones estadísticas.

b) Bosqueje las dos distribuciones de probabilidades en un mismo gráfico y describa comparativamente sus características.

c) Calcule, para cada equipo, la proporción de mediciones de la tensión patrón que se encuen-

tran dentro del intervalo especificado.

d) Teniendo en cuenta simultáneamente la exactitud, la precisión y la proporción de mediciones dentro del intervalo especificado, indique qué multímetro recomendaría. Justifique su elección.

e) Si se desea aumentar la proporción de mediciones de la tensión patrón que se encuentran dentro del intervalo $(11,995;\ 12,005)$ V, indique qué parámetro/s de la distribución de cada equipo sería conveniente modificar y explique por qué. Si considera que la información disponible no es suficiente para responder, indique qué información adicional sería necesaria. Justifique su respuesta.

6. En la Actividad 24, Capítulo 2, se consideró un tipo de válvulas para las cuales su espesor (en pulgadas) debía cumplir con las especificaciones $(2;\ 2,2)$. Suponga que se conoce que el espesor de estas válvulas se comporta según el modelo Normal.

   a) Indique qué valores deberían tener los parámetros matemáticos del modelo mencionado si se pretende que como mínimo el 99 % de las válvulas cumpla con las especificaciones relativas al espesor.

   b) Si el mínimo pretendido fuera de 99,5 %, ¿qué parámetro debería modificarse? ¿cuál debería ser su valor máximo en este caso?

   c) Suponga que se conoce que el espesor promedio de las válvulas es 2,105", ¿cómo procedería para obtener el máximo valor de la desviación estándar si se pretende como mínimo un 99 % de válvulas dentro de especificaciones? (Observación: Este ítem no se puede resolver analíticamente ya que, al no estar centrado el proceso en el punto medio del intervalo de especificaciones, las proporciones de válvulas con espesor menor que el límite inferior y mayor que el límite superior de especificaciones no coinciden. Se puede utilizar Excel o `R`, entre otros, para encontrar el valor máximo pedido computacionalmente).

7. En las vacaciones de invierno, se agregan colectivos de la línea 23 que conectan el centro de la ciudad con la zona de teatros y museos. El tiempo de retraso Y (en minutos) respecto del tiempo establecido para completar el recorrido (2 horas) de los colectivos de esta línea durante las vacaciones, se puede modelar con una distribución Uniforme entre -5 min y 15 min.

   a) Indique la población en estudio, la variable que se mide (clasifíquela) y la población estadística.

   b) Bosqueje la función de densidad de probabilidad de la variable y comente qué información brinda sobre la población estadística.

   c) ¿Qué porcentaje de las vueltas se adelantan (retraso menor de 0 min) en este período de vacaciones?

   d) Calcule la probabilidad de que en una vuelta completa el retraso sea de al menos 8 min.

   e) ¿Qué proporción de vueltas se retrasan entre 10 min y 20 min?

   f) Defina un intervalo de valores (y_1, y_2) tal que el valor de la probabilidad de que una vuelta completa presente retraso dentro de dicho intervalo coincida con el obtenido en el ítem *(e)*. Justifique su respuesta.

   g) ¿Qué valor del tiempo de retraso es superado por el 10 % de las vueltas completas?

8. Considere la situación del Problema 6 (Sección 1.4, página 7) y suponga que el espesor (en centímetros) de los puntos de cierto tramo de ruta (Y), se comporta según un modelo Uniforme. A continuación se presenta la siguiente información, obtenida con `R`:

```
punif(21,20,24)
0.25

punif(23.4, 20, 24, lower.tail = FALSE)
0.15

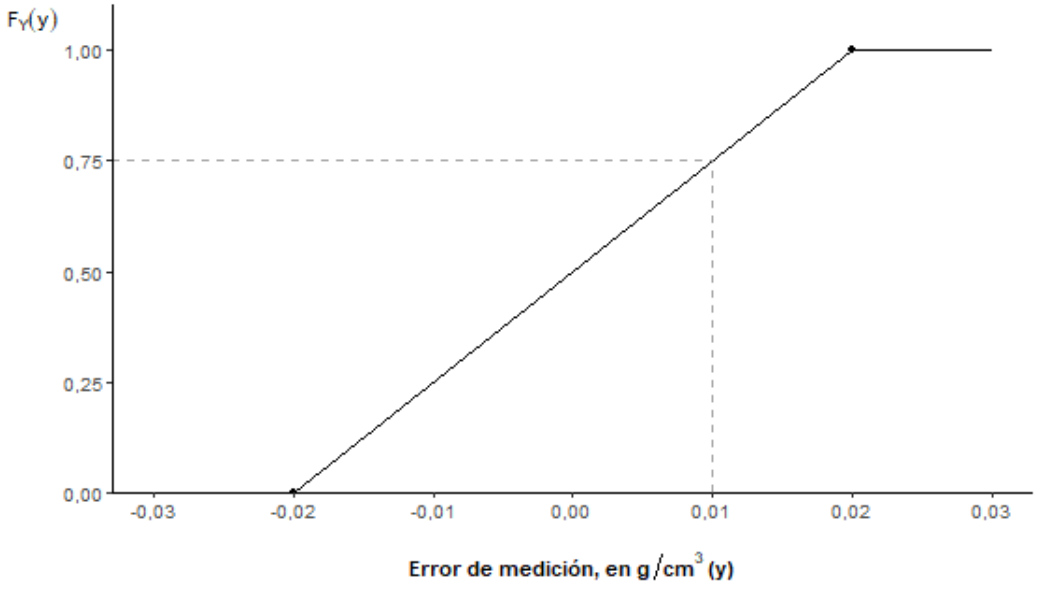
qunif(0.20,20,24)
20.8

qunif(0.75, 20,24)
23.0

qunif(0.20,20,24,lower.tail = FALSE)
23.2
```

a) Mencione la población en estudio, la variable de interés y la población estadística.
b) Indique cuánto valen los parámetros matemáticos del modelo Uniforme. Interprete ambos valores en términos del problema.
c) Exprese formalmente las probabilidades incluidas en la salida. Represéntelas gráficamente.
d) ¿Algun/os de los valores de la salida se corresponde con un percentil? Expréselo/s formalmente y represéntelo/s en el gráfico de la función de distribución acumulada, $F_Y(y)$.
e) Complete las siguientes afirmaciones:
   1) El 20 % de los puntos de ese tramo de ruta tienen espesor menor a.....
   2) La proporción de puntos con espesor mayor a 21 cm vale.....
   3) La máxima diferencia que se observa en el espesor del 50 % central de los puntos es .....cm.
   4) Considere los valores de la variable y_1 y y_2, de los que se conoce que $y_2 - y_1 = 1$. $P(y_1 \leq Y \leq y_2)$ = ...

9. El error que se comete al medir la densidad de una sustancia (en gramos por cm^3) es una variable aleatoria continua Y con distribución Uniforme en el intervalo $[-0,02; b]$. A continuación se presenta el gráfico de la función de distribución acumulada de la variable de interés con información sobre la misma.

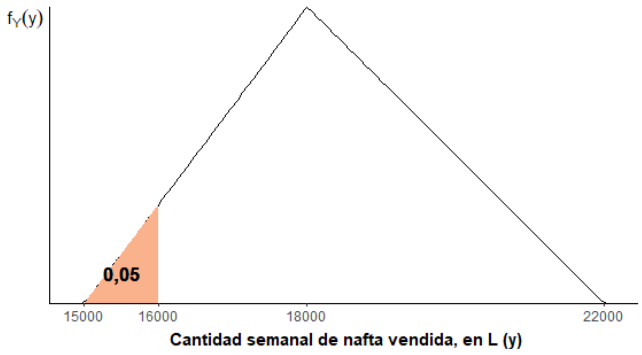


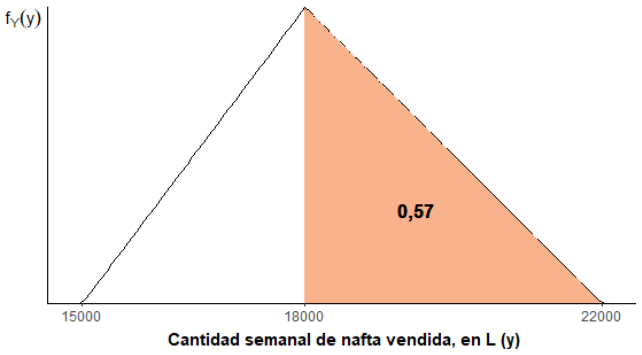
a) Exprese formalmente la probabilidad representada en el gráfico y, a partir de dicha expresión, determine el valor del parámetro b.
b) Obtenga la media y el desvío estándar de los errores de las mediciones de la densidad.
c) Indique cuánto valen los cuartiles (Q_{1Y}, Q_{2Y} y Q_{3Y}) y señálelos en el gráfico. Interprételos en contexto.
d) Calcule la proporción de mediciones cuyo error es al menos 0,015 g/cm^3.
e) A partir del resultado obtenido en el ítem anterior, indique cuánto vale la probabilidad de que el error de una medición sea a lo sumo $-0,015$ g/cm^3. Justifique adecuadamente.
f) Calcule la probabilidad de que en una medición no se cometa error. Explique el porqué del resultado obtenido.

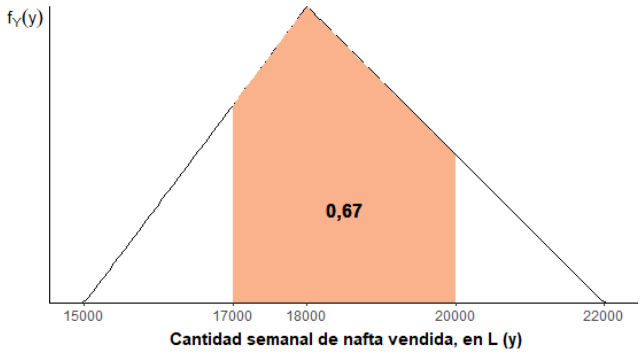
10. En un gran centro de reparaciones de equipos electrónicos se conoce que el costo de reparación (en pesos) de los equipos de cierto tipo es una variable aleatoria C que depende del tiempo T (en horas) que insume dicha tarea: $C = 8000 + 1000T$. Suponga que es razonable pensar que $T \sim U(2,20)$. Una empresa que cuenta con muchos esquipos electrónicos de ese tipo ofrece pagar un monto fijo de $25000 por cada reparación. Los encargados del centro de reparaciones firmarían un contrato con esa empresa solo si saben que la proporción de reparaciones más costosas es menor a 0,10.
    a) Defina la población y las variables de interés.
    b) Calcule e interprete la esperanza y el desvío estándar de ambas variables.
    c) ¿Recomendaría Ud. al centro de reparaciones firmar el contrato con la empresa? Justifique.

11. Una obra social cuenta con un servicio de atención telefónico de sus afiliados, dedicado exclusivamente a la autorización de prácticas médicas. Por estudios previos se sabe que, para cada llamada realizada por un afiliado, el tiempo de espera en línea (en segundos), X, hasta que esta es atendida por un operador varía aleatoriamente con distribución Triangular cuyos parámetros matemáticos son 20 s, 30 s y 50 s.
    a) Indique la población en estudio, la variable que se mide (clasifíquela) y la población estadística.
    b) Bosqueje la función de densidad de probabilidad de la variable y comente qué información brinda sobre esta población estadística.
    c) ¿Cuál es la probabilidad de que en una llamada de un afiliado el tiempo de espera en línea sea de a lo sumo 35 s?
    d) ¿En qué proporción de llamadas de afiliados el tiempo de espera en línea es al menos 40 s?
    e) ¿Cuál es la probabilidad de que en la llamada de un afiliado el tiempo de espera en línea oscile entre 25 y 45 s ?
    f) ¿Cuál es el tiempo de espera en línea superado por el 20 % de las llamadas de afiliados?
    g) Represente gráficamente los valores obtenidos en los ítems c) al f).

12. El dueño de una estación de servicio necesita describir la cantidad de nafta (en litros) vendida por semana. Los registros de las ventas anteriores indican que cada semana se venden un mínimo de 15000 L y un máximo de 22000 L, y la mayoría de las semanas se venden 18000 L. Al no contar con información precisa de como se comporta dicha variable se modela mediante una distribución Triangular.
    Se cuenta con los siguientes gráficos construidos con `R`:







a) Mencione cuál es la población en estudio, la variable de interés y la población estadística.

b) Indique cuánto valen los parámetros matemáticos del modelo Triangular en este caso. Interprete estos valores en términos del problema.

c) Exprese formalmente las probabilidades incluidas en la salida.

d) ¿Cuál es la cantidad de nafta máxima vendida en el 43 % de las semanas de menor venta?. Exprese formalmente ese valor.

13. En un hospital de la ciudad, a partir de los registros de pacientes entre 20 y 40 años de edad de los últimos 10 años, se ha determinado que la presión sanguínea sistólica (en milímetros de mercurio) en ese grupo etario puede modelarse mediante una distribución Triangular. Se conoce que el mínimo y el máximo son, respectivamente, 80 mm Hg y 150 mm Hg, y que la presión sanguínea promedio de la población es 115 mm Hg.

    a) A partir de la información brindada, determine el valor de los parámetros matemáticos de la distribución.

    b) El parámetro matemático c, ¿con qué medida/s coincide? Interprete una de ellas en términos del problema.

    c) Bosqueje la función de densidad de probabilidad. ¿Qué particularidad observa? Comente.

    d) ¿Cuál es la probabilidad de que una persona de entre 20 y 40 años seleccionada al azar tenga una presión sanguínea sistólica que supere los 145 mm Hg.?

    e) ¿Qué porcentaje de personas de entre 20 y 40 años tiene una presión arterial de entre 100 mm Hg y 120 mm Hg.?

    f) ¿Cuánto vale el desvío estándar de la presión sistólica para las personas de entre 20 y 40 años, si es válido el modelo indicado en el hospital? ¿Se trata de un parámetro o de un estadístico? Justifique.

14. Considere la situación del Problema 6 (Sección 1.4, página 7). Suponga que se conoce que el comportamiento del espesor (en centímetros) de los puntos de ese tramo de ruta se comporta según un modelo Triangular simétrico, con parámetros matemáticos $a = 20$ cm y $b = 24$ cm.

Suponga también que la empresa encargada de pavimentar ese tramo de ruta debe asegurar que como mínimo el 95 % de los puntos tenga espesor superior a 21 cm.
Si esto no se cumple, el organismo de control debe sancionar económicamente a la empresa y exigirle la repavimentación del tramo.

a) En base a la información disponible, plantee exhaustivamente el problema.
b) Informe qué debe hacer el organismo de control.
c) En caso de no cumplirse lo requerido, piense cómo se deberían modificar los parámetros matemáticos para que se cumpla lo pedido. ¿Qué acciones cree que se deberían llevar a cabo en la pavimentación para que esto ocurra?

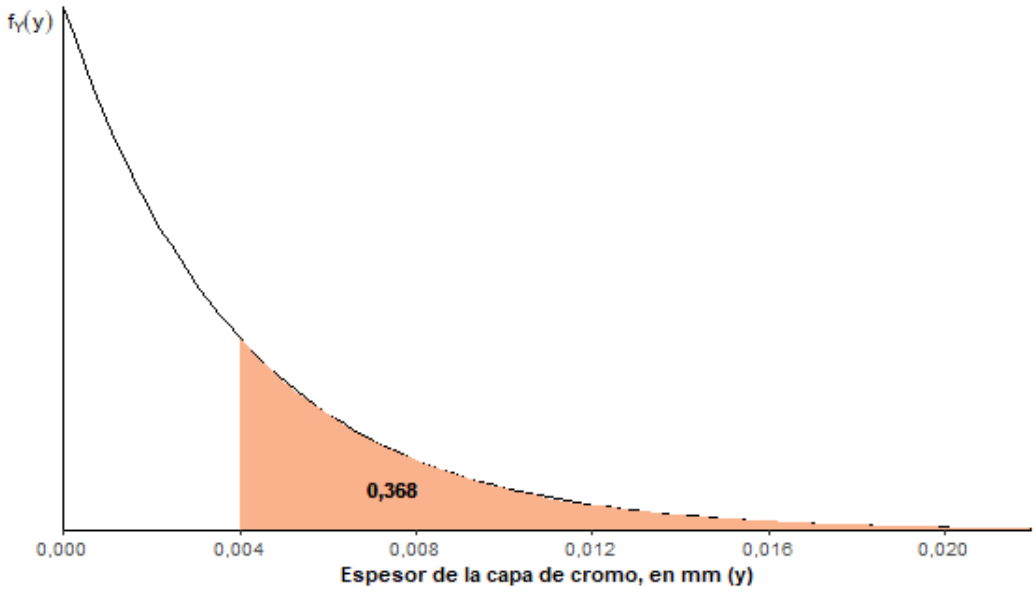
15. Los administradores de un sitio web conocen, por experiencia, que el tiempo (en segundos) que transcurre entre una visita y la siguiente se comporta según el modelo Exponencial, con parámetro $\alpha = 0,2\ s^{-1}$.
    a) Indique la población en estudio, la variable que se mide (clasifíquela) y la población estadística.
    b) Bosqueje la función de densidad de probabilidad de la variable y comente qué información brinda sobre esta población estadística.
    c) Para un par de visitas consecutivas, seleccionadas al azar, ¿cuál es la probabilidad de que transcurran a lo sumo 4 s entre ellas?
    d) Calcule la probabilidad de que entre una visita y la siguiente transcurran como mínimo 2 s.
    e) ¿En qué proporción de pares de visitas consecutivas, transcurren entre 1 y 4 s?
    f) ¿Qué valor del tiempo transcurrido entre dos visitas consecutivas no es superado por el 20 % de los pares?
    g) Represente gráficamente los valores obtenidos en los ítems *c*) al *f*).

16. Una empresa se especializa en la producción de un cierto tipo de herramientas de corte, a las cuales reviste con una capa de cromo. Una característica de interés es el espesor de esta capa, Y (en milímetros).
    Para los usuarios, una herramienta resulta de calidad aceptable si el espesor de la capa de cromo es un valor mayor que 0,001 mm. Para los fabricantes resulta antieconómico comercializar herramientas que tengan una capa de cromo con espesor mayor a 0,005 mm.
    Conociendo el modelo apropiado para describir el comportamiento del espesor de cromo, se obtuvo lo siguiente, con `R`:

```
pexp(0.001, 250)
0.221
```

```
pexp(0.005, 250, lower.tail=FALSE)
0.287
```

```
qexp(0.50, 250)
0.0028
```



a) Mencione cuál es la población en estudio, la variable de interés y la población estadística.
b) Según la información de la salida de R, ¿a qué distribución se aproxima el comportamiento del espesor de la capa de cromo?
c) Obtenga e interprete E(Y) y D(Y) en términos del problema.
d) ¿Alguno de los valores obtenidos con R se corresponde con un percentil? Expréselo formalmente y señalelo en el gráfico de la distribución. Interprételo en términos del problema.
e) Exprese formalmente a las probabilidades obtenidas con R e interprételas en términos del problema.
f) Considerando el espesor de la capa de cromado, ¿cuál es la proporción de herramientas que no conviene comercializar (por lo económico) entre las que son aceptables para los clientes?
   1) Exprese formalmente a esta proporción y obtenga su valor.
   2) El valor obtenido en el ítem anterior, ¿coincide con alguna de las probabilidades obtenidas con R? ¿Por qué?

17. La duración (en horas) de un cierto tipo de componente electrónico es una variable aleatoria T con distribución Exponencial. Se conoce que el desvío estándar de la duración para estos componentes vale $4347,82$ h.
   a) Determine el valor del parámetro matemático de la distribución de probabilidades de las duraciones de los componentes. Justifique.
   b) Obtenga e interprete el promedio y la mediana de la duración de los componentes. Justifique la relación entre ambos valores.
   c) Calcule e interprete el rango intercuartílico.
   d) Para los siguientes ítems, responda sin realizar cálculos. Justifique su respuesta.
      1) Obtenga la probabilidad de que un componente dure más de 3013,7 h.
      2) Indique el valor de la probabilidad de que un componente dure más de 4013,7 h, si ya ha durado 1000 h. Exprese formalmente a esta probabilidad.

18. En una metalúrgica producen alambres de diferentes tipos y calidades. Para los alambres de calidad alta, se está estudiando la ocurrencia de imperfecciones. Se sabe por experiencia que la distancia (en metros) entre una imperfección y la siguiente se distribuye según el modelo

Exponencial, con $\alpha = 0,004$ m^{-1}. Para que en los rollos no se presenten imperfecciones, cuando aparece una, se corta el alambre y se enrolla, obteniéndose rollos de longitud variable.

a) Si se define como población de interés a la totalidad de rollos de alambre de calidad alta, obtenga la longitud promedio por rollo y el desvío estándar. Interprete ambas medidas, indicando si se trata de parámetros o estadísticos.

b) Los clientes pretenden que los rollos tengan como mínimo 100 m. Un analista, observando el valor de la longitud promedio de los rollos, afirma que con seguridad casi la totalidad de ellos va a cumplir con esa condición.

1) Usted, ¿qué opina de esa afirmación?
2) Mencione otra medida que informe sobre el cumplimiento de la pretensión de los clientes y obtenga su valor.

c) En la empresa pretenden que más del 80 % de sus rollos superen los 100 m, aunque solo iniciarían acciones correctivas en el proceso si no se alcanza el 70 %. Interesa saber si deben iniciarse acciones correctivas.

Plantee brevemente el problema (defina población, variable, parámetro de interés y objetivo en términos de dicho parámetro) e informe sus conclusiones.

19. Lea atentamente las siguientes situaciones, asociadas a algunos de los problemas planteados al comienzo del libro (Sección 1.4, página 7).

- (Problema 1) En el proceso de producción de barras de acero, se pretende observar si una barra elegida al azar tiene longitud dentro de las especificaciones o no.
- (Problema 4) Se observa si una devolución seleccionada al azar se debe a problemas de retraso en la entrega o no.
- (Problema 6) En un tramo de ruta se observa si un punto elegido al azar cumple o no cumple con los requerimientos en cuanto a su resistencia.
- (Problema 8) Se registra si un dispositivo de cierto tipo elegido al azar dura más de 150 horas o no.

a) Para cada situación defina la población e indique cuál es la variable cualitativa de interés. ¿Qué tienen en común estas variables?

b) A partir de la variable cualitativa mencionada, defina una variable aleatoria dicotómica donde el valor 1 se asocie al resultado que se considere de interés en cada caso.

c) ¿Con qué información se necesitaría contar para construir la distribución de probabilidades de cada una de las variables aleatorias definidas en el ítem anterior?

20. Considere la situación del Problema 6 (Sección 1.4) y suponga que la empresa que se ocupa de la pavimentación de la ruta afirma que el 98 % de los puntos cumple con los requerimientos en cuanto a su resistencia. Se define la variable C que toma el valor 1 si el punto cumple y 0 en caso contrario.

a) ¿Cuál es la población y la población estadística asociada a la variable C?

b) ¿Qué distribución de probabilidades tiene C? Construya su función de probabilidad puntual y represéntela gráficamente.

c) Calcule e interprete la esperanza y el desvío estándar de la variable C.

21. Considere una variable Y con distribución Bernoulli de parámetro π.

a) Analice entre qué valores puede estar comprendido su desvío estándar e indique cuál es el valor máximo que puede tomar esta medida. ¿Para qué valor de π se da dicho máximo?

b) Si $\pi = 0,05$, ¿cuánto valen la media y el desvío estándar de Y?
c) Si $\pi = 0,95$, ¿cuánto valen la media y el desvío estándar de Y?

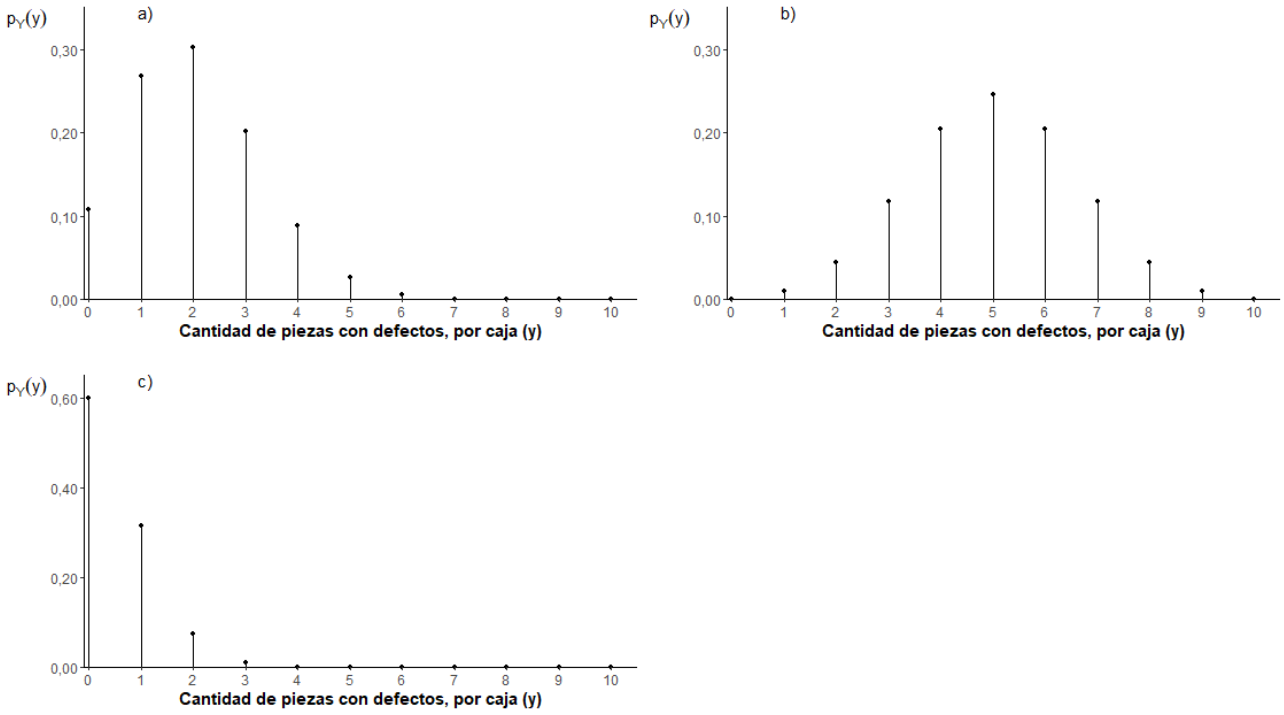
22. Una distribuidora de productos alimenticios ha adoptado la política de hacer un descuento del 10 % a los clientes que paguen en efectivo en vez de hacerlo con tarjeta de crédito. Su experiencia indica que el 50 % de los clientes adoptan el descuento. Sea *X: número de clientes que aceptan el descuento entre los próximos 20.*

a) Indique la población en estudio, la variable que se mide (clasifíquela) y la población estadística.
b) En esta situación, ¿parecen razonables las hipótesis para considerar que la variable sigue una distribución Binomial? Justifique.
c) Suponiendo que el modelo Binomial es razonable, bosqueje la función de probabilidad puntual de la variable y comente qué información brinda sobre esta población estadística.
d) Encuentre la probabilidad de que exactamente 5 de entre los próximos 20 clientes acepten el descuento.
e) ¿Cuál es la proporción de grupos de 20 clientes en los cuales a lo sumo 6 acepten el descuento?
f) Si se elige un grupo de 20 clientes al azar, ¿cuál es la chance de que al menos 10 clientes acepten el descuento?
g) Encuentre la probabilidad de que entre 8 y 12 clientes (incluidos ambos) de entre los próximos 20 acepten el descuento.
h) ¿Qué cantidad de clientes que aceptan el descuento no es superada por el 20 % de los grupos?
i) Represente gráficamente los valores obtenidos en los ítems *d*) al *h*).

23. Una empresa fabrica tres modelos diferentes de piezas de porcelanato para pisos. Las piezas se comercializan en cajas de n unidades.
Una de las características más importantes de las piezas es la presencia de defectos en la superficie. Suponga que, para cada modelo se conoce la proporción de piezas con defectos.
Interesa estudiar el comportamiento de la cantidad de piezas con defectos por caja, para cada modelo (Y_i). Se cuenta con la siguiente información, obtenida con `R`:

Modelo A	Modelo B	Modelo C

`pbinom(1,10,0.05)`
`0.91`
 	
`pbinom(5,10,0.05)`
`0.999`
	`pbinom(5,10,0.20,`
	`lower.tail = FALSE)`
	`0.006`
	
	`dbinom(0,10,0.20)`
	`0.107`
	
	`dbinom(1,10,0.20)`
	`0.268`
		`pbinom(5,10,0.50,`
		`lower.tail = FALSE)`
		`0.377`
		
		`dbinom(2,10,0.50)`
		`0.0439`
		
		`dbinom(3,10,0.50)`
		`0.117`
		
		`dbinom(4,10,0.50)`
		`0.205`
		
		`dbinom(5,10,0.50)`
		`0.246`



a) Defina para cada modelo la variable, la población y la población estadística correspondientes.

b) Suponiendo que el modelo Binomial es razonable, identifique cuál es el gráfico correspondiente al número de piezas con defectos por caja, para cada uno de los modelos de porcelanato. Justifique.

c) Complete la tabla que se presenta a continuación, con las probabilidades de los sucesos definidos en la primera columna, para cada modelo. Exprese formalmente los cálculos realizados. Comente brevemente qué se concluye de la lectura de la tabla.

Suceso	Modelo A	Modelo B	Modelo C
La cantidad de piezas con defectos es a lo sumo 1 ($Y_i \leq 1$)			
La cantidad de piezas con defectos está entre 2 y 5, ambos incluidos ($2 \leq Y_i \leq 5$)			
La cantidad de piezas con defectos es mayor que 5 ($Y_i > 5$)			

24. Una compañía pequeña utiliza un servicio de paquetería para enviar los pedidos de tabla de quesos especiales que son para regalo. El servicio contratado garantiza que el 95 % de los pedidos se entregan a tiempo a los clientes. Además se sabe que para grupos de n pedidos, el número promedio de pedidos que se entregan a tiempo es 19.
    a) Identifique la variable de interés y proponga un modelo para describir su comportamiento en probabilidad.
    b) Defina la población asociada a la variable de interés.
    c) Sin graficar la distribución de probabilidades de la variable, indique qué forma tiene. Justifique.
    d) Calcule dos medidas de variabilidad e interprételas en términos del problema.

25. Reconsidere la situación del Problema 6 (Sección 1.4, página 7), relativa al proceso de pavimentación de la ruta. Suponga que un ente de control va a seleccionar una muestra de n = 30 puntos en cada tramo y si el número de puntos que cumple con las normas en relación a la resistencia, X, es 25 o menos, le exigirá al comitente la repavimentación de ese tramo de ruta.
    a) Si la proporción de puntos que cumplen con las normas en un tramo determinado, π, vale $0,98$, ¿cuál es la chance de que ese tramo deba ser repavimentado?
    b) ¿Qué sucederá con esta chance si disminuye el valor de π? Explique.

26. Suponga que este mes hay 2500 nuevos inscriptos del Gran Rosario en Netflix, 125 lo hicieron con una tarjeta de crédito particular. Se está interesado en estudiar el perfil de esos nuevos clientes para ofrecerles un pack. Se eligen al azar un grupo de 50 clientes e interesa el número de clientes que pagaron con esa tarjeta.
    a) Indique la población en estudio, la variable que se mide (clasifíquela) y la población estadística.
    b) Bosqueje la función de probabilidad puntual de la variable y comente qué información brinda sobre esta población estadística.
    c) Encuentre la probabilidad de que se elijan exactamente 5 nuevos inscriptos que hayan pagado con esa tarjeta.
    d) ¿Cuál es la proporción de grupos de 50 clientes en los cuales a lo sumo 6 paguen con esa tarjeta?
    e) Si se elige un grupo de 50 clientes al azar, ¿cuál es la chance de que al menos 10 clientes paguen con esa tarjeta?
    f) Encuentre la probabilidad de que entre 8 y 12 clientes (incluidos ambos) del grupo paguen con esa tarjeta.

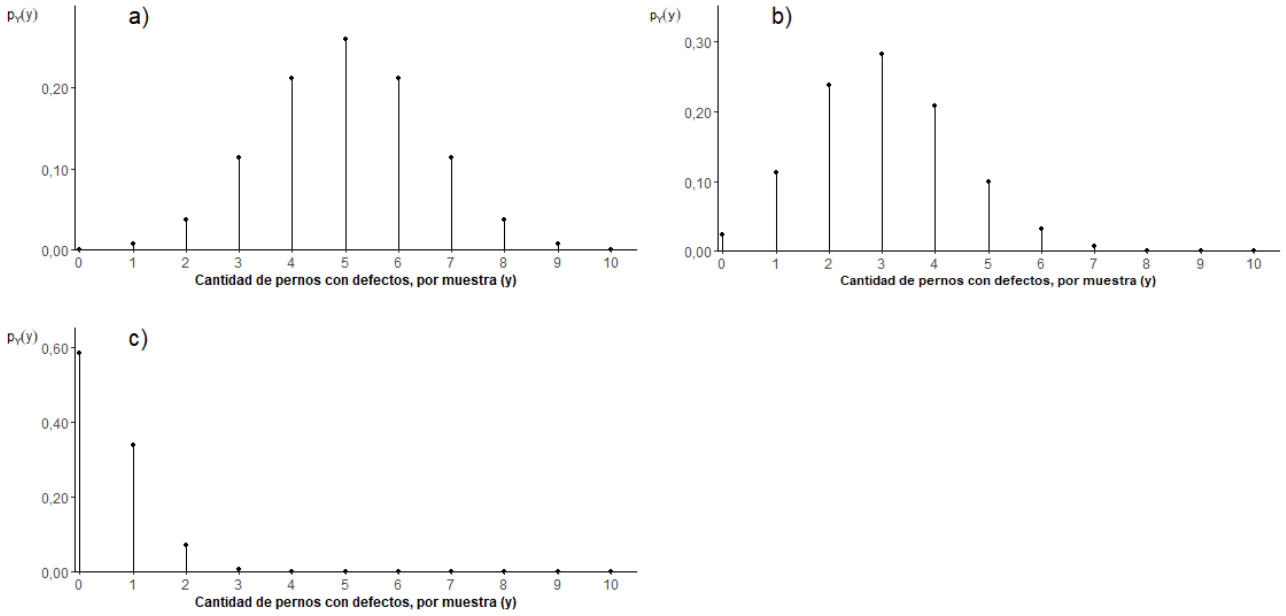
g) ¿Qué cantidad de clientes que pagan con esa tarjeta es superado por el 20 % de los grupos?
h) Represente gráficamente los valores obtenidos en los ítems *c*) al *g*).

27. Una empresa fabrica pernos de precisión para uso industrial y los envía a sus clientes en lotes de 100 unidades. Algunos pernos pueden presentar defectos de fabricación detectables únicamente mediante una inspección técnica.
Antes de despachar un lote, la empresa selecciona al azar y sin reposición una muestra de 10 pernos y registra la cantidad de pernos defectuosos encontrados en ella.
Suponga que, en un lote particular, hay 5 pernos defectuosos.
Se cuenta con la siguiente información, obtenida con R:

```
dhyper(0,5,95,10)
0.584

dhyper(1,5,95,10)
0.339

phyper(2,5,95,10, lower.tail = FALSE)
0.0066
```



a) Defina la variable aleatoria de interés, la población y la población estadística, suponiendo que se repite el procedimiento de muestreo sobre lotes con las características mencionadas.
b) Indique cómo se distribuye la variable aleatoria e identifique el gráfico correspondiente. Justifique su elección.
c) ¿En qué proporción de las muestras posibles hay a lo sumo un perno defectuoso? Exprese formalmente los cálculos realizados.
d) ¿Cuál es la probabilidad de que una muestra tenga entre 2 y 5 pernos defectuosos?
e) ¿Qué representa el valor 0,0066 obtenido con R? Expréselo formalmente e interprételo en contexto.

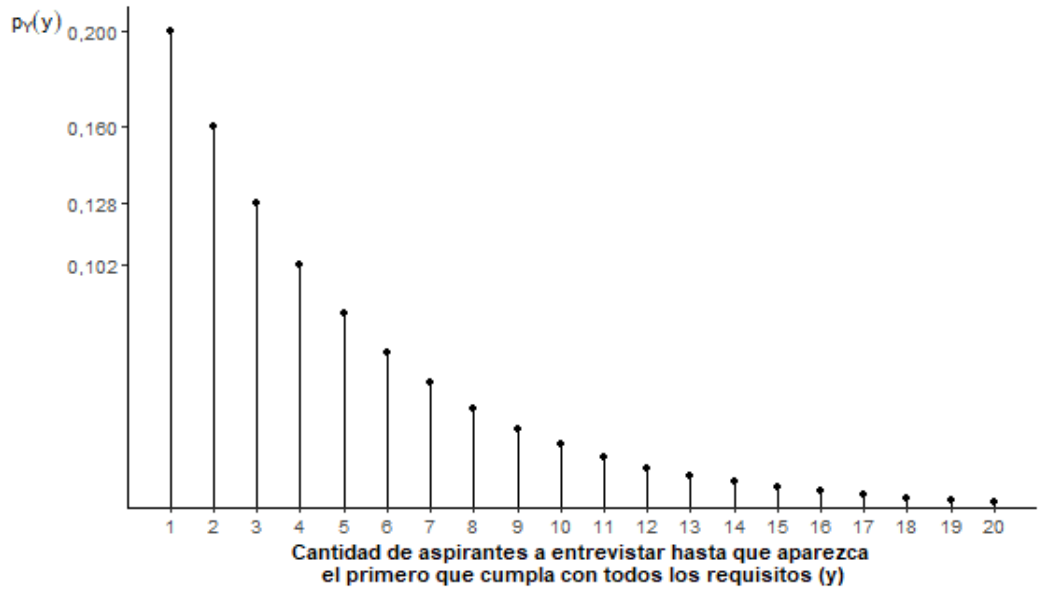
28. Una automotriz fabrica una de las autopartes por tandas de a 50 unidades y se van seleccionando 5 de ellas para que pasen a la línea de ensamblado. Para una tanda particular, se sabe que el número promedio de autopartes defectuosas entre las 5 es 0,2.
    a) Identifique la variable de interés y proponga un modelo para describir su comportamiento en probabilidad.
    b) Determine el total de defectuosas para esa tanda de 50 autopartes.
    c) Sin graficar la distribución de probabilidades de la variable, indique qué forma tiene. Justifique.
    d) Calcule dos medidas de variabilidad e interprételas en términos del problema.

29. Suponga que una vendedora arma lotes de 30 piezas con 2 que son defectuosas. Un comprador utiliza el siguiente plan de muestreo para la aceptación de cada lote:
    Si en una muestra de 7 piezas, seleccionadas sin reposición, encuentra al menos dos piezas defectuosas, rechaza el lote; de lo contrario lo acepta.
    La vendedora considerará adecuado el plan de muestreo de ese comprador si como máximo le rechaza el 8 % de los lotes que ella le envía.
    a) Calcule la probabilidad de que un lote con las características mencionadas sea aceptado.
    b) ¿Le parece que el plan de muestreo para la aceptación es adecuado para la vendedora? Justifique.

30. Reconsidere la situación descripta en la Actividad 22, referida a una distribuidora de productos alimenticios. Suponga que interesa estudiar el comportamiento del número de clientes que abonan hasta que se registra el primero que paga en efectivo, este incluido, (Y).
    a) Defina la población y la población estadística asociadas a la variable de interés.
    b) Si se sigue considerando que el 50 % de los clientes pagan en efectivo, ¿qué distribución puede asignársele a la variable Y? Justifique adecuadamente, indicando las hipótesis que deben verificarse para que el modelo mencionado sea razonable.
    c) Bosqueje la función de probabilidad puntual de la variable y comente qué información brinda sobre esta población estadística.
    d) ¿Cuánto vale la cantidad media de clientes que se presentan hasta que uno paga en efectivo? ¿Y el desvío estándar?
    e) ¿Qué proporción de las secuencias[3] tiene 3 clientes?
    f) ¿Cuánto vale la probabilidad de que una secuencia tenga como mínimo 2 clientes?
    g) ¿Cuál es la chance de que en una secuencia haya como máximo 2 clientes?
    h) La proporción de secuencias que tienen entre 2 y 4 clientes, ambos incluidos, ¿cuánto vale?
    i) ¿Qué valor de la variable es superado por el 60 % de las secuencias?

31. El departamento de selección de personal de una empresa sabe que ante cada convocatoria para cubrir un puesto gerencial, solo el 20 % de los aspirantes cumple con todos los requisitos exigidos. Se hace una convocatoria para cubrir el puesto de un gerente por una reciente jubilación y se entrevista a los aspirantes uno a uno. Interesa la cantidad de aspirantes que se deberán entrevistar hasta encontrar uno que cumpla con todos los requisitos exigidos (Y). Considere la siguiente información obtenida con `R`:

[3] Se refiere a secuencias de clientes que llegan a la empresa hasta que uno paga en efectivo, este incluido



a) Mencione cuál es la población en estudio, la variable de interés y la población estadística.

b) ¿A qué distribución se aproxima el comportamiento de la variable de interés? ¿Qué condiciones se deben cumplir para que esta distribución sea la apropiada?

c) Observe la función de probabilidad puntual representada en el gráfico y analice, a partir de ella, la información que proporciona sobre la población estadística, considerando la forma general de la distribución, los valores posibles de la variable y los valores con mayor probabilidad de ocurrencia, entre otros aspectos.

d) Obtenga e interprete E(Y) y D(Y) en términos del problema.

e) Indique cuánto vale la probabilidad de que la tercera persona entrevistada sea la primera que cumpla con todos los requisitos de la convocatoria. Exprésela formalmente.

f) Indique cuánto vale la probabilidad de que se deban entrevistar al menos a 4 personas para encontrar a la primera que cumple con todos lo requisitos exigidos. Exprésela formalmente.

32. Una distribuidora cuenta con un sistema informático para gestionar los pedidos de los clientes. Este sistema falla aleatoriamente y queda sin funcionar algunos minutos, lo cual suele traer inconvenientes. En la empresa registran para cada día si el sistema ha fallado o no y estudian el comportamiento de las secuencias de días que transcurren hasta que un día el sistema falla. Por experiencia se conoce que, en promedio pasan 20 días hasta que un día el sistema falla.

a) Identifique la variable de interés y la población asociada.

b) ¿Cómo se distribuye dicha variable? ¿Cuánto vale el parámetro matemático? Interprete su valor en términos del problema.

c) ¿Qué significa, en términos del problema, $P(Y = 15)$? Obtenga la probabilidad mencionada e interprete su valor en contexto.

33. En los gráficos de control, utilizados para monitorear un proceso de producción de bienes o servicios, se señalan periódicamente puntos obtenidos con información de muestras del proceso. Si un punto cae fuera de los límites de control se considera una señal de alarma.
Suponga que cierto gráfico de control fue diseñado para cumplir con lo siguiente:

- si el proceso funciona adecuadamente, la probabilidad (π) de que un punto caiga fuera de los límites (es decir, de que ocurra una falsa alarma) es 0,0027.
- si el proceso funciona mal, la probabilidad (π) de que un punto caiga fuera de los límites

(es decir, de que suene bien la alarma) es 0,90.

Interesa la cantidad de puntos que se deben señalar en el gráfico hasta que uno cae fuera de los límites (Y), que en el control de procesos se denomina **"longitud de corrida"**.

a) Considere que el proceso está funcionando adecuadamente.

1) ¿Cómo se comporta la variable Y_1: longitud de corrida? Esquematice la distribución.
2) ¿Cuál es la población asociada a la variable mencionada?
3) ¿Cuánto vale el promedio de la longitud de corrida (o longitud de corrida promedio)? Interprete ese valor en contexto.

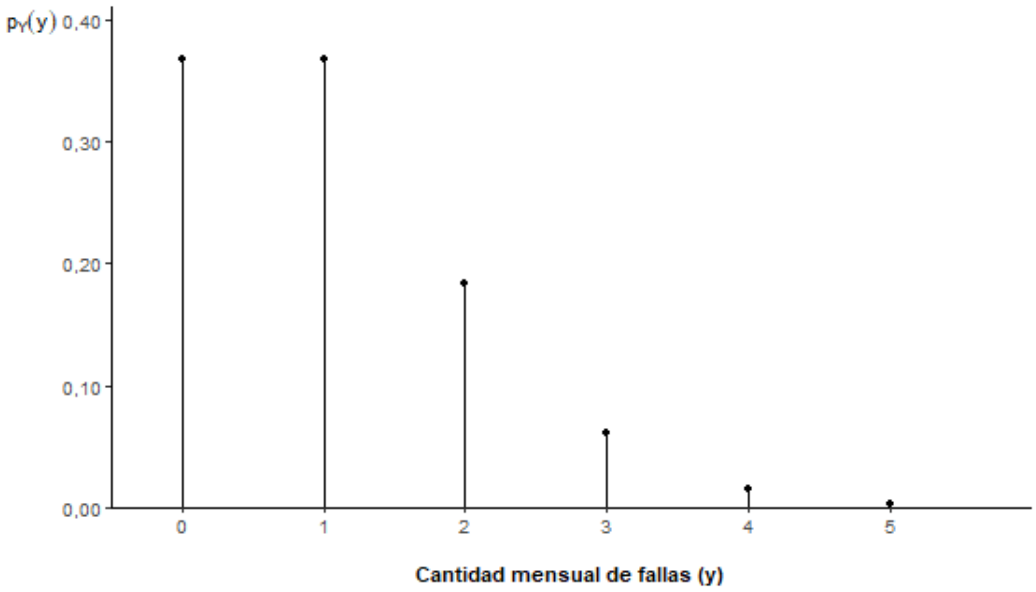
b) Considere ahora que el proceso está funcionando mal.

1) ¿Cómo se comporta la variable Y_2: longitud de corrida? Esquematice la distribución.
2) ¿Cuál es la población asociada a la variable mencionada?
3) ¿Cuánto vale el promedio de la longitud de corrida (o longitud de corrida promedio)? Interprete ese valor en contexto.

34. Una compañía aseguradora recibe reclamos por robos de cubiertas de vehículos en una determinada ciudad. El comportamiento del número de reclamos se puede aproximar con una distribución Poisson a razón de $2,25$ por semana.

a) Indique la población en estudio, la variable que se mide (clasifíquela) y la población estadística.
b) En esta situación, ¿parecen razonables las hipótesis para considerar que la variable sigue una distribución Poisson? Justifique.
c) Suponiendo que el modelo Poisson es razonable, bosqueje la función de probabilidad puntual de la variable y comente qué información brinda sobre esta población estadística.
d) Encuentre la probabilidad de que se reciban exactamente 5 reclamos en la próxima semana.
e) ¿Cuál es la proporción de semanas en las que se hacen 2 reclamos?
f) Si se elige una semana al azar, ¿cuál es la chance de que al menos haya un reclamo por robo de cubiertas?
g) Encuentre la probabilidad de que en una semana haya entre 2 y 5 reclamos (incluidos ambos).
h) ¿Qué cantidad de reclamos es superado en el 20 % de las semanas?
i) Represente gráficamente los valores obtenidos en los ítems *d*) al *h*).

35. En una empresa distribuidora de energía eléctrica de la provincia de Santa Fe interesa estudiar el comportamiento de la cantidad mensual de fallas en el proceso de distribución de la energía (Y). Suponga que Y se comporta según el modelo Poisson, con un promedio de α fallas por mes.
El gráfico de la distribución se presenta a continuación y luego se indican algunos valores, obtenidos con `R`:




```
ppois(2,1)
0.92

dpois(2,1)
0.184

ppois(4,1)
0.996

qpois(0.5,1)
1

qpois(0.25,1)
0

qpois(0.75,1)
2
```

a) Indique cuánto vale el parámetro matemático de la distribución (α). Obtenga e interprete $E(Y)$ y $V(Y)$.

b) Observe el gráfico de la función de probabilidad puntual correspondiente a la cantidad mensual de fallas y complete las siguientes afirmaciones:

- Para la mayoría de los meses, la cantidad mensual de fallas oscila entre ..... y ..... (aprox.)
- Las cantidades de fallas mensuales más probables son ..... y .....
- La distribución de la cantidad mensual de fallas, en cuanto a su simetría, es .....
- Es prácticamente imposible observar ..... fallas en un mes cualquiera, si el modelo es válido.

c) A partir de los valores obtenidos con R, complete las siguientes afirmaciones:

- La proporción de meses en los que se observan 2 fallas vale .....
- $P(Y \geq 5) =$ .....

- La probabilidad de que en un mes elegido al azar se observen a lo sumo 2 fallas es .....
- La máxima diferencia en el 50 % central de los meses es de ..... fallas.

36. En un centro de monitoreo de una empresa de telecomunicaciones se registran inconsistencias en los paquetes de información que llegan al servidor principal. La experiencia indica que las inconsistencias ocurren aleatoriamente en el tiempo y de manera independiente unas de otras. En este contexto se define la variable aleatoria Y: número de inconsistencias registradas en intervalos de 1 minuto, la cual se comporta según el modelo Poisson. Se conoce además que la desviación estándar, $D(Y)$, es igual a $1,25$.
    a) ¿Bajo qué hipótesis resulta razonable modelizar esta situación mediante una distribución de Poisson?
    b) Suponga que la distribución Poisson es adecuada y que se quiere estudiar el comportamiento de la variable Y_5*: Número de inconsistencias en períodos de 5 min*.
        1) Defina la población asociada a Y_5 e indique cómo se distribuye esta variable. Justifique.
        2) ¿Cuál es la probabilidad de que haya al menos 5 inconsistencias en un período de 5 min?
        3) ¿Cuál es la probabilidad de que haya a lo sumo 10 inconsistencias en un período de 5 min?
        4) Calcule e interprete la esperanza y la variancia de la variable Y_5 en el contexto de la situación planteada.

37. Sobre un equipo electrónico se conoce que el número de fallas que presenta se distribuye según un proceso de Poisson. Se conoce además que el número promedio de fallas por hora es 0,1. Considere las siguientes variables aleatorias:
    - X_1: número de fallas por hora
    - T: tiempo entre dos fallas consecutivas (en h)

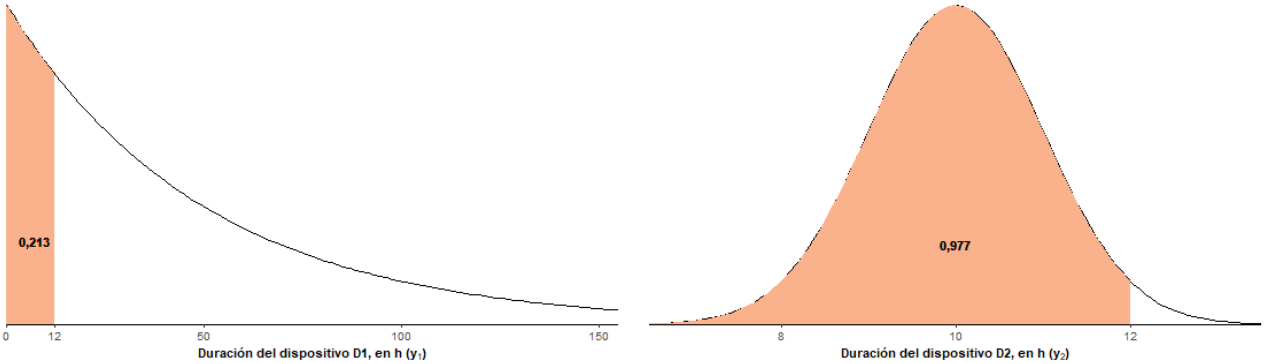
    a) Defina las poblaciones asociadas a las dos variables mencionadas.
    b) Obtenga la esperanza y el desvío estándar de cada una de ellas.
    c) Interesa que pase como mínimo una hora entre dos fallas consecutivas.
        1) Exprese al suceso de interés en función de la variable T. Obtenga e interprete su probabilidad.
        2) Realice lo mismo que en el ítem anterior, pero ahora en función de la variable X_1.
    d) Bosqueje las distribuciones de ambas variables y señale las probabilidades obtenidas.
    e) Obtenga e inteprete en términos del problema el valor de $P(T > 10)$.
    f) Exprese a la probabilidad del suceso considerado en el ítem anterior, en función de una variable con distribución Poisson. Justifique.

38. Una máquina fabrica cables de fibra óptica con ocasionales defectos de manufactura. El número de defectos de ese tipo, en tramos de "d" metros lineales (X_d) es una variable aleatoria con distribución Poisson con una media de 3 defectos cada 100 m de cable producido.
    a) Se consideran las cantidades de defectos en tramos de cable de 50 m. Defina la variable de interés e indique su distribución de probabilidades.
    b) ¿Qué proporción de tramos tendrán más de 2 defectos de manufactura?
    c) ¿Qué distribución tiene la variable aleatoria D: Distancia (en m) entre dos defectos consecutivos? Justifique.

d) Obtenga la probabilidad de que pasen más de 50 m entre un defecto y el siguiente.

e) Exprese a la probabilidad obtenida en el ítem anterior en función de la variable definida al inicio de esta actividad. Justifique.

39. Un sistema consta de dos dispositivos (D1 y D2) que funcionan simultánea e independientemente. La duración (en horas) del dispositivo D1, Y_1, es una variable aleatoria con distribución Exponencial de parámetro $0,02\ \text{h}^{-1}$, mientras que la del dispositivo D2, Y_2, tiene distribución Normal con parámetros $\mu = 10$ h y $\sigma = 1$ h. Se cuenta con los siguientes gráficos construidos con R:



a) ¿Cuál es la probabilidad de que al menos uno de los dispositivos dure menos de 12 h?

b) Suponga que los dispositivos están conectados en serie. Obtenga la probabilidad de que el sistema dure más de 12 h. Interprete este valor como una frecuencia relativa en el límite.

c) Suponga ahora que los dispositivos están conectados en paralelo y obtenga la misma probabilidad que en el ítem anterior.

d) ¿Cuál es la probabilidad de que el dispositivo D1 dure más de 20 h si se sabe que ya superó las 8 h de duración?

40. Una planta se dedica a la producción de envases para productos de perfumería, los cuales pueden presentar defectos tanto en la tapa como en la etiqueta. Para los envases de uno de los perfumes en particular, se conoce que la cantidad de defectos en la tapa (X) se comporta según una distribución Poisson con promedio $0,01$ y que la cantidad de defectos en la etiqueta (Y) se comporta según una distribución Poisson con promedio $0,03$. Debido al proceso de producción, es razonable suponer que ambas variables son independientes entre sí.
En la planta consideran que un envase de perfume es defectuoso cuando presenta al menos un defecto, sea en la tapa o en la etiqueta, dado que se trata de una fragancia de lujo.
Las siguientes probabilidades se obtuvieron con R:

```
dpois(0,0.01)
0.99

dpois(0, 0.03)
0.970
```

a) Detalle las poblaciones, las variables y las poblaciones estadísticas bajo estudio.

b) Esquematice la distribución de probabilidades para ambas variables definidas.

c) Informe cuánto vale la proporción de envases defectuosos.

d) Los envases se envían en lotes de 50 unidades a la empresa que produce el perfume. Informe la proporción de lotes en los cuales hay a lo sumo 4 envases defectuosos? Antes de hacer el cálculo explicite la variable y la distribución bajo estudio y exprese formalmente a dicha proporción.

41. En una empresa se fabrican ejes para dispositivos de almacenamiento óptico y se conoce por experiencia que el diámetro (en centímetros) de los mismos se distribuye Normal con media $0,652$ y desviación estándar $0,003$. Uno de sus mejores clientes requiere ejes que cumplan con la siguiente especificación para el diámetro: $0,650 \pm 0,005$ cm.
    a) ¿Cómo piensa que la empresa obtuvo la información acerca de la distribución de los diámetros de los ejes?
    b) ¿Qué proporción de los ejes fabricados por este proceso no cumple con la especificación?
    c) El cliente compra ejes en paquetes de 10 unidades. Para verificar si se cumple con la especificación pedida, cuando le llega un envío, selecciona un paquete al azar y mide los diámetros de los 10 ejes. Si encuentra al menos 2 fuera de las especificaciones, rechaza el envío. ¿Cuál es la probabilidad de que acepte un envío?
    d) Cuando le devuelven un pedido, el fabricante debe pagar los costos de envío, por lo que decide recalibrar el proceso para ajustarse a la especificación del cliente. Sabe que, luego de las modificaciones, la media de los diámetros es $0,65$ cm. ¿Cuál debería ser el máximo valor de la desviación estándar para que como mínimo el $99\,\%$ de los ejes cumpla con la especificación?

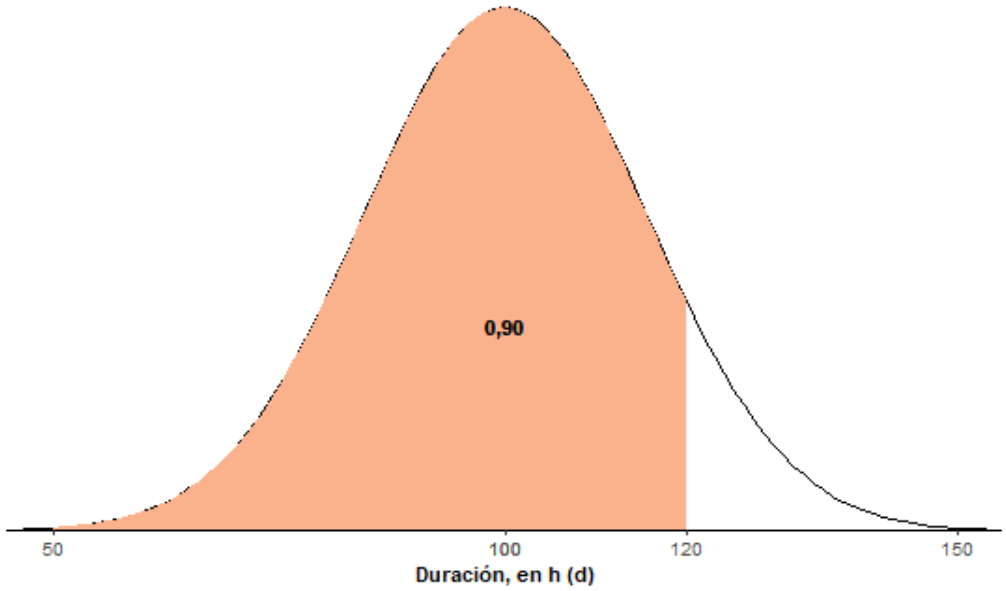
42. Se consideran tres variables aleatorias X, Y y L, que verifican lo siguiente: $X \sim N(3; 0,2)$, $Y \sim U(2,4; 3,6)$ y $L \sim Bi(10; 0,3)$
    a) Complete la siguiente tabla:

Medidas de interés	$X \sim N(3; 0,2)$	$Y \sim U(2,4; 3,6)$	$L \sim Bi(10; 0,3)$
Promedio			
Desvío estándar			
Primer cuartil			
Segundo cuartil			
Tercer cuartil			
Rango intercuartílico			
Proporción de valores menores que 3			
Proporción de valores iguales a 3			

    b) Indique si las medidas obtenidas en el ítem anterior son estadísticos o parámetros. Justifique su respuesta.
    c) Esquematice de manera comparativa las funciones de densidad de probabilidad y de probabilidad puntual de las variables consideradas en esta actividad.
    d) Proponga, a modo de ejemplo, una variable aleatoria que pueda tener una distribución como la de la variable L. Interprete dos de las medidas obtenidas para dicha variable en el ítem a
    e) En una industria metalúrgica se producen piezas especiales para hornos industriales con dos máquinas, A y B. Suponga que las variables X e Y corresponden a los diámetros

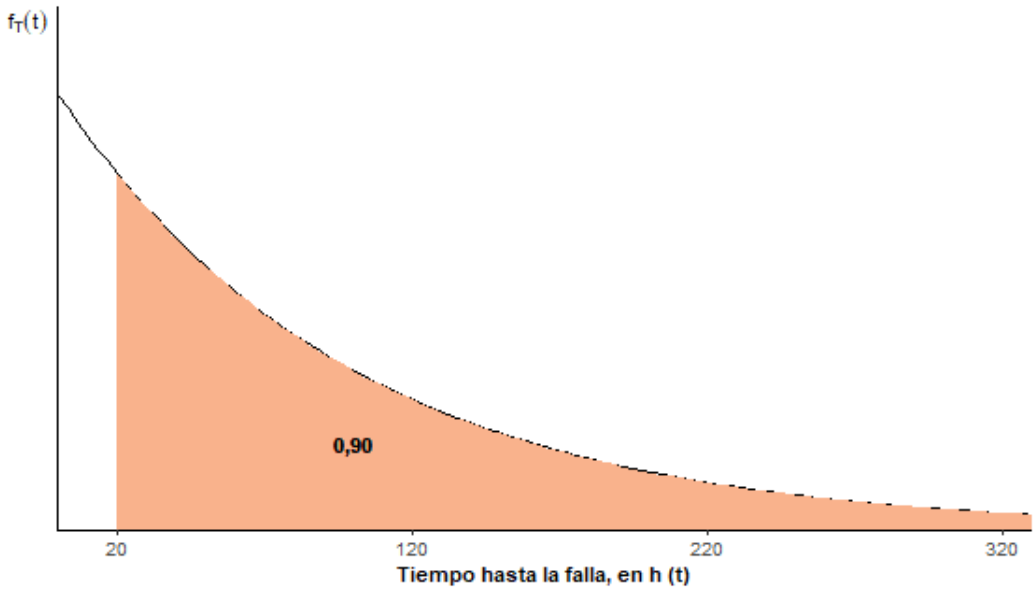
de las piezas producidas con cada una de ellas. Usted es un posible comprador de esas piezas especiales. Dadas las distribuciones de probabilidades de X e Y presentadas en este problema, plantee una situación en la cual le resulte conveniente elegir las piezas producidas por la máquina B.

43. La duración (en horas) de un tipo de componente que se utiliza para el armado de un dispositivo electrónico, D, tiene distribución Normal.
Se cuenta con el siguiente gráfico construido con `R`:



a) ¿Qué proporción de componentes tendrán una duración de al menos 110 h?
b) El dispositivo se arma con 5 componentes. Al fallar una cualquiera de las componentes deja de funcionar el dispositivo. ¿Cuál es la probabilidad de que el dispositivo dure menos de 110 h?
c) ¿Debió realizar algún supuesto para responder al ítem anterior? Justifique.

44. Se cuenta con componentes de cierto tipo, de las cuales se sabe que el tiempo (en h) hasta su falla, T, es una variable aleatoria con distribución Exponencial de media 100 h.
Se dispone de un sistema formado por n componentes de este tipo, conectadas en serie. Se supone que los tiempos de falla de las componentes son independientes entre sí.



a) Si $n = 4$, ¿cuál es la probabilidad de que el sistema funcione sin fallas después de 20 h de trabajo?

b) Indique la población asociada a la probabilidad obtenida en el ítem anterior e interprétela como una frecuencia relativa poblacional.

5.6 Estadística en R

El paquete `stats` de R, que es parte de la *base* del software, implementa numerosas funciones para la realización de cálculos asociados a distintas distribuciones de probabilidad. Entre las utilizadas más comunmente podemos citar:

Variables continuas	
Distribución	**Nombre en R**
Normal	`norm`
Uniforme	`unif`
Gamma	`gamma`
Exponencial	`exp`
Beta	`beta`
Weibull	`weibull`
Variables discretas	
Distribución	**Nombre en R**
Binomial	`binom`
Hipergeométrica	`hyper`
Binomial Negativa	`nbinom`
Poisson	`pois`

Para cada distribución, R dispone de cuatro funciones. Se puede acceder a cada una de ellas simplemente precediendo el nombre de la distribución que figura en la tabla anterior por la letra que se indica a continuación:

- `d`: devuelve el valor de la función de densidad o de probabilidad puntual en cierto valor de la variable;
- `p`: calcula la probabilidad acumulada hasta el valor especificado. Mediante la opción `lower.tail=FALSE` calcula la probabilidad anti-acumulada o a la derecha;[4]
- `q`: permite conocer el valor de la variable que acumula cierta probabilidad dada. En el caso de variables aleatorias discretas, indica el menor valor de la variable para el cual se verifica que $F_Y(y) \geq$ probabilidad dada. Con la función `q` se obtienen percentiles de interés. Mediante la opción `lower.tail=FALSE` se obtiene el valor de la variable que anti-acumula cierta probabilidad;[5]
- `r`: función para simular datos con la distribución en cuestión.

Cabe destacar que cada distribución de probabilidades requiere la especificación de los valores de sus parámetros matemáticos. A continuación se desarrollan algunos ejemplos.

[4] Con la opción mencionada, R informa el valor de $(1-F(a))$. Para variables aleatorias continuas, $1-F(a) = P(Y > a) = P(Y \geq a)$. Para variables aleatorias discretas, $1-F(a) = P(Y > a) = P(Y \geq a+1)$

[5] En el caso de variables aleatorias discretas, con la opción mencionada, R brinda el mayor valor de la variable para el cual se verifica que $(1-F_Y(y)) \geq$ probabilidad dada.

5.6.1 Distribución Normal

Si Y es una variable aleatoria con distribución Normal con media μ y desviación estándar σ, es decir, $Y \sim N(\mu, \sigma)$. Entonces:

- `dnorm(`y`,`μ`,`σ`)` devuelve el valor de la función de densidad Normal en el valor y.
  Por ejemplo, `dnorm(249.9,250,0.2)` calcula el valor que toma la función de densidad de una variable con distribución Normal con media 250 y desvío estándar 0,2 (como la longitud de las barras producidas por la empresa metalúrgica, página 216) cuando la variable vale 249,9. El resultado es 1,76.

- `pnorm(`y`,`μ`,`σ`)` calcula la probabilidad acumulada hasta y, es decir, $F_Y(y) = P(Y \leq y)$.
  Por ejemplo, `pnorm(250.142,250,0.2)` calcula la probabilidad de que una variable con distribución Normal con media 250 y desvío estándar 0,2 tome un valor menor o igual a 250,142, es decir, $F_Y(250,142) = P(Y \leq 250,142)$. El resultado es 0,7611 (Item 2.a, página 216).
  Si se desea encontrar $P(Y > 250,276)$, basta con ejecutar `1-pnorm(250.276,250,0.2)` o bien `pnorm(250.176,250,0.2, lower.tail=FALSE)` que resulta 0,0838 (Item 2.b, página 216).
  Si se quiere calcular $P(249,668 \leq Y \leq 250,08)$ se debe ejecutar `pnorm(250.08,250,0.2) - pnorm(249.668, 250,0.2)`. Esto resulta 0,6069 (Item 2.c, página 216).

- `qnorm(`p`,`μ`,`σ`)` busca el mínimo de los valores de la variable tal que la probabilidad acumulada hasta él sea mayor o igual a p, esto es, $min\{y : P(Y \leq y) \geq p\}$. En el caso de distribuciones continuas como la Normal, se obtiene el valor de la variable que acumula exactamente la probabilidad p pedida.
  Por ejemplo, `qnorm(0.95,250,0.2)` busca el valor de una variable con distribución Normal con media 250 y desvío estándar 0,2 que acumula una probabilidad de 0,95. El resultado es 250,329, ya que $F_Y(250,329) = P(Y \leq 250,329) = 0,95$.
  Si, en cambio, se quiere encontrar el valor que anti-acumula cierta probabilidad, es decir, $min\{y : P(Y \geq y) \leq p\}$, entonces se utiliza `qnorm(`p`,`μ`,`σ`, lower.tail=FALSE)`. En el caso de distribuciones continuas, se obtiene el valor que anti-acumula exactamente la probabilidad pedida.
  Por ejemplo, `qnorm(0.20,250,0.2, lower.tail=FALSE)` busca el valor de una variable con distribución Normal con media 250 y desvío estándar 0,2 tal que la probabilidad de que la variable tome un valor mayor o igual a él es 0,20. El resultado es 250,168, ya que $P(Y > 250,168) = 0,20$ (Item 3, página 216).

- `rnorm(`n`,`μ`,`σ`)` considera una variable con distribución Normal con media μ y desvío estándar σ y extrae de ella una muestra de n observaciones.
  Si se ejecuta, por ejemplo, `rnorm(100,250, 0.2)`, se obtiene un conjunto de 100 valores extraídos en forma aleatoria de una población en la cual la variable Y tiene una distribución Normal con media 250 y desvío estándar 0,2.

En todos los casos vistos anteriormente, si no se especifican los valores de los parámetros μ y σ, R considera que son iguales a 0 y 1 respectivamente, es decir, considera que se trata de la distribución Normal estándar.

En el ítem 4, página 216, interesa encontrar el valor de la variable estandarizada Z que acumula una probabilidad igual a 0,25. Es decir, el valor z^* que verifica $F_Z(z^*) = P(Z < z^*) = 0,25$. Este valor, que

es el percentil de orden 0,25 de la distribución N(0, 1), se simboliza $z_{0,25}$. Para obtenerlo se utiliza la función `qnorm(0.25)` que da como resultado el valor $-0,67$ ya que $F_Z(-0,67) = P(Z \leq -0,67) = 0,25$.

Además, el primer valor de la función puede reemplazarse por un conjunto de valores y obtener el resultado deseado para cada uno de ellos.

Así, a modo de ejemplo, si se desea obtener las probabilidades acumuladas en los valores -3, -1, 0, 1, 2 y 3 para una variable con distribución Normal Estándar, se puede ejecutar:

```
x <- seq(-3,3,1)
pnorm(x)
```

En la primera línea, se genera un conjunto de valores (*vector*) con los números del -3 al 3 en saltos de una unidad y, en la segunda línea, se piden las probabilidades acumuladas en cada uno de ellos. El resultado será el conjunto de valores: 0,0013; 0,0228; 0,1587; 0,5000; 0,8413; 0,9772 y 0,9987.

Utilizando funciones similares a las vistas junto con el paquete `ggplot2`, es posible obtener representaciones de las funciones de densidad y de probabilidad acumulada de variables con distribución Normal. El siguiente *script* permite obtener la función de densidad de una variable con distribución Normal, con promedio 10 y desvío estándar 2, y la gráfica para los valores de la variable comprendidos entre 0 y 20:

```
ggplot(data.frame(y = c(0, 20)), aes(x = y)) +
  stat_function(fun = dnorm, args = list(10, 2))+
  #Nombre de los ejes
  labs(x = "y", y = expression(f[Y](y))) +
  #Configuraciones de formato
  scale_y_continuous(expand=c(0,0), labels = scales::label_number
  (accuracy = 0.01, decimal.mark = ','))+
  #Estilo
  theme_classic()+
  #Fuente para los ejes
  theme(axis.title.x = element_text(face="bold", colour="black",
  size = 12, hjust = 1),
        axis.title.y = element_text(face="bold", colour="black",
        size = 12, angle = 0,
        vjust = 1, hjust = 0.5))
```

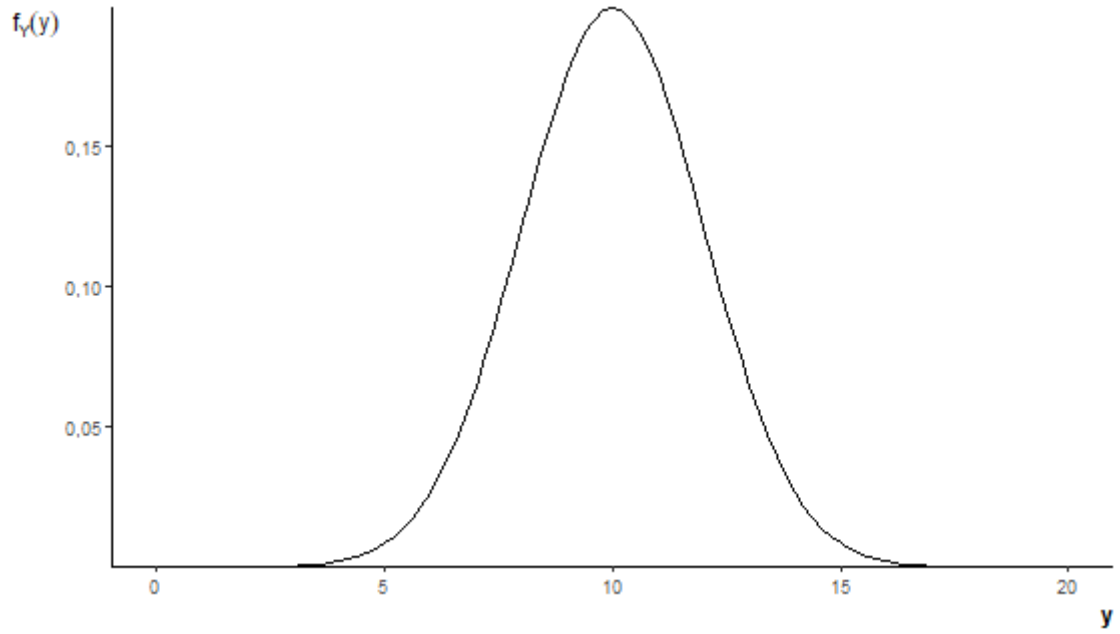


Figura 5.35 *Función de densidad de probabilidad, $f_Y(y)$, para una variable con distribución Normal con media 10 y desvío estándar 2*

Agregando al código anterior la sentencia:

```
geom_area(stat = "function", fun = dnorm, args = list(10, 2),
fill="#f9b28c", xlim = c(12,20))+
```

se obtiene el gráfico de densidad con el área bajo la curva para los valores de la variable que van de 12 en adelante (particularmente, se elige como límite superior el valor 20, dado que el área correspondiente a valores mayores es casi nula).

```
ggplot(data.frame(y = c(0, 20)), aes(x = y)) +
  stat_function(fun = dnorm, args = list(10, 2))+
  geom_area(stat = "function", fun = dnorm, args = list(10, 2),
            fill="#f9b28c", xlim = c(12,20))+
    #Nombre de los ejes
  labs(x = "y", y = expression(f[Y](y))) +
  #Configuraciones de formato
  scale_y_continuous(expand=c(0,0), labels = scales::label_number
  (accuracy = 0.01, decimal.mark = ','))+
  #Estilo
  theme_classic()+
  #Fuente para los ejes
  theme(axis.title.x = element_text(face="bold", colour="black",
   size = 12, hjust = 1),
        axis.title.y = element_text(face="bold", colour="black",
        size = 12, angle = 0,
        vjust = 1, hjust = 0.5))
```

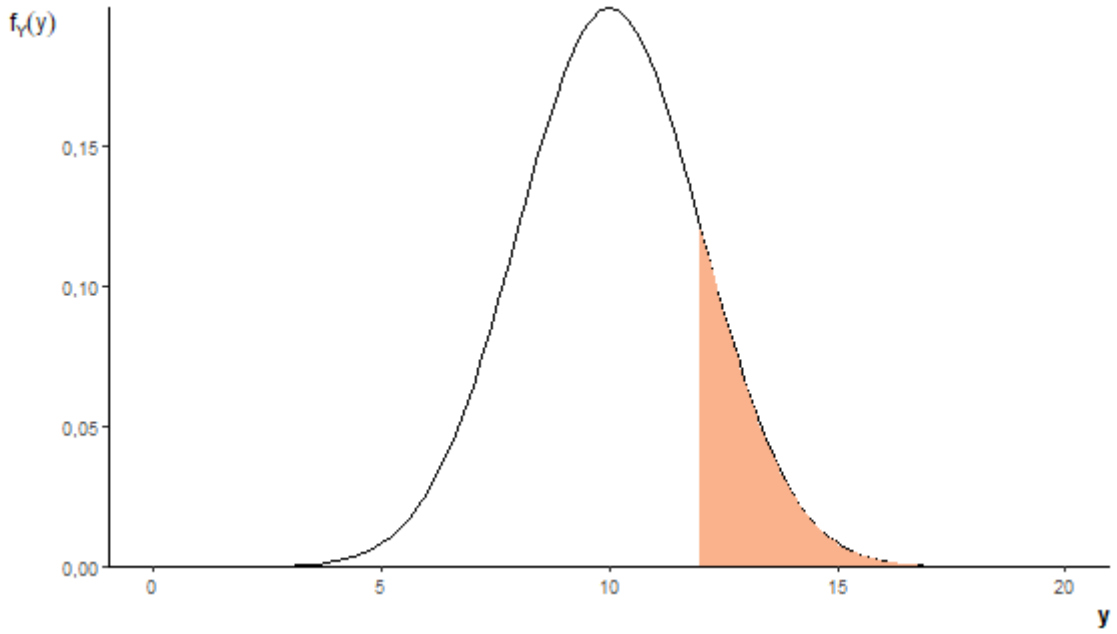


Figura 5.36 *Función de densidad de probabilidad, $f_Y(y)$, para una variable con distribución Normal con media 10 y desvío estándar 2 y área bajo la curva para valores mayores a 12*

En forma muy similar, solo reemplazando `fun = dnorm` por `fun = pnorm`, se obtiene el gráfico de la función de probabilidad acumulada. En este caso, en el título del eje de ordenadas se cambia $f_Y(y)$ por $F_Y(y)$.

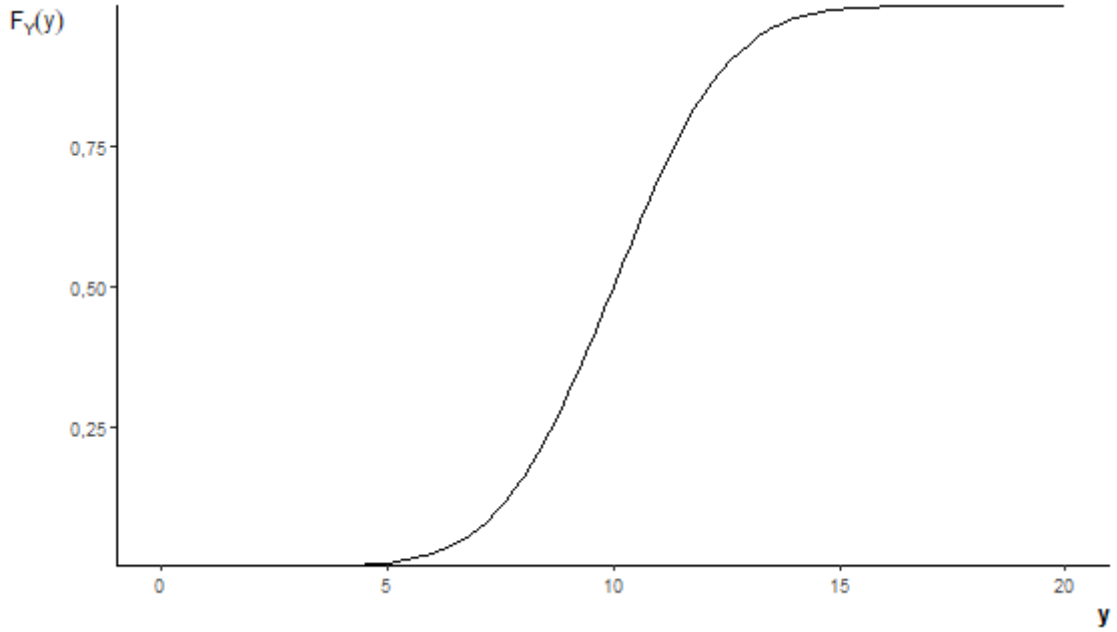


Figura 5.37 *Función de probabilidad acumulada, $F_Y(y)$, para una variable con distribución Normal con media 10 y desvío estándar 2*

5.6.2 Distribución Uniforme

Si Y una variable con distribución Uniforme en el intervalo (a, b). Entonces:

- `dunif(`y`,`a`,`b`)` devuelve el valor de la función de densidad Uniforme en el valor y.
  Por ejemplo, `dunif(250,249.655,250.345)` calcula el valor que toma la función de densidad de una variable con distribución Uniforme con mínimo en 249,655 y máximo en 250,345 (como la longitud de las barras producidas por la empresa metalúrgica, página 225) cuando la variable vale 250. El resultado es 1,45.

- `punif(`y`,`a`,`b`)` calcula la probabilidad acumulada hasta y, es decir, $F_Y(y) = P(Y \leq y)$.
  Por ejemplo, `punif(250.142,249.655,250.345)` calcula la probabilidad de que una variable con distribución Uniforme con mínimo en 249,655 y máximo en 250,345 tome un valor menor o igual a 250,142, es decir, $F_Y(250,142) = P(Y \leq 250,142)$. El resultado es 0,71 (Item 2.a, página 225).
  Si se desea encontrar $P(Y > 250,276)$, basta con ejecutar `1-punif(250.276,249.655,250.345)` o `punif(250.142,249.655,250.345, lower.tail=FALSE)`. El resultado es 0,10 (Item 2.b, página 225).
  Si se quiere calcular $P(250 \leq Y \leq 250,138)$ se debe ejecutar `punif(250.138, 249.655, 250.345) - punif(250, 249.655, 250.345)`, lo que resulta 0,20 (Item 2.c, página 225).

- `qunif(`p`,`a`,`b`)` busca el mínimo de los valores de la variable tal que la probabilidad acumulada hasta él sea mayor o igual a p, esto es, $min\{y : P(Y \leq y) \geq p\}$. En el caso de distribuciones continuas como la Uniforme, se obtiene el valor de la variable que acumula exactamente la probabilidad p pedida.
  Por ejemplo, `qunif(0.60,249.655,250.345)` busca el valor de una variable con distribución Uniforme con mínimo en 249,655 y máximo en 250,345 que acumula una probabilidad de 0,60. El resultado es 250,069 ya que $F_Y(250,069) = P(Y \leq 250,069) = 0,60$. Si, en cambio, se quiere encontrar el valor que anti-acumula cierta probabilidad, es decir, $min\{y : P(Y \geq y) \leq p\}$, entonces se utiliza `qunif(`p`,`a`,`b`, lower.tail=FALSE)`. En el caso de distribuciones continuas, se obtiene el valor que anti-acumula exactamente la probabilidad pedida.
  Por ejemplo, si se quiere encontrar aquel valor tal que la probabilidad de que la variable asuma dicho valor o uno mayor es 0,20, se puede ejecutar `qunif(0.20,249.655,250.345, lower.tail=FALSE)` que da como resultado 250,207, ya que $P(Y \geq 250,207) = 0,20$ (Item 3, página 225).

- `runif(`n`,`a`,`b`)` considera una variable con distribución Uniforme en el intervalo (a, b) y extrae de ella una muestra de n observaciones.
  Si se ejecuta, por ejemplo, `runif(100,249.655,250.345)`, se obtiene un conjunto de 100 valores extraídos en forma aleatoria de una población en la cual la variable Y tiene una distribución Uniforme con mínimo en 249,655 y máximo en 250,345.

En todos los casos vistos anteriormente, si no se especifican los valores de los parámetros a y b, R considera que son iguales a 0 y 1 respectivamente, es decir, considera que se trata de la distribución Uniforme en el intervalo $(0, 1)$.

Con una modalidad equivalente a la vista en el caso de la distribución Normal, es posible obtener las gráficas correspondientes para la distribución Uniforme.

5.6.3 Distribución Binomial

Si Y es una variable aleatoria con distribución Binomial de parámetros n (tamaño de la muestra) y π (probabilidad de que cada unidad seleccionada presente la categoría de interés), entonces:

- `dbinom(`k,n,π`)` calcula la probabilidad puntual $p_Y(k) = P(Y = k)$.
  Si $Y \sim Bi(n = 100, \pi = 0,90)$ (como la cantidad de barras que cumplen con las especificaciones, entre 100, página 242), entonces `dbinom(100,100,0.90)` presenta la probabilidad de que la variable Y tome el valor 100, es decir, $p_Y(100) = P(Y = 100)$ la que resulta igual a 0 (Item 2.a, página 242).

- `pbinom(`k,n,π`)` calcula la probabilidad acumulada hasta el valor k, $F_Y(k) = P(Y \leq k)$.
  Si $Y \sim Bi(n = 100, \pi = 0,90)$ entonces `pbinom(94,100,0.90)` presenta la probabilidad de que la variable Y tome un valor menor o igual a 94 (o menor que 95), es decir, $F_Y(94) = P(Y \leq 94) = P(Y < 95)$, la que resulta igual a 0,942 (Item 2.b, página 242).
  Haciendo `1 - pbinom(89,100,0.90)` o `pbinom(89,100,0.90, lower.tail=FALSE)` se obtiene $P(Y \geq 90)$ que resulta igual a 0,583 (Item 2.c, página 242).
  Si interesa la probabilidad de que la variable tome valores entre 95 y 100, ambos incluidos, se debe ejecutar `pbinom(100,100,0.90) - pbinom(94,100,0.90)`, lo que da como resultado 0,058 (Item 2.d, página 242).

- `qbinom(`a,n,π`)` busca el mínimo de los valores de la variable tal que la probabilidad acumulada hasta él sea mayor o igual a a, esto es, $min\{x : P(X \leq x) \geq a\}$.
  Por ejemplo, para $Y \sim Bi(n = 100, \pi = 0,90)$, `qbinom(0.30,100,0.90)` busca el menor valor de Y que acumula una probabilidad mayor o igual a 0,30, siendo este valor 89. En este caso se puede comprobar que $P(Y \leq 88) = 0,297$ y que $P(Y \leq 89) = 0,417$, por lo que se verifica que 89 es el menor valor que acumula una probabilidad mayor o igual que 0,30.
  `qbinom(`a,n,π`, lower.tail=FALSE)` busca el mínimo de los valores de la variable tal que la probabilidad anti-acumulada en él sea menor o igual a a, esto es, $min\{x : P(X > x) \leq a\}$.
  Si se ejecuta `qbinom(0.20,100,0.90, lower.tail=FALSE)` se obtiene el valor 93. En este caso se puede comprobar que $P(Y \geq 93) = 0,207$ y $P(Y \geq 94) = 0,118$, por lo que se verifica que 93 es el menor valor que anti-acumula una probabilidad menor o igual que 0,20 (Item 3, página 242).

- `rbinom(`a,n,π`)` considera una variable con distribución poblacional Binomial con parámetros n y π y extrae de ella una muestra de a observaciones.
  Si se ejecuta, por ejemplo, `rbinom(100,100,0.90)`, se obtiene un conjunto de 100 valores extraídos en forma aleatoria de una población en la cual la variable Y tiene una distribución Binomial con $n = 100$ y $\pi = 0,90$.

Para representar gráficamente la distribución de probabilidades de cualquier variable discreta, se utilizan los gráficos de bastones. Para representar la distribución acumulada, se emplean los gráficos escalonados. Para poder hacerlo mediante `ggplot2`, es necesario crear un *vector* con los valores de la variable, otro con las probabilidades puntuales y uno con las probabilidades acumuladas, y unirlos en un conjunto de datos para ser utilizados en forma similar a lo mostrado en el capítulo anterior con las tablas de frecuencias.

Por ejemplo, si se desea representar la distribución de probabilidades de una variable $Y \sim Bi(n = 10, \pi = 0{,}40)$, que tiene recorrido $R_Y = \{0, 1, 2, \ldots, 10\}$, entonces el siguiente *script* permite obtener una tabla cuya primera columna incluya los valores de este recorrido, la segunda presente las probabilidades puntuales y la tercera contenga las probabilidades acumuladas:

```
x <- seq(0,10,1)
p <- dbinom(x,10,0.40)
F <- cumsum(p)
Tabla <- cbind.data.frame(x,p,F)
```

A continuación, se utiliza la tabla obtenida para obtener la representación gráfica de las probabilidades puntuales:

```
ggplot(data=Tabla) +
  geom_segment(aes(x=x,y=0,xend=x, yend=p)) +
  geom_point(aes(x,p),size=1.5) +
  labs(x = "y", y = expression(p[Y](y))) +
  #Configuraciones de formato
  #Estilo
  theme_classic()+
  #Fuente para los ejes
  theme(axis.title.x = element_text(face="bold", colour="black",
  size = 11, hjust = 1, vjust=5),
  axis.title.y = element_text(face="bold", colour="black",
  size = 11, angle = 0, vjust = 1))+
  scale_x_continuous(expand=c(0,0), breaks = seq(0,10,1)) +
  scale_y_continuous(expand=c(0,0), limits=c(0,0.3),
        labels = scales::label_number(accuracy = 0.01, decimal.mark = ','))
```

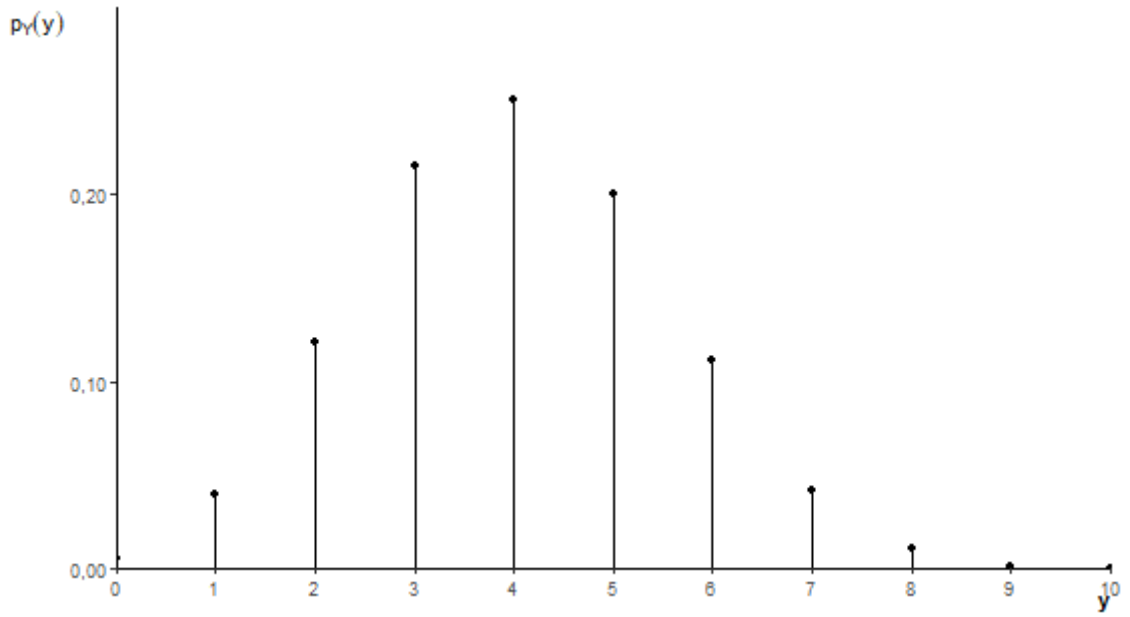


Figura 5.38 *Función de probabilidad puntual, $p_Y(y)$, para una distribución Binomial con $n = 10$ y $\pi = 0,40$*

Así mismo, es posible utilizar el conjunto de sentencias empleado en el capítulo anterior para obtener el gráfico escalonado que representa la distribución acumulada:

```
ggplot(data=Tabla) +
  geom_segment(aes(x=x,y=F, xend=x+1, yend=F)) +
  geom_segment(aes(10,1,xend=10.5,yend=1)) +
  geom_point(aes(x,F),size=1.5, shape=21, fill = "black") +
  labs(x = "y", y = expression(F[Y](y))) +
  #Configuraciones de formato
  #Estilo
  theme_classic()+
  #Configuración fuente de ejes
  theme(axis.title.x = element_text(face="bold", colour="black",
  size = 11, hjust = 1, vjust=5),
   axis.title.y = element_text(face="bold", colour="black",
   size = 11, angle = 0, vjust = 1))+ #Límites de los ejes
  scale_x_continuous(expand=c(0,0), breaks = seq(0,10,1)) +
  scale_y_continuous(expand=c(0,0), limits = c(0,1.05),
                     labels = scales::label_number(accuracy = 0.01,
                                                   decimal.mark = ','))
```

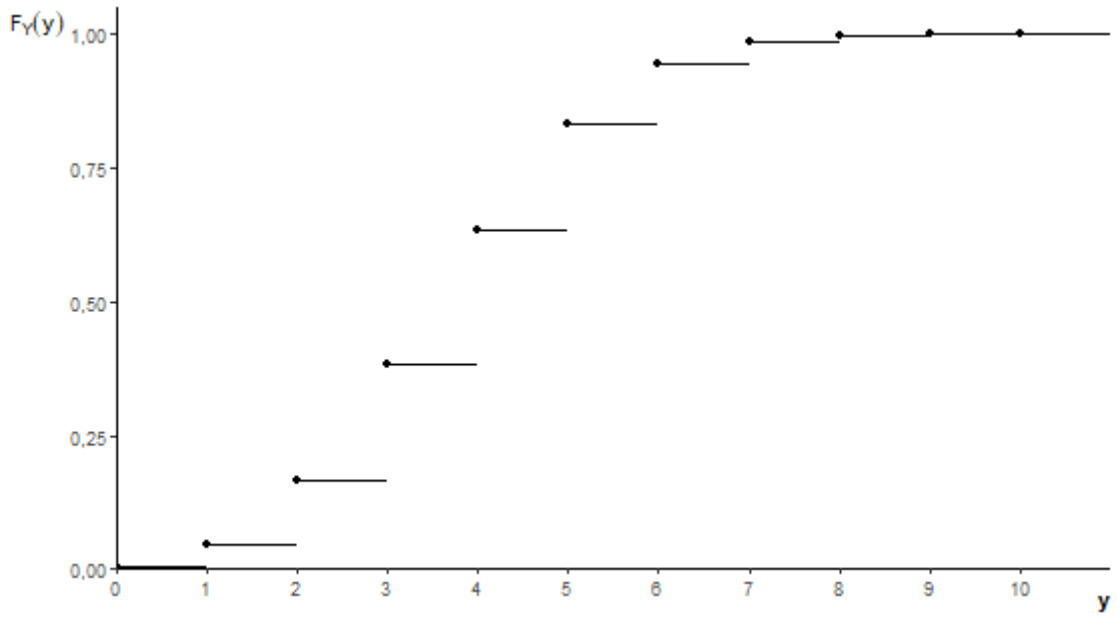


Figura 5.39 *Función de probabilidad acumulada, $F_Y(y)$, para una distribución Binomial con $n = 10$ y $\pi = 0,40$*

5.6.4 Otras distribuciones

Con criterios similares a los vistos para las distribuciones Normal, Uniforme y Binomial, es posible trabajar con otras distribuciones estudiadas en este libro. Para eso, es importante conocer cuáles son los parámetros que se deben especificar en cada caso.

- **Distribución Exponencial**: se debe especificar el parámetro α (*rate*), correspondiente a la inversa de la esperanza.
  Por ejemplo, si Y es una variable aleatoria con distribución Exponencial con $E(Y) = 500$, entonces $\alpha = 1/500 = 0,002$ y para calcular $F_Y(100) = P(Y \leq 100)$ se ejecuta `pexp(100,0.002)`. Esto resulta 0,1813 (Item 2.a, página 235).

- **Distribución Hipergeométrica**: se debe mencionar el número de unidades en la población que presentan cierta característica de interés (N_E), el número de unidades de la población que no presentan esa característica ($N - N_E$) y el tamaño de la muestra (n)
  Si, por ejemplo, se extraen muestras aleatorias sin reposición de tamaño 5 de una población de 20 elementos entre los cuales hay 4 que poseen cierta característica de interés, entonces, la variable X: cantidad de elementos que presentan la característica de interés en muestras de tamaño 5 extraídas de una población finita como la descripta, sigue una distribución Hipergeométrica: $X \sim Hipergeom(N = 20, N_E = 4, n = 5)$. En este contexto, `dhyper(2,4,16,5)` calcula la probabilidad de que la variable tome el valor 2, siendo esta igual a $p_X(2) = P(X = 2) = 0,2167$.
- **Distribución Poisson**: se debe indicar el valor del parámetro α.
  Por ejemplo, si $X \sim Po(0,20)$, ejecutando `dpois(2,0.20)` se encuentra $p_X(2) = P(X = 2)$ que en este caso es, aproximadamente, 0,0163.

- **Distribuciones no especificadas**: es posible definir mediante *funciones* otras distribuciones de probabilidad no incluidas en `R`. Una vez definida la función de densidad, se puede integrar a fin de obtener probabilidades. Así mismo, se la puede graficar utilizando `ggplot2`.

Por ejemplo, si se está trabajando con una variable Y con distribución Triangular simétrica entre 249 y 251, su función de densidad se puede expresar como:

```
triangular <- function(x) {
ifelse(x < 249 | x > 251,0,
ifelse(x >= 249 & x <= 250, x-249, 251-x))
}
```

Si se pretende obtener $P(249,6 < Y < 250)$, se podrá ejecutar:

```
integrate(triangular, lower = 249.6, upper = 250)
```

Esto resulta igual a 0,32. Para realizar la gráfica de la función de densidad, se trabaja en forma similar a lo visto para otras distribuciones continuas, indicando en `stat_function(fun= )` el nombre de la función creada. Por ejemplo:

```
ggplot(data.frame(x = c(248.5, 251.5)), aes(x = x)) +
  stat_function(fun = triangular)+
  #Nombre de los ejes
  labs(x = "y", y = expression(f[Y](y))) +
  #Configuraciones de formato
  #Estilo
  theme_classic()+
  #Fuente para los ejes
  theme(axis.title.x = element_text(face="bold", colour="black", size = 11,
  hjust = 1, vjust=5),
  axis.title.y = element_text(face="bold", colour="black", size = 11,
  angle = 0, vjust = 1))+ #Límites de los ejes
  scale_y_continuous(expand=c(0,0),
                     labels = scales::label_number(accuracy = 0.01,
                                                   decimal.mark = ',')) +
  scale_x_continuous(expand=c(0,0),
                     labels = scales::label_number(accuracy = 1,
                                                   decimal.mark = ','))
```

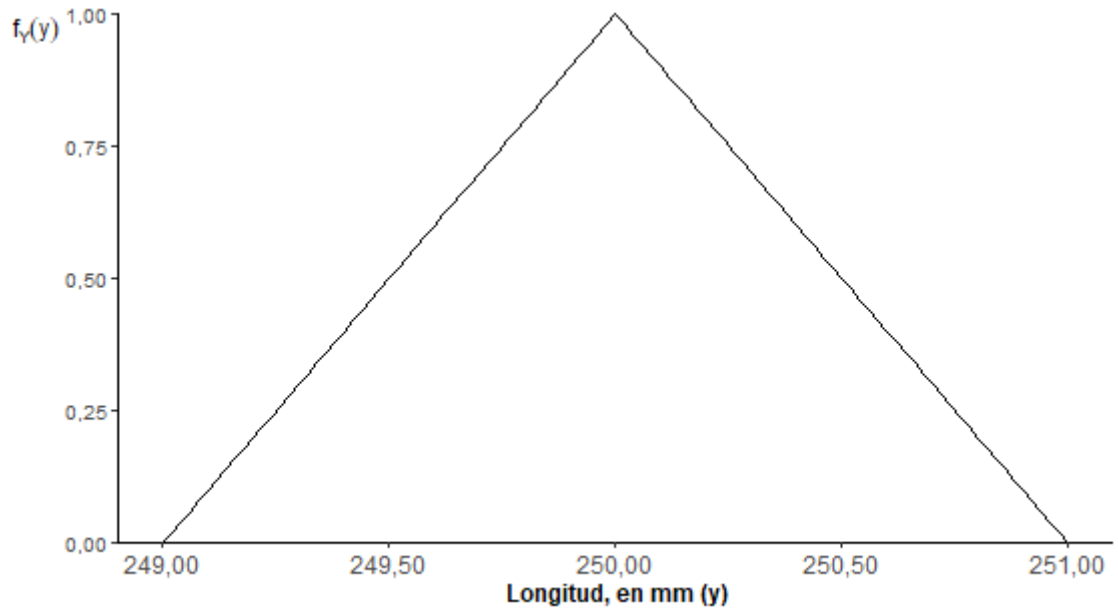


Figura 5.40 *Función de densidad de probabilidad, $f_Y(y)$, para la distribución Triangular simétrica con mínimo en 249 y máximo en 251*

5.7 Apéndice

Uso de la aplicación `Probability Distributions`

Esta aplicación, desarrollada por Matthew Bognar del Departamento de Estadística y Ciencia Actuarial de la Universidad de Iowa (Estados Unidos), permite obtener probabilidades, así como valores de la variable que verifican cierta probabilidad, para un conjunto de distribuciones de probabilidades discretas y continuas.

La aplicación `Probability Distributions` está disponible como aplicación móvil para iOS (iPhone) y Android, por lo que puede instalarse directamente en teléfonos y tabletas. En ambos casos, el nombre de la aplicación es Probability-Distributions.

- En iPhone/iOS, está lista para descargar gratuitamente en la App Store.
- En Android, también se puede descargar gratuitamente desde la Google Play Store.

Una vez definido el modelo y sus parámetros matemáticos, la aplicación muestra automáticamente el gráfico de la función de densidad de probabilidad ($f_Y(y)$) o de la función de probabilidad puntual ($p_Y(y)$). Cabe destacar que, en el caso discreto, las probabilidades puntuales se representan mediante barras y no bastones. A continuación, el usuario puede definir un valor de la variable o una probabilidad.

- Si define un valor y, la aplicación le permite obtener:
  - $P(Y = y)$ (solo para variables discretas);
  - $P(Y \leq y) = F_Y(y)$;
  - $P(Y \geq y)$.

  En todos los casos, la probabilidad calculada se muestra también de forma gráfica.
- Si define un valor para una probabilidad, ya sea $P(Y \leq y)$ o $P(Y \geq y)$, la aplicación le brinda el valor y que lo verifica.
  - Cuando la probabilidad se define como $P(Y \leq y)$, el valor y obtenido corresponde a un percentil.
  - En el caso de variables discretas, puede no existir un y que verifique exactamente la probabilidad requerida. En estas situaciones, la aplicación proporciona los dos valores y correspondientes a probabilidades que acotan, inferior y superiormente, la probabilidad solicitada.

Adicionalmente, la aplicación brinda información de interés para cada modelo, como las expresiones de la función de densidad de probabilidad (o de probabilidad puntual), de la media y de la variancia.

A modo de ejemplo se utiliza la aplicación para obtener las probabilidades pedidas en la página 216, correspondientes a la distribución Normal.

En la Figura 5.41 se muestran las capturas de pantalla utilizadas para obtener las probabilidades pedidas en los ítems 2.a, 2.b y 2.c; mientras que en la Figura 5.42 se muestran las capturas de pantalla utilizadas para obtener los valores de la variable pedidos en los ítems 3 y 4. En este último ítem no se conoce el valor de la desviación estándar, σ, y por eso se trabaja con la variable estandarizada.

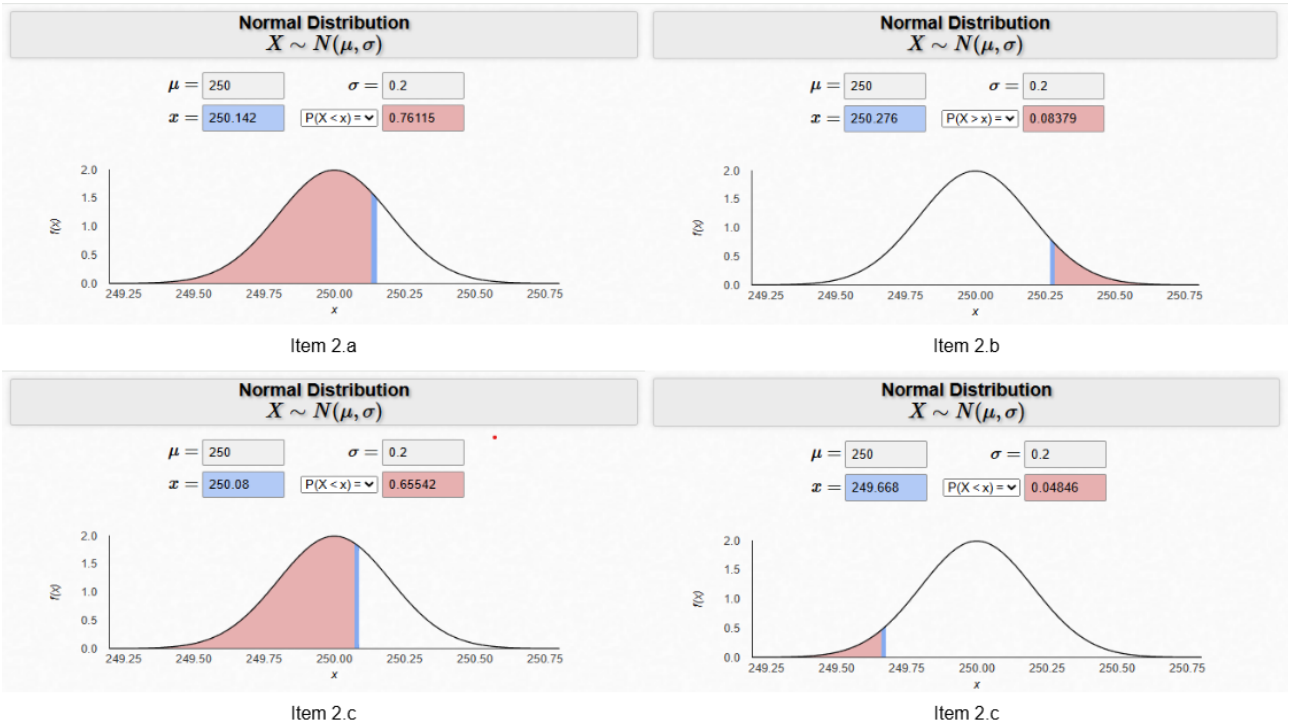


Figura 5.41 *Capturas de pantalla de la aplicación "Probability Distributions" para obtener las probabilidades correspondientes a los ítems 2.a, 2.b y 2.c (página 216) suponiendo* $Y \sim N(250; 0,2)$

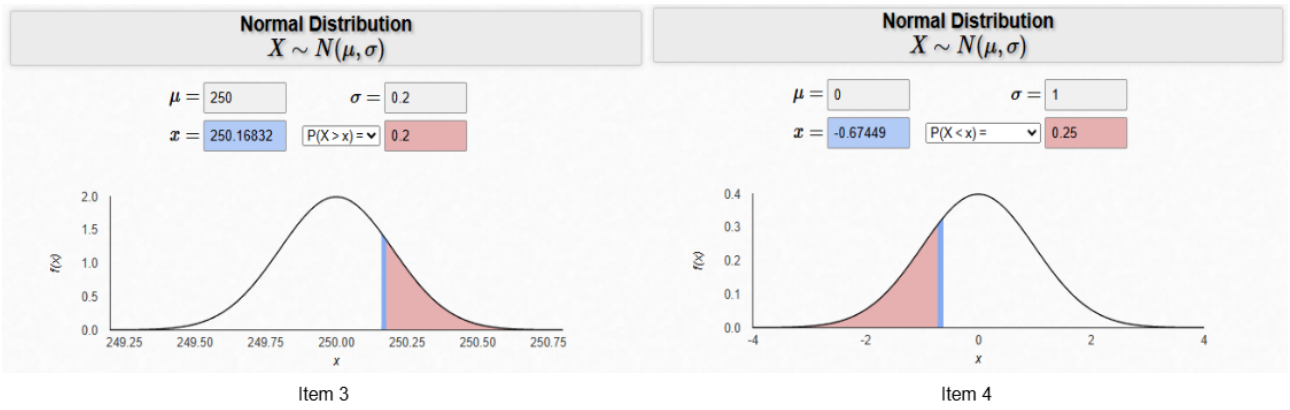


Figura 5.42 *Capturas de pantalla de la aplicación "Probability Distributions" para obtener los valores de la variables correspondientes a los ítems 3 (suponiendo* $Y \sim N(250; 0,2)$*) y 4 (suponiendo* $Y \sim N(250; \sigma$ *no conocido)) (página 216)*



6. Funciones de un vector aleatorio

6.1 Introducción

En los Capítulos 3 y 5 se han tratado problemas en cuya resolución se involucra el modelo de comportamiento de una única variable aleatoria discreta o continua. Muy frecuentemente, sin embargo, la resolución de un problema implica el estudio simultáneo de un conjunto de variables aleatorias o "vector aleatorio". Puede interesar estudiar el modelo de comportamiento conjunto y conocer la relación entre las variables del vector.

En este libro no se tratan problemas multivariados, pero sí se abordan situaciones donde el objeto de estudio es una nueva variable, función de las variables del vector. En este caso, el conocimiento del comportamiento en probabilidad de la nueva variable se puede deducir del comportamiento de las variables aleatorias que integran dicho vector.

En el presente capítulo se propone hacer una breve introducción al estudio de los vectores aleatorios, sus distribuciones de probabilidades y la relación entre las variables que lo componen, para luego centrarse fundamentalmente en propiedades útiles para el conocimiento de la distribución de probabilidades de variables aleatorias unidimensionales que son funciones de las variables aleatorias del vector.

El objetivo general de este capítulo es proporcionar algunos resultados que permitan abordar problemas de Ingeniería en los cuales la variable de interés es función de las variables de un vector aleatorio. Los objetivos específicos de este capítulo son:

- Definir el vector aleatorio, clasificarlo según el número y el tipo de variables que lo integran y presentar básicamente su comportamiento en probabilidad.
- Definir variables aleatorias como funciones de las variables de un vector aleatorio.
- Estudiar algunas propiedades que permiten conocer el modelo probabilístico de variables que se pueden definir como combinaciones lineales de las variables del vector.
- Aplicar algunas herramientas de simulación para obtener probabilidades y otros parámetros especialmente para el caso de variables aleatorias que se definen como funciones no lineales de las variables del vector.

6.2 Vector aleatorio y su comportamiento en probabilidad

Cuando interesa estudiar simultáneamente dos o más variables aleatorias en las unidades que conforman una población determinada, se define un vector de variables aleatorias o vector aleatorio $(Y_1, Y_2, \ldots, Y_k)$. El recorrido de este vector está constituido por todas las combinaciones posibles de valores de sus componentes; por lo tanto, $R_{(Y_1,Y_2,\ldots,Y_k)} = R_{Y_1} \times R_{Y_2} \times \cdots \times R_{Y_k}$.

Clasificación de los vectores aleatorios:

- Según el tipo de variables, los vectores aleatorios pueden ser discretos, si todas las variables aleatorias son discretas; continuos, si todas son continuas o mixtos [1] si incluyen algunas variables aleatorias discretas y otras continuas.
- Según el número de variables aleatorias que incluya el vector, se tienen vectores aleatorios bivariados (2 variables aleatorias) o multivariados (3 o más variables aleatorias).



Problema 1 (pág. 7): *Considerando las barras de acero producidas en la empresa metalúrgica (Problema 1, Sección 1.4), se definen a continuación tres vectores aleatorios.*

1. *Se registra en cada barra de acero, la cantidad de poros (W_1) y la cantidad de manchas (W_2). El vector aleatorio es (W_1, W_2).*
   Si $R_{W_1} = \{0,1,2,3\}$ y $R_{W_2} = \{0,1\}$, el recorrido del vector aleatorio (W_1, W_2) resulta entonces ser $R_{(W_1,W_2)} = \{(0,0),(0,1),(1,0),(1,1),(2,0),(2,1),(3,0),(3,1)\}$.
2. *Se registra de cada barra de acero el diámetro de la sección transversal, (X) y la longitud, (Y), ambas medidas en mm. El vector aleatorio es (X, Y).*
   Si $R_X = \{x : 0 < x < 10\}$ y $R_Y = \{y : 0 < y < 500\}$, el recorrido del vector aleatorio (X, Y) resulta entonces ser $R_{(X,Y)} = \{(x, y) : 0 < x < 10,\ 0 < y < 500\}$.
3. *En la empresa comercializan las barras en grupos de 10 unidades. La longitud de cada barra del grupo, en mm, es una variable aleatoria Y_i y el vector aleatorio que da información sobre las longitudes de todas las barras de cada grupo es $(Y_1, Y_2, Y_3, \ldots, Y_{10})$. $R_{Y_1} = R_{Y_2} = \cdots = R_{Y_{10}} = \{y : 0 < y < 500\}$.*
   Entonces, el recorrido del vector aleatorio $(Y_1, Y_2, Y_3, \ldots, Y_{10})$ resulta:
   $R_{(Y_1,Y_2,Y_3,\ldots,Y_{10})} = \{(y_1, y_2, \ldots, y_{10}) : 0 < y_i < 500,\ i = 1, \ldots, 10\}$.

*Observe que en el primer vector se consideran dos variables aleatorias discretas: se trata de un **vector aleatorio bivariado discreto** que asume un número finito de pares ordenados [2]. En el segundo, las dos variables aleatorias son continuas: se trata de un **vector aleatorio bivariado continuo** y en el tercero, se consideran diez variables aleatorias continuas. Se trata de un **vector aleatorio multivariado continuo** que puede asumir cualquier k-upla de valores en la región definida por el recorrido. En este último caso, además, las diez variables tienen la misma distribución de probabilidades puesto que las barras que integran cada grupo pertenecen a la misma población.*

[2] Como ya se mencionó en el Capítulo 3, las variables aleatorias discretas también pueden asumir un número infinito numerable de valores, de modo que los vectores aleatorios bivariados discretos tendrían en ese caso un número infinito numerable de pares ordenados.

Si para cada unidad de la población se registran simultáneamente los valores de varias variables aleatorias es porque interesa observar el comportamiento conjunto de estas variables. Surge así la distribución de probabilidad conjunta, de la que se desprenden distribuciones marginales y condicionales.

[1] En este libro no se estudia el comportamiento en probabilidad de vectores aleatorios mixtos.

La distribución conjunta de un vector aleatorio discreto se denomina distribución de probabilidad puntual conjunta y la de un vector aleatorio continuo, se denomina función de densidad de probabilidad conjunta.



En el Apéndice (Sección 6.8.1) se definen y ejemplifican ambas distribuciones (así como las distribuciones marginales y condicionales correspondientes) para el caso de vectores bivariados discretos o continuos. En este libro no se presentan modelos bivariados ni multivariados de uso frecuente (como la distribución normal multivariada por ejemplo).

A modo de ejemplo, a continuación se presenta la distribución de probabilidad puntual conjunta del primer vector bivariado definido para el proceso de producción de barras de acero (Sección 6.2).



Problema 1 (pág. 7): *En este caso, interesa estudiar el comportamiento de la cantidad de poros (W_1) y de la cantidad de manchas (W_2) en forma simultánea. Esto da origen al vector aleatorio bivariado (W_1, W_2): (Cantidad de poros, cantidad de manchas). La población está conformada por las infinitas barras producidas por el proceso y la población estadística por los infinitos pares ordenados de observaciones (w_1, w_2).*

Tabla 6.1 *Distribución de probabilidades conjunta de la cantidad de poros (W_1) y de la cantidad de manchas (W_2) por barra*

W_2	W_1				Total
	0	**1**	**2**	**3**	
0	0,3577	0,2628	0,0876	0,0219	**0,73**
1	**0,1323**	0,0972	0,0324	0,0081	0,27
Total	0,49	**0,36**	0,12	0,03	1

$P(W_1 = 0, W_2 = 1) = p_{W_1,W_2}(0,1) = 0,1323$
Pensando en una gran cantidad de barras producidas, aproximadamente el 13,23 % de ellas no presenta poros y presenta una sola mancha. Si se selecciona al azar una barra del proceso, la chance de que esta no presente poros y presente una sola mancha es 0,1323.

$P(W_1 = 1) = p_{W_1}(1) = 0,36$
Pensando en una gran cantidad de barras producidas, aproximadamente el 36 % de ellas presenta un solo poro. Si se selecciona al azar una barra del proceso, la chance de que la misma presente un solo poro es 0,36.

$P(W_2 = 0) = p_{W_2}(0) = 0,73$
Pensando en una gran cantidad de barras producidas, aproximadamente el 73 % de ellas no presenta manchas. Si se selecciona al azar una barra del proceso, la chance de que la misma no presente manchas es 0,73.

Las dos últimas probabilidades interpretadas corresponden a las distribuciones marginales de las variables W_1 y W_2 respectivamente, es decir a las distribuciones de probabilidad de cada una de esas variables consideradas individualmente.

6.2.1 Sobre la relación entre las variables del vector

Un punto importante a tener en cuenta cuando se trabaja con un vector aleatorio es la relación que pueden tener las variables del vector. Si no tienen ningún tipo de relación, se dice que estas son independientes entre sí. Si tienen algún tipo de relación, se podrá analizar de qué tipo es y su intensidad.

¿Cómo se entiende el concepto de independencia?

En el Capítulo 4 se vio que el hecho de que dos sucesos A y B sean independientes indica que la ocurrencia de uno de ellos no modifica la probabilidad de ocurrencia del otro. Llevado a variables aleatorias, que dos variables sean independientes significa que el comportamiento en probabilidad de una de ellas no depende del valor que tome la otra, o, dicho de otra manera, la probabilidad de que una de las variables tome un valor cualquiera de su recorrido no se modifica al variar el valor que toma la otra.

Por ejemplo, si se consideran infinitas piezas metálicas de diferentes espesores fabricadas con cierta aleación, y se define el vector (X_1, X_2) donde X_1: espesor y X_2: resistencia a la rotura, es razonable pensar que ambas variables no son independientes, es decir, que la resistencia a la rotura de las piezas depende de su espesor. Por ejemplo, si las piezas tienen un espesor alto, es más probable que su resistencia a la rotura sea mayor y si tienen un espesor pequeño, es más probable que su resistencia a la rotura sea menor.

Si, en cambio, se consideran infinitas piezas plásticas almacenadas en el depósito y se define el vector (X_1, X_2) donde X_1: diámetro y X_2: cantidad de días de almacenamiento, es razonable pensar que ambas variables son independientes, es decir, que el comportamiento del diámetro no se modifica con el tiempo de almacenamiento.



En el Apéndice (Sección 6.8.2) se presentan las definiciones formales de independencia de variables, derivadas de las definiciones de independencia de sucesos (Capítulo 4) y se analiza la independencia de las variables correspondientes a los dos primeros vectores presentados al inicio de esta sección. Cuando las variables no son independientes se comenta cómo evaluar el sentido y la intensidad de la relación entre ambas.

6.3 Variables definidas como funciones de un vector aleatorio

A continuación se definen algunas situaciones relacionadas con los vectores aleatorios presentados al inicio de la Sección 6.2 y relativos a las barras producidas por la empresa metalúrgica.



Problema 1 (pág. 7):

1. *Se registra en cada barra de acero, la cantidad de poros (W_1) y la cantidad de manchas (W_2). El vector aleatorio es (W_1, W_2).*
   - *Suponga ahora que cierto cliente está preocupado por la cantidad total de poros y manchas que las barras presentan (T). Esta es una nueva variable aleatoria que puede definirse en función de las variables del vector: $T = W_1 + W_2$.*

- *Suponga que el costo (en $) de eliminar cada poro que se presenta en las barras es de $350 y que el costo de eliminar cada mancha es de $200. ¿Cómo se define una nueva variable que dé información sobre el costo de eliminar los poros y las manchas de cada barra? Ese costo (C) se define de la siguiente manera: $C = 350W_1 + 200W_2$, es decir, se expresa en función de W_1 y W_2.*

2. *Se registra de cada barra de acero el diámetro de la sección transversal, (X) y la longitud, (Y), ambas medidas en mm. El vector aleatorio es (X, Y).*
   - *Una de las empresas que compran barras de acero en la metalúrgica requiere que las mismas cumplan con ciertas especificaciones relativas al volumen (en mm^3). Esta variable (J) puede definirse en función de las variables del vector: $J = \frac{1}{4}\pi X^2 Y$.*
   - *Se cuenta con unos soportes metálicos que tienen una perforación circular, de la cual se registra su diámetro, en mm (M). Para saber si la barra entra en el soporte, se deben comparar ambos diámetros. Esto puede hacerse a través de la variable $H = M - X$. Si esta nueva variable asume valor negativo, la barra no entra en el soporte.*
3. *En la empresa comercializan las barras en grupos de 10 unidades. La longitud de cada barra del grupo, en mm, es una variable aleatoria Y_i y el vector aleatorio que da información sobre las longitudes de todas las barras de cada grupo es $(Y_1, Y_2, Y_3, \ldots, Y_{10})$.*
   - *Una arquitecta está interesada en la longitud total de las barras de cada grupo (L). Esta nueva variable puede definirse de la siguiente manera: $L = Y_1 + Y_2 + Y_3 + \cdots + Y_{10}$ representa la longitud total de las barras.*
   - *Como parte del control del proceso de producción de las barras se realiza lo siguiente: se selecciona cada hora un conjunto de 10 barras y se obtiene la longitud promedio del mismo. Si este valor es menor que 249,8 mm o mayor que 250,2 mm, se revisan todas las barras producidas en esa hora. La variable de interés en este caso es $M = \frac{L}{10} = \frac{1}{10}(Y_1 + Y_2 + Y_3 + \cdots + Y_{10})$ que representa la longitud promedio de las barras de cada grupo.*

En las situaciones presentadas se cuenta con un vector aleatorio $(Y_1, Y_2, Y_3, \ldots, Y_k)$ y a partir de él surge una nueva variable aleatoria que resulta función del vector. Podría ocurrir el caso inverso, es decir, contar con una variable y considerarla como función de un vector asociado a ella. Por ejemplo, el peso (P) de bolsas que contienen 50 tornillos especiales se puede considerar de la siguiente manera: $P = T_1 + T_2 + \cdots + T_{50} + Z$ donde T_i es el peso del tornillo i, con $i = 1, 2, \ldots, 50$ y Z es el peso del envase. El vector aleatorio correspondiente a cada bolsa es $(T_1, T_2, \ldots, T_{50}, Z)$.



Con lo visto hasta ahora se pueden resolver las Actividades 1 y 2, Sección 6.6.



Cualquiera sea el caso, ¿cómo se obtienen la distribución de probabilidades y los parámetros de estas nuevas variables, funciones del vector aleatorio?

Para dar respuesta a lo planteado[3] se tomará como ejemplo la situación relativa al total de poros y manchas de las barras de acero.

[3]En este texto solo se va a responder a estas preguntas de manera analítica para el caso de variables que resulten funciones lineales del vector aleatorio. Para el caso de otras funciones, como por ejemplo, el del volumen de las barras se pueden obtener probabilidades y valores de otros parámetros de interés aplicando herramientas de simulación.

Ejemplo 6.1 Como se mencionó al inicio de la presente sección (página 6.3), en las barras de acero se registra la cantidad de poros (W_1) y la cantidad de manchas (W_2). Suponga ahora que el cliente, preocupado por la cantidad total de poros y manchas que las barras presentan (T), solo hará una compra si la proporción de barras con más de 2 de estos defectos es menor que 0,10. A partir de la distribución de probabilidades conjunta presentada en la Tabla 6.1 (página 299), se puede construir la distribución de probabilidades de T y analizar si se puede satisfacer el requerimiento de este cliente.

Observe que W_1 toma valores enteros entre 0 y 3 y W_2 entre 0 y 1, de modo que la variable T puede tomar valores enteros entre 0 y 4, con las probabilidades que se detallan en la Tabla 6.2. Para obtener la probabilidad de cada valor posible de la variable T, que en este caso es una suma de dos variables aleatorias, se consideran todos los pares (w_1, w_2) del vector cuyos valores, sumados, dan como resultado ese valor.

Tabla 6.2 *Construcción de la distribución de probabilidades de la cantidad total de defectos (poros o manchas) (T) por barra a partir de la distribución conjunta de la cantidad de poros (W_1) y de la cantidad de manchas (W_2)*

T **vale:**	**cuando:**	**Cálculo de las probabilidades**	$p_T(t)$
0	W_1 y W_2, ambas toman valor 0	$p_T(0) = P(T=0) = P(W_1=0,\ W_2=0)$	0,3577
1	W_1 toma el valor 1 y W_2 el valor 0 o viceversa	$p_T(1) = P(T=1) = P(W_1=1,\ W_2=0) +$ $+ \mathrm{P}(W_1=0,\ W_2=1) = 0,2628 + 0,1323$	0,3951
2	W_1 toma el valor 2 y W_2 el valor 0 o cuando W_1 y W_2 ambas toman valor 1	$p_T(2) = P(T=2) = P(W_1=2,\ W_2=0) +$ $+ \mathrm{P}(W_1=1,\ W_2=1) = 0,0876 + 0,0972$	0,1848
3	W_1 toma el valor 3 y W_2 el valor 0 o cuando W_1 toma el valor 2 y W_2 el valor 1	$p_T(3) = P(T=3) = P(W_1=3,\ W_2=0) +$ $+ \mathrm{P}(W_1=2,\ W_2=1) = 0,0219 + 0,0324$	0,0543
4	W_1 toma el valor 3 y W_2 el valor 1	$p_T(4) = P(T=4) = P(W_1=3,\ W_2=1) =$ $= 0,0081$	0,0081
Total:			1

En la primera y en la última columna de la Tabla 6.2 se muestra la distribución de probabilidades de la variable T, que se representa gráficamente en la Figura 6.1. A continuación de la figura se presenta el cálculo del promedio, la variancia y la desviación estándar de esta variable.

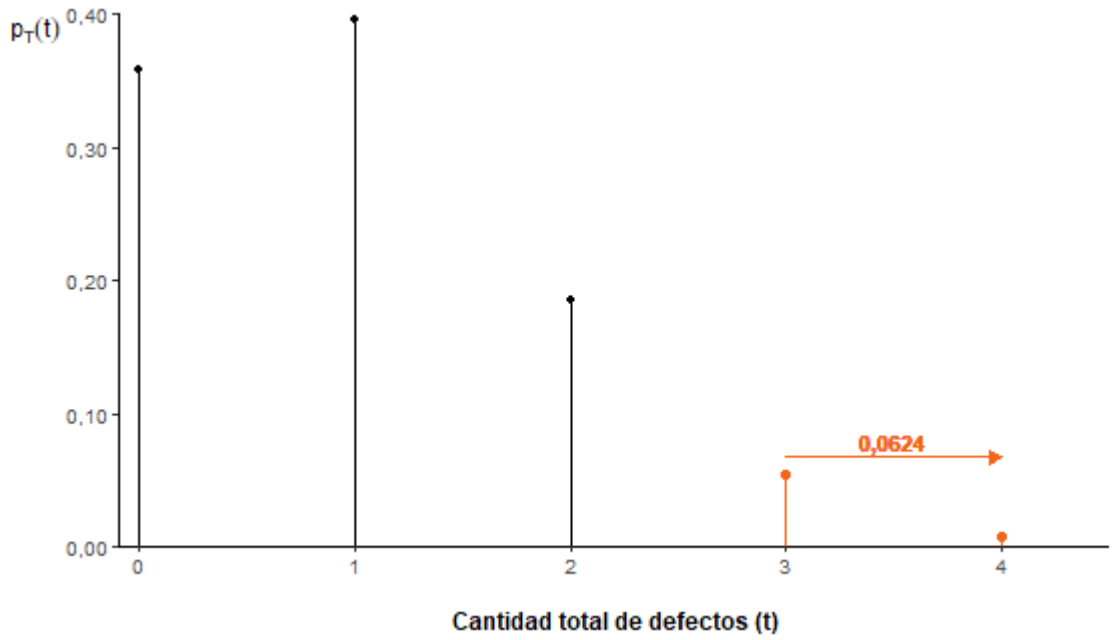


Figura 6.1 *Función de probabilidad puntual de T, cantidad total de defectos (poros o manchas) por barra. La suma de las alturas de los bastones en naranja indica la probabilidad de que una barra presente más de dos de estos defectos*

$$E(T) = 0 \cdot 0,3577 + 1 \cdot 0,3951 + 2 \cdot 0,1848 + 3 \cdot 0,0543 + 4 \cdot 0,0081 = 0,96$$

$$V(T) = E(T^2) - \left[E(T)^2\right] = \left(0^2 \cdot 0,3577 + 1^2 \cdot 0,3951 + 2^2 \cdot 0,1848 + 3^2 \cdot 0,0543 + 4^2 \cdot 0,0081\right) - 0,962 = 1,7526 - 0,9216 = 0,831$$

$$D(T) = \sqrt{0,831} = 0,911$$

De la Tabla 6.2 se puede obtener la proporción de barras con más de dos defectos (poros o manchas):

$$P(T > 2) = P(T = 3) + P(T = 4) = 0,0543 + 0,0081 = 0,0624$$

Como esta resulta ser menor que 0,10, el cliente no debería tener inconvenientes y podría comprar las barras en la metalúrgica.

Repasando todo lo actuado puede decirse que se definió una nueva variable T como suma de las dos variables del vector y se pudo construir, de manera "artesanal", su distribución de probabilidad y a partir de ella, obtener los valores de algunos parámetros. En este caso pudo hacerse de esta manera porque las dos variables del vector podían asumir un número muy reducido de valores, pero esto no es lo que ocurre en la mayoría de los casos.

Para un caso general, dada una nueva variable aleatoria G, función del vector aleatorio $(Y_1; Y_2; \ldots; Y_k)$ se requieren propiedades que faciliten la tarea de obtener su comportamiento en probabilidad. En la Sección 6.3.1 se presentan las propiedades que permiten obtener los valores de $E(G)$ y $V(G)$;

mientras que en la Sección 6.3.2 se presentan las que involucran a las distribuciones de las variables del vector, permitiendo determinar, o al menos aproximar, una distribución de probabilidades para la nueva variable. En ambas secciones se consideran variables que pueden definirse como funciones lineales de un vector de variables independientes entre sí; aunque se dan algunos lineamientos para obtener probabilidades y otros parámetros a través de simulación cuando la función G no es lineal.

6.3.1 Esperanza y desviación estándar de la nueva variable

Se define una nueva variable G que resulta una combinación lineal de las variables del vector aleatorio: $G = a_1 Y_1 + a_2 Y_2 + \cdots + a_k Y_k$ con $a_i \in R$, para $i = 1, 2, \ldots, k$.

1. Esperanza de la nueva variable G

Dada la variable G, puede demostrarse que $E(G) = a_1 E(Y_1) + a_2 E(Y_2) + \cdots + a_k E(Y_k)$ (Apéndice, Sección 6.3.1).

En otras palabras, la misma función que vincula a G con las variables del vector aleatorio, también vincula a $E(G)$ con sus esperanzas. En la Tabla 6.3 se muestra cómo obtener $E(G)$ para casos particulares de una combinación lineal de variables aleatorias.

Tabla 6.3 *$E(G)$ para algunos casos particulares de una combinación lineal de variables aleatorias*

Casos particulares	G	E(G)
G es una suma de variables aleatorias (es decir, $a_i = 1\ \forall i$)	$G = Y_1 + Y_2 + \cdots + Y_k$	$E(Y_1) + E(Y_2) + \cdots + E(Y_k)$
G es una suma de variables aleatorias igualmente distribuidas (es decir, $a_i = 1\ \forall i$ y $E(Y_i) = E(Y)\ \forall i$)	$G = Y_1 + Y_2 + \cdots + Y_k$ Y_i idénticamente distribuidas	$E(Y_1) + \cdots + E(Y_k) = kE(Y)$
G es una resta de variables aleatorias (es decir, $a_1 = 1$ y $a_2 = -1$)	$G = Y_1 - Y_2$	$E(Y_1) - E(Y_2)$

2. Desviación estándar de la nueva variable G

En primer lugar se halla la variancia de la nueva variable y luego, su desviación estándar. Para hallar la variancia de G es importante conocer si las variables son independientes o no.

Para los puntos siguientes (determinación de $V(G)$ y de la distribución de probabilidades de G) se trabaja solo el caso de variables aleatorias independientes entre sí. En el Apéndice (Sección 6.3.2) se presenta el caso de la variancia de G para variables no independientes.

Si las variables del vector aleatorio son independientes entre sí, puede demostrarse que:

$V(G) = a_1^2 V(Y_1) + a_2^2 V(Y_2) + \cdots + a_k^2 V(Y_k)$.

Y, por lo tanto,
$D(G) = \sqrt{a_1^2 V(Y_1) + a_2^2 V(Y_2) + + a_k^2 V(Y_k)}$.

En la Tabla 6.4 se muestra cómo obtener $D(G)$ para casos particulares de una combinación lineal de variables aleatorias independientes entre sí.

Tabla 6.4 *$D(G)$ para algunos casos particulares de una combinación lineal de variables aleatorias independientes entre sí*

Casos particulares	G	$D(G)=\sqrt{V(G)}$
G es una suma de variables aleatorias (es decir, $a_i=1\ \forall i$)	$G=Y_1+Y_2+\cdots+Y_k$	$\sqrt{V(Y_1)+V(Y_2)+\cdots+V(Y_k)}$
G es una suma de variables aleatorias igualmente distribuidas (es decir, $a_i=1\ \forall i$ y $V(Y_i)=V(Y)\ \forall i$)	$G=Y_1+Y_2+\cdots+Y_k$ Y_i idénticamente distribuidas	$\sqrt{V(Y_1)+\cdots+V(Y_k)}=\sqrt{kV(Y)}$
G es una resta de variables aleatorias (es decir, $a_1=1$ y $a_2=-1$)	$G=Y_1-Y_2$	$\sqrt{1^2\,V(Y_1)+(-1)^2\,V(Y_2)}=$ $=\sqrt{V(Y_1)+V(Y_2)}$

En el Apéndice (Sección 6.8.3) se demuestran las propiedades que permiten expresar a la esperanza y a la variancia de una combinación lineal de variables del vector aleatorio en función de las esperanzas y variancias de dichas variables.

Con lo visto hasta ahora se puede resolver la Actividad 3, Sección 6.6.

6.3.2 Distribución de probabilidades de la nueva variable

¿Cómo obtener la distribución de probabilidades de la nueva variable a partir de la distribución de probabilidades de las variables del vector aleatorio?

Para dar respuesta a esta pregunta se puede recurrir en algunos casos a la Propiedad Reproductiva para algunos modelos y en otros, al Teorema Central del Límite. También pueden obtenerse probabilidades y otros parámetros de la distribución por simulación.

1. Propiedad reproductiva de la suma de variables aleatorias independientes

Para algunas distribuciones de probabilidades es válida la siguiente propiedad: "si dos (o más) variables aleatorias independientes distribuidas según cierto modelo, se suman, la nueva variable aleatoria se distribuye según el mismo modelo que el de los sumandos".

Esta propiedad recibe el nombre de **propiedad reproductiva** y se establece para algunas distribuciones como Normal, Binomial y Poisson. A continuación se presenta y ejemplifica esta propiedad para el modelo Normal; en el Apéndice (Sección 6.8.4) se presenta y ejemplifica esta propiedad para los modelos Binomial y Poisson. En el caso del modelo Normal esta propiedad no solo vale para sumas de variables aleatorias sino también para combinaciones lineales. Para otras distribuciones, como la Uniforme o la Exponencial, esta propiedad no es válida.

Sea $(Y_1, Y_2, \ldots, Y_k)$ un vector aleatorio cuyas variables Y_i son independientes entre sí y tales que $Y_i \sim N(\mu_i, \sigma_i)$ para todo $i = 1, \ldots, k$.

Se define una nueva variable aleatoria

$$G = a_1 Y_1 + a_2 Y_2 + \cdots + a_k Y_k.$$

Entonces,

$$G \sim N\left(a_1 \mu_1 + \cdots + a_k \mu_k,\ \sqrt{a_1^2 \sigma_1^2 + \cdots + a_k^2 \sigma_k^2}\right),$$

por la propiedad reproductiva de la distribución Normal.

A continuación se presentan algunos ejemplos en los que se aplica la propiedad reproductiva de la distribución Normal para obtener la distribución de probabilidades de la nueva variable. Algunos de ellos están relacionados con variables aleatorias definidas en la Sección 6.3, relativas a las barras de acero producidas en la empresa metalúrgica.

Ejemplo 6.2 **En relación a la situación de las barras y los soportes, Sección 6.3**:

Un cliente interesado en las barras que produce la empresa metalúrgica, las debe conectar en soportes metálicos que tienen una perforación circular. Al comprador le interesa que la mayoría de las barras entren adecuadamente en los soportes, para no tener que hacer trabajos adicionales como pulido o rellenado de la perforación, que conllevarían un aumento de los costos.

El diámetro de la perforación de un soporte, en mm, es una variable aleatoria M, de la que se conoce que está distribuida normalmente con promedio 16 y desvío estándar 0,8; mientras que el diámetro de la sección transversal de una barra, en mm, es una variable aleatoria X, de la que se conoce que está distribuida normalmente con promedio 15 y desvío estándar 0,2. La variable $H = M - X$ compara ambos diámetros e informa si la barra puede entrar en el soporte:

- Si $H < 0$, el diámetro de la barra es mayor que el de la perforación del soporte y la barra no entra. En ese caso se requiere pulir la barra para que entre finalmente en el soporte.
- Si $H > 1,2$, la barra entra pero no queda firme. En ese caso se requiere rellenar la perforación del soporte con un material especial para que la barra no se mueva.
- Si $0 < H < 1,2$, la barra entra y queda firme. En ese caso no se requiere ningún trabajo adicional.

Si interesa, por ejemplo, calcular la probabilidad de que no se requiera ningún trabajo adicional, se debe conocer la distribución de probabilidades de la variable H. Esta variable se definió como la diferencia (en mm) entre los diámetros de un soporte y de una barra, $M - X$.

De la primera de las variables (M), se conoce que se comporta según el modelo Normal, con parámetros 16 y 0,8; de la segunda (X), se conoce que se comporta según el modelo Normal con parámetros 15 y 0,2. Además, dado que una variable se refiere a soportes y la otra a barras metálicas, es razonable suponer que se comportan de manera independiente.

Aplicando las expresiones relativas a la esperanza de una diferencia de variables aleatorias (Tabla 6.3) y a la desviación estándar de una diferencia de variables aleatorias independientes (Tabla 6.4), se obtiene que:

$E(H) = E(M - X) = E(M) - E(X) = 16 - 15 = 1.$

$V(H) = V(M - X) = V(M) + V(X) = 0,8^2 + 0,2^2 = 0,68.$

$D(H) = \sqrt{0,68} = 0,825.$

Pero no alcanza con conocer la esperanza y la desviación estándar de la variable H; se requiere conocer su distribución de probabilidades. En este caso, dado que H es una combinación lineal de dos variables independientes que se distribuyen según el modelo Normal, se puede aplicar la Propiedad Reproductiva de la distribución Normal y afirmar que H también se distribuye según dicho modelo.

En resumen:

Sobre las variables del vector aleatorio	**Sobre la nueva variable, expresada como función del vector**
Para cada barra: X: diámetro de la sección transversal, en mm. Población asociada a la variable X: todas las barras metálicas. $X \sim N(15;\ 0,2)$ Para cada soporte: M: diámetro de la perforación, en mm. Población asociada a la variable M: todos los soportes. $M \sim N(16;\ 0,8)$	Para un soporte y una barra, seleccionados al azar, $H = M - X$ es la diferencia entre el diámetro de la perforación del soporte y el diámetro de la sección transversal de la barra, en mm. Población asociada a la variable H: todos los pares formados por un soporte y una barra seleccionados al azar. $H \sim N(1;\ 0,825)$

Conocido el modelo, se pueden obtener diferentes probabilidades y caracterizar el comportamiento de la nueva variable.

En la Figura 6.2 se presenta la distribución de la diferencia entre ambos diámetros y se señalan algunas probabilidades.

Pensando en una gran cantidad de pares (soporte-barra), se observa que el 48,31 % de las veces, la barra entrará adecuadamente en el soporte metálico, sin necesidad de hacer retrabajos; mientras que el 11,27 % de las veces la barra deberá ser pulida para poder entrar finalmente en el soporte y en el 40,42 % restante, la barra entrará en el soporte pero se requerirá de un relleno adicional para que quede firme.

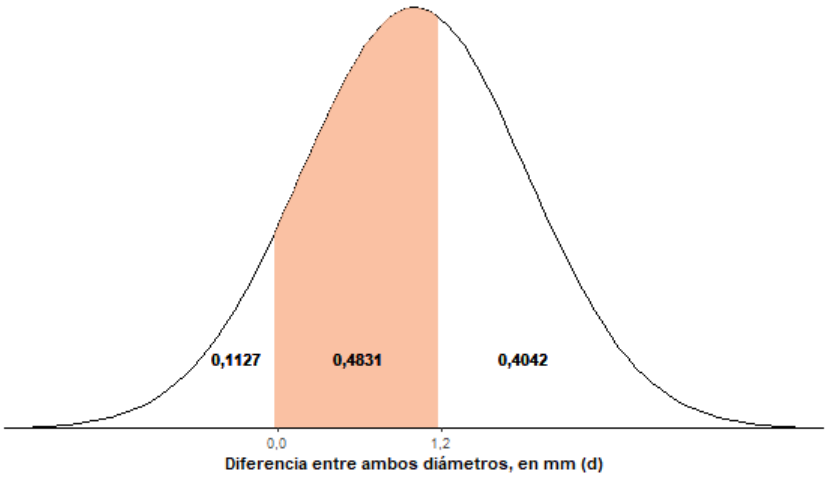


Figura 6.2 *Función de densidad de probabilidad de D, diferencia entre el diámetro de la perforación del soporte y el de la sección transversal de la barra, y probabilidades de interés*

¿El posible comprador de las barras estará conforme con esta información? ¿Qué otras cuestiones deberá tener en cuenta para decidir si comprarlas o no?

Ejemplo 6.3 **En relación a la situación de las barras comercializadas en grupos de 10, Sección 6.3**:

Para una aplicación particular interesa la longitud total de las barras de cada grupo. Suponga que se conoce que la longitud de una barra, en mm, (Y) sigue una ley Normal con promedio 250 y desvío estándar 0,2. Ahora bien, ¿cómo se comporta la longitud total de las barras de cada grupo?

A esa variable, que puede denominarse L, se la puede expresar en función de las longitudes de cada una de las barras de un grupo:

$$L = Y_1 + Y_2 + \cdots + Y_{10}.$$

Dado que las barras se van agrupando a medida que salen del proceso de producción, es razonable pensar que sus longitudes son independientes entre sí y si se aplica la propiedad reproductiva de la distribución Normal, L también puede considerarse normalmente distribuida, por ser una suma de variables normales e independientes.

Ahora bien, ¿cuánto vale el promedio de la nueva variable? ¿y su desviación estándar?

Se pueden aplicar las expresiones para obtener la esperanza de una suma de variables aleatorias (Tabla 6.3) y la desviación estándar de una suma de variables aleatorias independientes (Tabla 6.4). Además, al ser todas las barras del mismo tipo, las variables Y_i pueden considerarse igualmente distribuidas. Entonces, se obtiene que:

$$E(L) = E(Y_1 + Y_2 + \cdots + Y_{10}) = E(Y_1) + E(Y_2) + \cdots + E(Y_{10}) = 10E(Y) = 10 \cdot 250 = 2500$$

$$V(L) = V(Y_1 + Y_2 + \cdots + Y_{10}) = V(Y_1) + V(Y_2) + \cdots + V(Y_{10}) = 10V(Y) = 10 \cdot 0,04 = 0,4$$

$$D(L) = \sqrt{0,4} = 0,632$$

En resumen:

Sobre las variables del vector aleatorio	**Sobre la nueva variable, expresada como función del vector**
Para la barra i, con $i = 1, 2, \cdots, 10$ Y_i: longitud de una barra i, en mm. Población asociada a las variables Y_i: todas las barras metálicas. $Y_i \sim N(250;\ 0,2) \forall i$	Para cada grupo de 10 barras, $L = Y_1 + Y_2 + \cdots + Y_{10}$ se define como la suma de las longitudes de las barras del grupo, en mm. Población asociada a la variable L: todos los grupos de 10 barras. $L \sim N(2500;\ 0,632)$

Conocida la distribución de L, pueden responderse varias preguntas. Suponga, por ejemplo, que una arquitecta usa las barras en la construcción de ciertas estructuras, uniéndolas para obtener barras de longitud mayor. Si requiere armar una barra que tenga una longitud mínima de 2500 mm de largo, ¿le alcanza con comprar un grupo de 10 barras a la metalúrgica? (para esta situación, se supone que la unión entre las barras no aumenta la longitud de la barra ensamblada).

En la Figura 6.3 se presenta la distribución de probabilidades de L y se señala la probabilidad de que la longitud mínima de la barra ensamblada a partir de las 10 barras del grupo supere los 2500 mm. Observe que solo la mitad de las veces, a la arquitecta le alcanzará con un grupo de 10 barras para cumplir con su objetivo.

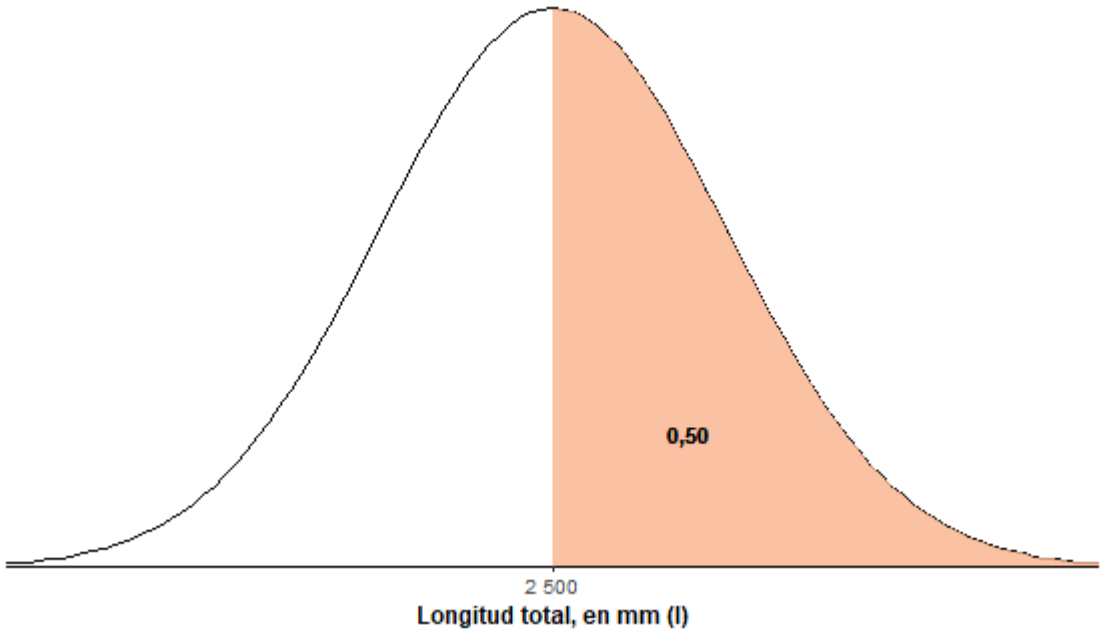


Figura 6.3 *Función de densidad de probabilidad de L, longitud total de grupos de 10 barras. El área de la región en naranja indica la probabilidad de que L supere los 2500 mm*

Ejemplo 6.4 Una empresa distribuye bolsas con mezclas de frutos secos en supermercados y dietéticas. Para armar las bolsas cuenta con 4 máquinas dosificadoras que trabajan de la siguiente manera: la primera vierte una cierta cantidad de nueces peladas; la segunda, una cierta cantidad de almendras; la tercera una cierta cantidad de avellanas y la última, una cierta cantidad de castañas. Luego una máquina mezcla el contenido de cada bolsa y las cierra. El vector aleatorio está formado por las cantidades vertidas por cada dosificadora, (X_1, X_2, X_3, X_4). Estas variables se pueden considerar independientes entre sí, ya que las dosificadoras trabajan independientemente unas de otras. A continuación se presentan las distribuciones de estas variables junto con los precios de costo de cada producto. El costo del envase es de \$75 .

Cantidad vertida (en kg) de:	Distribución de probabilidades	Costo de cada producto, en \$/ kg
Nueces (X_1)	$N(0,5; 0,025)$	1270
Almendras (X_2)	$N(0,3; 0,01)$	2085
Avellanas (X_3)	$N(0,2; 0,022)$	1500
Castañas (X_4)	$N(0,1; 0,008)$	2400

La empresa desea fijar el precio de venta de sus bolsas de tal manera que la ganancia por bolsa sea superior a \$250 como mínimo en el 90 % de las mismas.

El precio de costo de cada bolsa con frutos secos (en \$) se puede definir de la siguiente manera:

$$C = 1270X_1 + 2085X_2 + 1500X_3 + 2400X_4 + 75$$

Dado que X_1, X_2, X_3 y X_4 son variables aleatorias normales e independientes entre sí, por propiedad reproductiva de la distribución Normal, C también puede considerarse distribuida normalmente con parámetros:

$$E(C) = 1270E(X_1) + 2085E(X_2) + 1500E(X_3) + 2400E(X_4) + 75 = 1270 \cdot 0,5 + 2085 \cdot 0,3 + 1500 \cdot 0,2 + 2400 \cdot 0,1 + 75 = 1875,50$$ [4]

$$V(C) = 1270^2\,V(X_1) + 2085^2\,V(X_2) + 1500^2\,V(X_3) + 2400^2\,V(X_4) =$$

$$1270^2 \cdot 0,000625 + 2085^2 \cdot 0,0001 + 1500^2 \cdot 0,000484 + 2400^2 \cdot 0,000064 = 2900,465$$ [5]

$$D(C) = \sqrt{2900,465} = 53,86$$

En resumen:

Sobre las variables del vector aleatorio	Sobre la nueva variable, expresada como función del vector
Para cada bolsa X_1: cantidad vertida de nueces, en kg. X_2: cantidad vertida de almendras, en kg. X_3: cantidad vertida de avellanas, en kg. X_4: cantidad vertida de castañas, en kg. Población asociada a X_i, con $i = 1,\ 2,\ 3,\ 4$: Todas las bolsas que se pueden comercializar. $X_1 \sim N(0,5; 0,025)$ $X_2 \sim N(0,3; 0,010)$ $X_3 \sim N(0,2; 0,022)$ $X_4 \sim N(0,1; 0,008)$	Para cada bolsa, $C = 1270\,X_1 + 2085\,X_2 + 1500\,X_3 + 2400\,X_4 + 75$ representa el costo de cada bolsa, en \$, incluyendo envase y contenido. Población asociada a la variable C: Todas las bolsas que se pueden comercializar. $C \sim N(1875,5; 53,86)$

En la Figura 6.4 se representa gráficamente la distribución del costo de las bolsas, C.

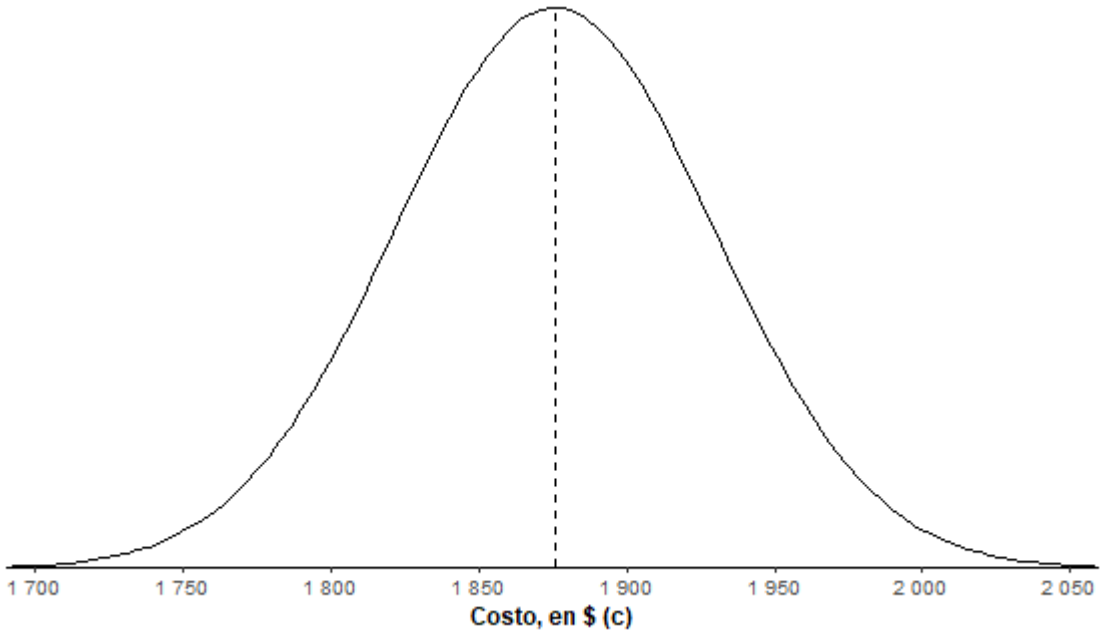


Figura 6.4 *Función de densidad de probabilidad de C, costo de bolsas de frutos secos* $C \sim N(1875,50;\ 53,86)$

Conocido el modelo, se puede describir el comportamiento del costo de las bolsas de frutos secos comercializados por la empresa y brindar alguna información de interés. Se puede afirmar que el costo promedio por bolsa es de \$1875,50 y que la desviación estándar vale \$53,86. Se puede decir también que prácticamente la totalidad de las bolsas (99,73 %) tiene un costo entre \$1713 y \$2037 (aplicando la Regla Empírica).

Ahora bien, en la empresa deben decidir el precio de venta de las bolsas, v, de tal manera que la ganancia por bolsa $(v-C)$ supere los \$250 con probabilidad de al menos 0,90.

La variable aleatoria Ganancia, en \$ por bolsa, (G), se puede definir de la siguiente manera: $G = v - C$, donde v es una constante y C es la variable aleatoria que da cuenta del costo de cada bolsa con frutos secos.

G se distribuye normalmente con $E(G) = v - E(C) = v - 1875,5$ y $D(G) = D(C) = 53,86$ (Ver página 216).

Planteada la pretensión de la empresa en relación a la ganancia, $P(G > 250) \geq 0,90$, se puede despejar el valor de v a partir de la siguiente expresión:

$P(Z > \frac{250-(v-1875,5)}{53,86}) \geq 0,90$

$P(Z > \frac{2125,5-v}{53,86}) \geq 0,90,$

de donde $\frac{2125,5-v}{53,86} \leq -1,28$ y por lo tanto, $v \geq 2194,44$.

Es decir que si se fija el precio de venta de cada bolsa, v, en un valor mayor o igual que \$ 2194,44, se logra que como mínimo el 90 % de las veces la ganancia por bolsa sea mayor que \$250. Si finalmente se fija el precio de venta en \$ 2200 por bolsa, la ganancia G (en \$/bolsa), se distribuye según el modelo $N(324,5; 53,86)$, el cual se representa en la Figura 6.5, junto con la probabilidad pretendida en la empresa.

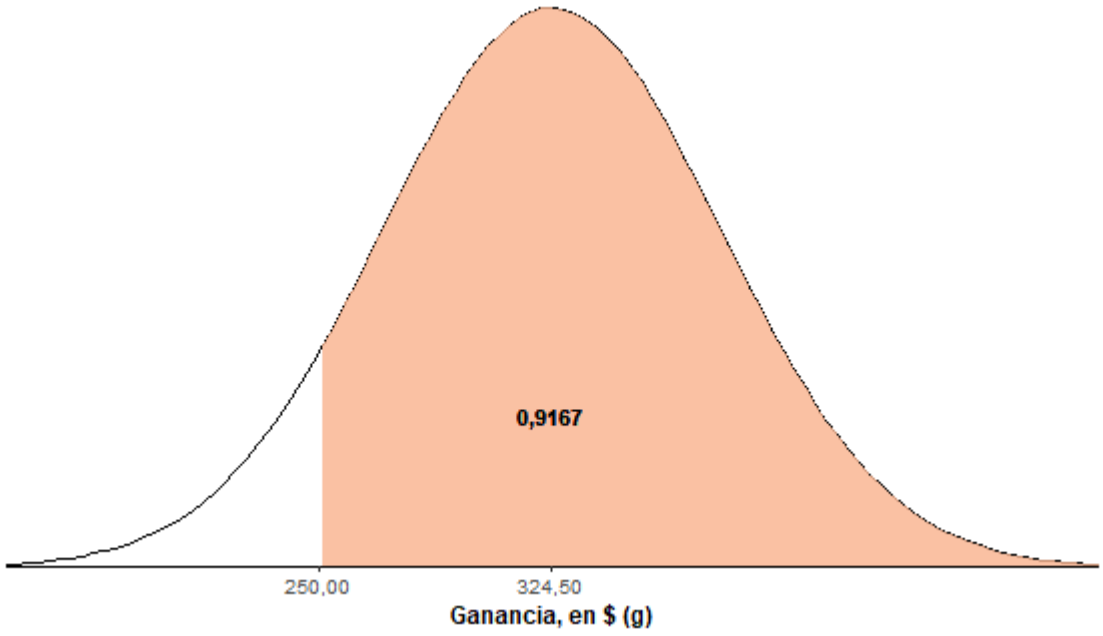


Figura 6.5 *Función de densidad de probabilidad de G, ganancia por bolsa si el precio de venta de cada una se fija en \$2200. El área de la región en naranja indica la probabilidad de que G supere los \$250*

[4] Recuerde que cuando a una variable aleatoria se le suma o resta una constante, la esperanza de la nueva variable se ve modificada por dicha constante: E(Y $\pm constante$) $= E(Y) \pm constante$

[5] Recuerde que cuando a una variable aleatoria se le suma o resta una constante, la variancia de la nueva variable no se modifica: V(Y $\pm constante$) $= V(Y)$



Con lo visto hasta ahora se pueden resolver las Actividades 4 a 7, Sección 6.6. También pueden resolverse las Actividades 8 y 9, relativas a la propiedad reproductiva para los modelos Binomial y Poisson una vez consultado el Apéndice, Sección 6.8.4.

2. Teorema Central del Límite

Este teorema permite obtener una distribución aproximada para la suma de k variables aleatorias independientes con distribuciones conocidas o desconocidas, dados sus promedios y sus variancias, siempre y cuando el número de variables que se sumen, k, sea *suficientemente* grande.

Sea $Y_1, Y_2, \ldots, Y_k$ un conjunto de k variables aleatorias independientes, no necesariamente idénticamente distribuidas, con media y varianza finitas. Es decir que $E(Y_i) = \mu_i$ y $V(Y_i) = \sigma_i^2$, para $i = 1, 2, \ldots, k$.

Sea $G = Y_1 + Y_2 + \cdots + Y_k$.

Entonces, bajo condiciones generales del Teorema Central del Límite, cuando k es suficientemente grande, la distribución de probabilidades de G puede aproximarse mediante una distribución Normal:

$$G \sim N\left(\sum_{i=1}^{k} \mu_i \, , \, \sqrt{\sum_{i=1}^{k} \sigma_i^2}\right) .$$

Algunas consideraciones:

- El número mínimo para obtener una buena aproximación depende de la distribución de las variables sumadas. Cuanto más asimétricas sean dichas distribuciones, más grande deberá ser el valor de k necesario para obtener una buena aproximación.
  Por ejemplo, si las variables sumadas tienen distribución Uniforme, la cual es una distribución simétrica, con sumar solamente 6 ya se obtiene una buena aproximación; pero si la distribución de los sumandos es asimétrica, es necesario sumar cantidades más grandes como 20 o 30 o en algunos casos puntuales de distribuciones muy asimétricas con colas pesadas, más de 100.
- Cuanto mayor es k, mejor se aproxima la distribución de la suma a la Normal. Por ejemplo, si se suman 20 variables, todas con una determinada distribución, la aproximación será mejor que si se suman 10 de esas variables.
  Teniendo en cuenta esta última observación, podemos decir también que la distribución de la suma de variables es más "parecida" a la Normal que la distribución de las variables que se suman. Por ejemplo, al sumar dos variables con distribución Uniforme, se obtiene una variable con distribución Triangular Simétrica. O al sumar dos variables con distribución Exponencial se obtiene una distribución con menor grado de asimetría, donde la moda no es cero.
  Se puede deducir entonces que muchas distribuciones no cumplen la propiedad reproductiva y solo algunas distribuciones conservan su forma al sumarse. Dicho en otras palabras, aunque la suma de variables Uniformes no es Uniforme, su distribución puede aproximarse por una Normal cuando el número de sumandos es suficientemente grande, en concordancia con el Teorema Central del Límite. Lo mismo ocurre con la Exponencial, la Triangular y otras distribuciones.
- El teorema puede aplicarse también si las variables sumadas son discretas; pero cabe aclarar que

una suma de variables aleatorias discretas también resulta una variable aleatoria discreta y es preferible cuando sea posible encontrar una distribución de probabilidades discreta para este tipo de variables, antes que el modelo Normal que corresponde a variables aleatorias continuas.

- Si las variables que se suman presentan diferentes distribuciones de probabilidad, el teorema es válido si sus variancias contribuyen de manera semejante a la variancia de la suma.
- Este teorema se puede aplicar también a cualquier combinación lineal de variables aleatorias ya que, por ejemplo, $G = a_1 X_1 + a_2 X_2 + \cdots + a_k X_k$ puede escribirse como $Y_1 + Y_2 + \cdots + Y_k$ siempre y cuando se cumplan los ítems anteriores.

A continuación se presentan algunos ejemplos en los que se aplica el Teorema Central del Límite para obtener la distribución de probabilidades de la nueva variable.

Ejemplo 6.5 Una empresa fabrica sistemas conectando 50 resistores en serie. Esto significa que la conexión parte del extremo de salida de un resistor al extremo de entrada del otro y no hay otros cables que se ramifiquen de los nodos entre ellos. La resistencia total del sistema se puede expresar como la suma de las resistencias de todos sus componentes, que funcionan independientemente uno del otro y la empresa debe asegurar que esta supere los 157500 Ohms. Por experiencia, se conoce que la resistencia promedio de los resistores es de 3200 Ohms y que la desviación estándar es 230 Ohms; pero no se conoce la distribución de probabilidades de cada X_i.

¿Cómo se comporta la resistencia total de estos sistemas? ¿Puede asegurar la empresa que dicha resistencia supere los 157500 Ohms o deberá incorporar más resistores a los sistemas? ¿Qué puede decir al respecto?

Para conocer el comportamiento de la resistencia total de los sistemas se necesita en primer lugar definir una nueva variable que totalice la resistencias (X) de sus 50 componentes. Esa variable, que puede denominarse T, puede definirse como la suma de las resistencias de cada uno de los componentes:

$T = \sum_{i=1}^{50} X_i$

En este caso, sobre las variables X_i no se conoce su distribución de probabilidades pero sí se tiene idea de los valores de sus esperanzas y variancias:

$E(X_i) = E(X) = 3200,\ i = 1,2,\ldots,50$

$V(X_i) = V(X) = 52900,\ i = 1,2,\ldots,50$

Para obtener la esperanza de la variable T y su variancia, se pueden aplicar las expresiones correspondientes (Tablas 6.3 y 6.4), considerando además que se trata de variables independientes e igualmente distribuidas.

$E(T) = E(\sum_{i=1}^{50} X_i) = \sum_{i=1}^{50} E(X_i) = 50\,E(X) = 50 \cdot 3200 = 160000$

$V(T) = V(\sum_{i=1}^{50} X_i) = \sum_{i=1}^{50} V(X_i) = 50\,V(X) = 50 \cdot 52900 = 2645000$

$D(T) = \sqrt{2645000} = 1626,3$

Resta ahora definir la distribución de probabilidades para la variable T pero, en este caso, no se

puede aplicar ninguna propiedad reproductiva, ya que se desconoce la distribución de las variables X_i. Como T se puede expresar como la suma de una gran cantidad de variables independientes ($k = 50$), se puede aplicar el Teorema Central del Límite para afirmar que esta variable se distribuye aproximadamente según el modelo Normal.

En resumen:

Sobre las variables del vector aleatorio	**Sobre la nueva variable, expresada como función del vector**
Para el resistor i, $i = 1, 2, \cdots, 50$, X_i: resistencia, en Ohms. Población asociada a cada una de estas variables: todos los resistores. $E(X_i) = 3200$ $V(X_i) = 52900$ Se desconoce la distribución de probabilidades de cada X_i.	Para cada sistema construido conectando 50 resistores en serie, $T = \sum_{i=1}^{50} X_i$ informa la resistencia total de este tipo de sistemas, en Ohms. Población asociada a la variable T: todos los sistemas de este tipo. $T \sim N(160000; 1626,3)$

Definida la distribución de probabilidades de la resistencia total de los sistemas, se puede conocer información sobre el comportamiento de los mismos. En este caso puede afirmarse que la resistencia promedio de los sistemas es de 160000 Ohms y su desvío estándar de 1626,3 Ohms. Aplicando la regla empírica se puede afirmar también que para aproximadamente la totalidad de los sistemas la resistencia oscila entre 155000 y 165000 Ohms, de modo que no todos los sistemas cumplirán con lo pretendido por su fabricante. La probabilidad de que un sistema cumpla con lo pretendido es de 0,94 aproximadamente, como se muestra en la Figura 6.6.

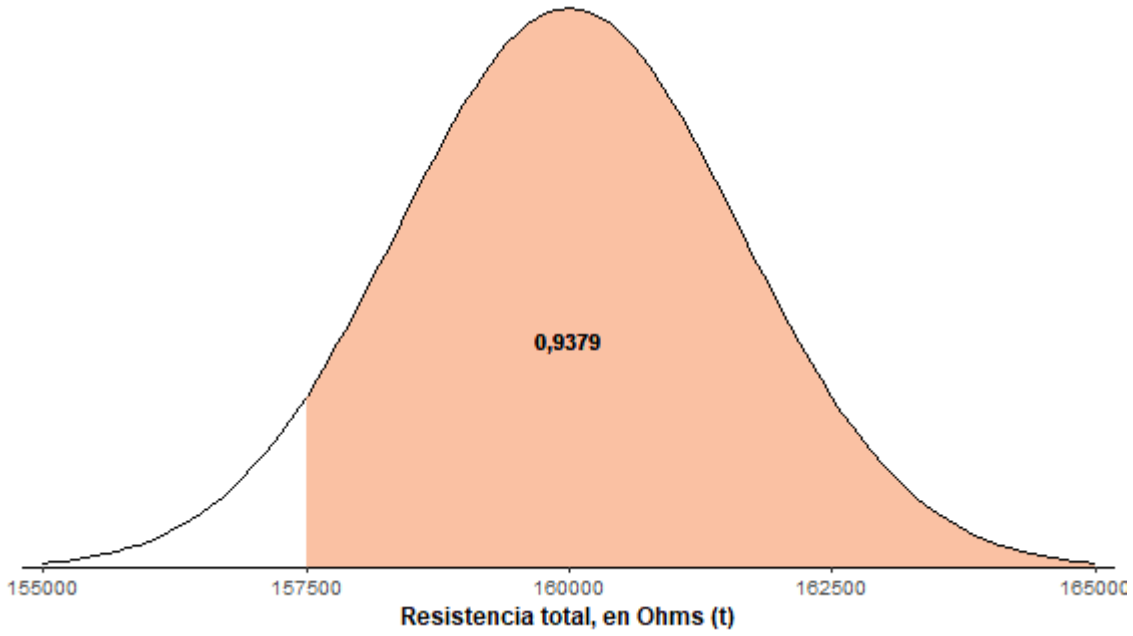


Figura 6.6 *Función de densidad de probabilidad de T, resistencia total de sistemas armados conectando 50 componentes en serie. El área de la región en naranja indica la probabilidad de que la resistencia total supere los 157500 Ohms*



De acuerdo a lo analizado, aproximadamente un 6 % de los sistemas construidos por la empresa conectando 50 resistencias en serie no cumplirán con lo pretendido. ¿Qué pasaría con estos sistemas si se conectara una resistencia más? ¿y si fueran dos?

Ejemplo 6.6 La fabricación de una máquina se realiza en 10 etapas. El tiempo de ejecución de cada etapa (en h) es una variable aleatoria distribuida uniformemente en el intervalo $(1,5)$ y por experiencia se considera razonable suponer que los tiempos de ejecución de cada etapa son variables aleatorias independientes. Si el tiempo total de fabricación supera las 36 h, los costos se incrementan considerablemente, por lo que la ingeniera a cargo está preocupada por esta situación.

¿Cuál es la probabilidad de que esto ocurra?

Para obtener esta probabilidad, en primer lugar se define una nueva variable que represente el tiempo total de ejecución de las 10 etapas. Esta variable, que se denota por T, se define como la suma de los tiempos de ejecución de cada una de las etapas:

$T = \sum_{i=1}^{10} E_i$

En este caso, las variables E_i se distribuyen según el modelo Uniforme con parámetros 1 h y 5 h, de donde se conoce que, para $i = 1,2,\ldots,10$:

$E(E_i) = E(E) = \frac{(5+1)}{2} = 3$

$V(E_i) = V(E) = \frac{(5-1)^2}{12} = 1,3333$

Para obtener la esperanza de la variable T y su variancia, se pueden aplicar las expresiones correspondientes (Tablas 6.3 y 6.4), considerando además que se trata de variables igualmente distribuidas.

$E(T) = E(\sum_{i=1}^{10} E_i) = \sum_{i=1}^{10} E_i = 10E(E) = 10 \cdot 3 = 30$

$V(T) = V(\sum_{i=1}^{10} E_i) = \sum_{i=1}^{10} V(E_i) = 10V(E) = 10 \cdot 1,3333 = 13,33$

$D(T) = \sqrt{13,33} = 3,65$

Resta ahora definir la distribución de probabilidades para la variable T pero, en este caso, no se puede aplicar la propiedad reproductiva, ya que para el modelo Uniforme esta no es válida.

Sin embargo, T corresponde a la suma de $k = 10$ variables aleatorias independientes e idénticamente distribuidas con distribución Uniforme. Debido a que este modelo es simétrico, el valor de k resulta suficiente para que, en virtud del Teorema Central del Límite, la distribución de T pueda aproximarse adecuadamente mediante una distribución Normal.

En resumen:

Sobre las variables del vector aleatorio	Sobre la nueva variable, expresada como función del vector
Para la etapa i, $i = 1, 2, \ldots, 10$ E_i: tiempo de ejecución, en h. Población asociada a cada una de estas variables: todas las veces que se realiza la etapa i. $E_i \sim U(1, 5)$	Para cada máquina, construida en 10 etapas $T = \sum_{i=1}^{10} E_i$ informa el tiempo total de construcción de la máquina, en h. Población asociada a la variable T: todas las veces que se fabrica este tipo de máquinas, o simplemente, todas las máquinas. $T \sim N(30; 3,65)$

Obtenida la distribución de probabilidades del tiempo de producción de las máquinas, T, se puede calcular la probabilidad pedida, $P(T > 36)$, la cual resulta 0,05.
Pensando en un gran número de máquinas, todas producidas en 10 etapas, aproximadamente en el 5 % de las mismas, el tiempo de producción será superior a 36 h. Este dato es de utilidad y la ingeniera a cargo del proceso verá si esto es tolerable o si se necesitan realizar ajustes para reducir los tiempos de producción.

En la Figura 6.7 se presenta la distribución de probabilidades de la variable T y se señala en ella la probabilidad de interés.



Figura 6.7 *Función de densidad de probabilidad de T, tiempo total de construcción para máquinas de cierto tipo. El área de la región en naranja indica la probabilidad de que T supere las 36 h*

Con lo visto hasta ahora se pueden resolver las Actividades 10 a 13, Sección 6.6.

3. Simulación

La simulación es un proceso que genera números aleatorios suponiendo cierto modelo probabilístico y, por lo tanto, permite obtener un conjunto de n observaciones de una población estadística que se comporta según dicho modelo.

Estas n observaciones se pueden considerar una muestra aleatoria de la población de referencia o bien los resultados de n ensayos si se trata de un estudio experimental; pero si n es un número considerablemente grande ($n > 10000$, por ejemplo) permiten aproximar el comportamiento de la población.

En el caso de variables aleatorias que resultan funciones de otras, la simulación puede utilizarse para obtener el valor de algunas probabilidades y otros parámetros, y también construir algún gráfico para tener una idea del modelo, siempre que se conozca el modelo probabilístico de cada una de las variables del vector y la relación entre ellas.

Si estas variables son independientes entre sí, se pueden generar las n observaciones para cada una de ellas y luego obtener las n observaciones de la nueva variable aplicando la función correspondiente (si las variables del vector no son independientes entre sí, se debe conocer la distribución conjunta y generar las n observaciones de cada variable a partir de la misma).

Con las n observaciones de la nueva variable se puede construir un histograma o gráfico de bastones, según corresponda, para tener una idea de su modelo probabilístico. También se pueden obtener medidas de resumen como promedio, mediana, desvío estándar, etc. y frecuencias relativas.

Dado que la simulación permite contar con un gran número de observaciones, las medidas obtenidas pueden aproximarse a los valores poblacionales, es decir, se interpretan como parámetros, y, en particular, las frecuencias relativas se interpretan como probabilidades.

La simulación resulta especialmente útil cuando la relación funcional entre la nueva variable y las variables del vector no es lineal, ya que en el caso de relaciones lineales suele ser posible obtener analíticamente sus parámetros y su distribución.

Ejemplo 6.7 **En relación a la situación del volumen de las barras, Sección 6.3**:

Un posible comprador de las barras de la empresa metalúrgica está interesado en que el volumen de las mismas sea superior a 42000 mm^3. En la empresa registran el diámetro de la sección transversal (X) y en la longitud de las barras (Y), ambas medidas en mm; de modo que el volumen, en mm^3, (J), se puede expresar en función de las variables del vector aleatorio (X, Y) de la siguiente manera: $J = \frac{1}{4}\pi X^2 Y$.

La función mencionada no es lineal y no se pueden aplicar las propiedades vistas en el presente texto. Sin embargo, puede obtenerse información sobre el comportamiento del volumen de las barras aplicando simulación.

Suponga que se conoce que la longitud de las barras se comporta según el modelo $N(250; 0,2)$ y que el diámetro de la sección transversal se comporta según el modelo $U(14,4; 15,6)$. Suponga también que es razonable pensar que ambas variables son independientes entre sí.

Se simulan 20000 valores de cada una de estas dos variables y se obtienen los valores de la variable J a partir de ellos.

En la Figura 6.8 se presenta un histograma, y en la Tabla 6.5 se muestran algunas medidas obtenidas a partir de los valores simulados.

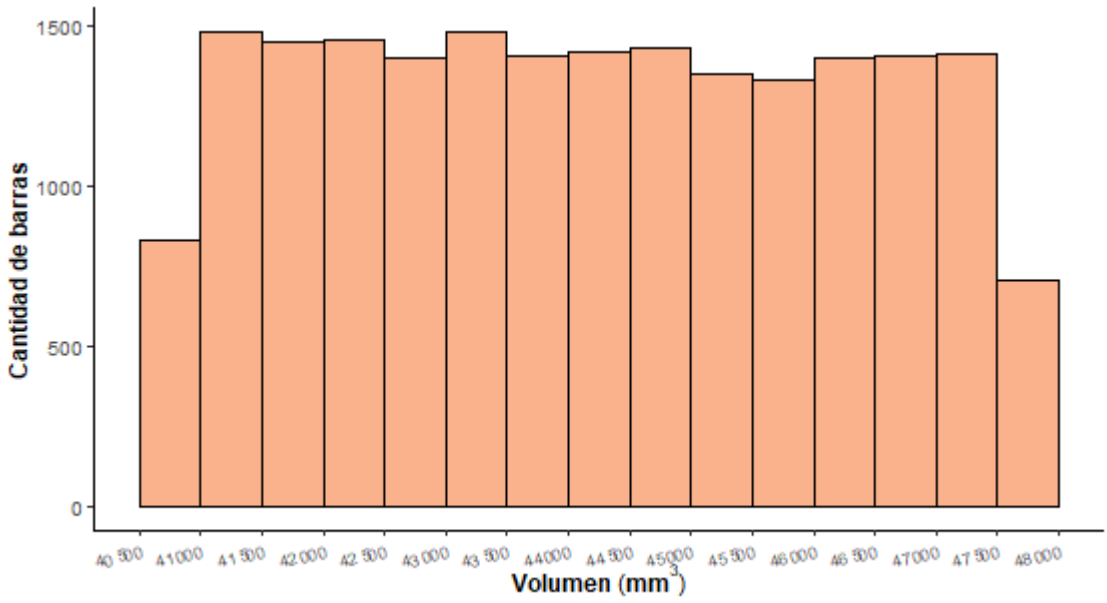


Figura 6.8 *Histograma obtenido a partir de 20000 valores del volumen de las barras de la empresa metalúrgica suponiendo los modelos* $U(14,4;\ 15,6)$ *para el diámetro de la sección transversal y* $N(250;\ 0,2)$ *para la longitud*

Tabla 6.5 *Medidas de resumen obtenidas a partir de 20000 valores del volumen de las barras de la empresa metalúrgica suponiendo los modelos* $U(14,4;\ 15,6)$ *para el diámetro de la sección transversal y* $N(250;\ 0,2)$ *para la longitud*

Medida	Valor (en mm^3)
Mínimo	40656,313
Máximo	47857,424
Promedio	44196,668
Cuartil 1	42429,028
Mediana	44168,744
Cuartil 3	45975,384
Desvío estándar	2040,754

Todos los resultados que se presentan a continuación deben interpretarse como aproximaciones, dado que se obtuvieron a partir de un estudio de simulación basado en 20000 observaciones y bajo el supuesto de validez de los modelos propuestos.

A partir de los resultados de la simulación, y asumiendo que los modelos propuestos para el diámetro y la longitud son adecuados, puede afirmarse que el volumen de las barras producidas en la empresa metalúrgica varía aproximadamente entre 40656,3 mm^3 y 47857,4 mm^3, con un valor medio de 44196,7 mm^3 y una desviación estándar de 2040,8 mm^3. El 50 % de las barras presenta un volumen de hasta 42429 mm^3. La proporción de barras que satisfacen las pretensiones del cliente en relación con su volumen es cercana a 0,81.

El comportamiento del volumen de las barras es aproximadamente simétrico, aunque no puede afirmarse que siga un modelo teórico particular, como el Normal.

Si, por ejemplo, el comportamiento del diámetro de la sección transversal, X, fuera $N(15; 0,2)$, los resultados serían los que se presentan en la Figura 6.11. En ese caso, la distribución del volumen puede asemejarse a una distribución Normal con promedio 44178,6 mm^3 y desvío estándar 1175,7 mm^3.

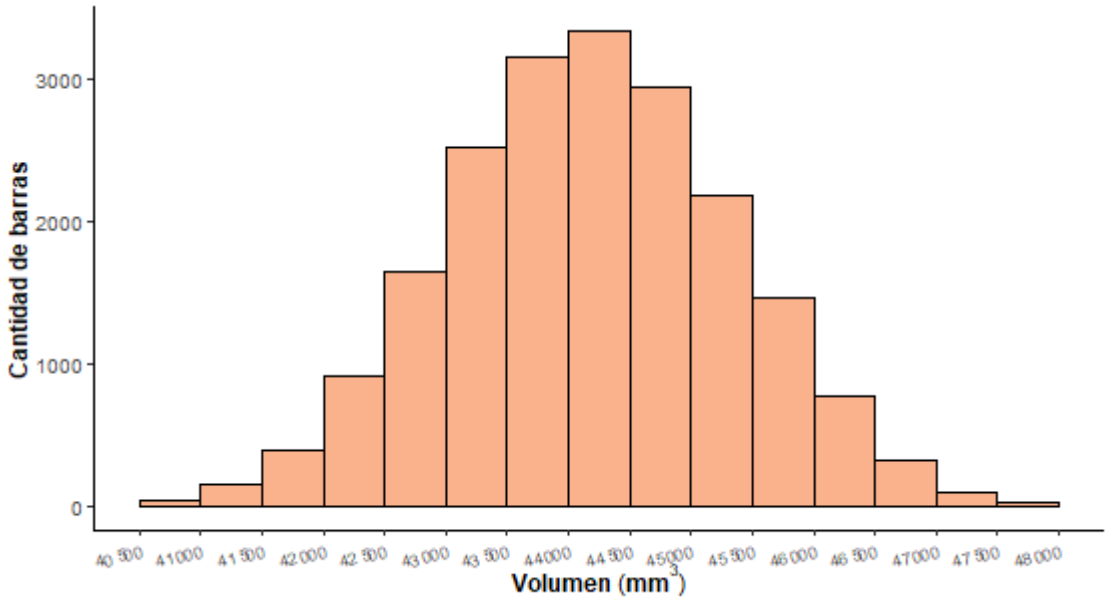


Figura 6.9 *Histograma obtenido a partir de 20000 valores del volumen de las barras de la empresa metalúrgica suponiendo los modelos $N(15; 0,2)$ para el diámetro de la sección transversal y $N(250; 0,2)$ para la longitud*

Tabla 6.6 *Medidas de resumen obtenidas a partir de 20000 valores del volumen de las barras de la empresa metalúrgica suponiendo los modelos $N(15; 0,2)$ para el diámetro de la sección transversal y $N(250; 0,2)$ para la longitud*

Medida	Valor (en mm^3)
Mínimo	39087,212
Máximo	48542,653
Promedio	44178,642
Cuartil 1	43380,204
Mediana	44165,790
Cuartil 3	44976,980
Desvío estándar	1175,722

Ejemplo 6.8 En relación a la situación de las barras y los soportes, Sección 6.3, se obtuvieron las expresiones de la esperanza y la variancia de la diferencia entre el diámetro de la perforación de un soporte y el diámetro de una barra. Se aplicó también la Propiedad Reproductiva del modelo Normal para encontrar que la diferencia mencionada se distribuye $N(1; 0,825)$ y se obtuvieron las probabilidades de interés (Ejemplo 6.2).

¿Podrían haberse obtenido estas probabilidades por simulación?

Conocidos los modelos para el diámetro de la perforación y el diámetro de las barras, se simulan 30000 valores de cada una de estas dos variables y se obtienen los valores de la diferencia entre ambas.

En la Figura 6.10 se presenta un histograma y algunas medidas obtenidas a partir de los 30000 valores simulados.

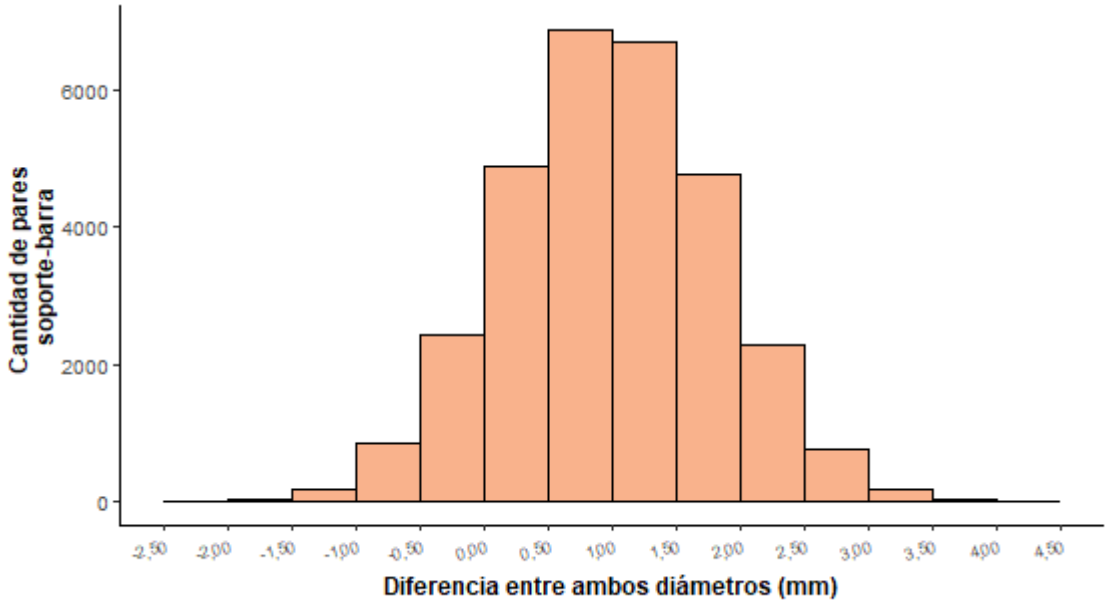


Figura 6.10 *Histograma obtenido a partir de 30000 valores de la diferencia entre el diámetro de la perforación del soporte y el de la barra, suponiendo los modelos* $N(16; 0,8)$ *y* $N(15; 0,2)$ *para cada una de esas variables, respectivamente*

En la Tabla 6.7 se comparan los parámetros correspondientes a la diferencia entre ambos diámetros obtenidos aplicando la propiedad reproductiva de la distribución Normal (Ejemplo 6.2) con los obtenidos a partir de simulación (Ejemplo 6.8). Se observa que los valores son similares entre sí.

Tabla 6.7 *Comparación entre los valores de algunos parámetros correspondientes a la diferencia entre el diámetro de las perforaciones y el de las barras, obtenidos aplicando propiedad reproductiva de la distribución Normal y simulación*

Parámetros asociados a la variable H	**Valores obtenidos aplicando propiedad reproductiva (Ejemplo 6.2)**	**Valores obtenidos por simulación (Ejemplo 6.8)**
$E(H)$	1 mm	$0,985$ mm
$D(H)$	$0,825$ mm	$0,825$ mm
$P(H < 0)$	$0,1127$	$0,1165$
$P(0 < H < 1,2)$	$0,4831$	$0,4846$
$P(H > 1,2)$	$0,4042$	$0,3989$

La Tabla 6.7 muestra que los resultados obtenidos por simulación son similares a los obtenidos analíticamente, lo cual avala la aplicación de esta herramienta computacional para conocer los parámetros de nuevas variables.

Ejemplo 6.9 En relación a la situación del costo de eliminar poros y manchas de las barras, Sección 6.3, suponga que las barras se envían a un cliente en lotes de 10 unidades pero previamente se eliminan los poros y las manchas que estas presentan. Interesa estudiar el comportamiento del costo de eliminar los poros y las manchas de todas las barras que componen cada lote, C_L.

Utilizando la información de la Tabla 6.1, que informa sobre el comportamiento conjunto de la cantidad de poros y de manchas de las barras, y sabiendo que la eliminación de cada poro cuesta \$ 350 y de cada mancha, \$ 200, se puede construir la distribución de probabilidades del costo de eliminar los poros y las manchas de cada barra, C_B, la cual se presenta en las dos últimas columnas de la Tabla 6.8 y en la Figura 6.11.

Tabla 6.8 *Comportamiento conjunto de la cantidad de poros y de manchas de las barras y distribución de probabilidades del costo de eliminar ambos defectos en cada barra*

Cant. de poros por barra (W_1)	**Cant. de manchas por barra** (W_2)	**Costo total de eliminación, en \$** (C_B) $C_B = 350W_1 + 200W_2$	**Probabilidad asociada** $P(CB_{=}c_B)$
0	0	0	0,3577
0	1	200	0,1323
1	0	350	0,2628
1	1	550	0,0972
2	0	700	0,0876
2	1	900	0,0324
3	0	1050	0,0219
3	1	1250	0,0081

Nota: Aunque el costo suele modelarse como una variable continua, en este caso la variable aleatoria C_B debe considerarse discreta, ya que solo puede asumir un conjunto finito de valores determinados por las posibles combinaciones de poros y manchas por barra.

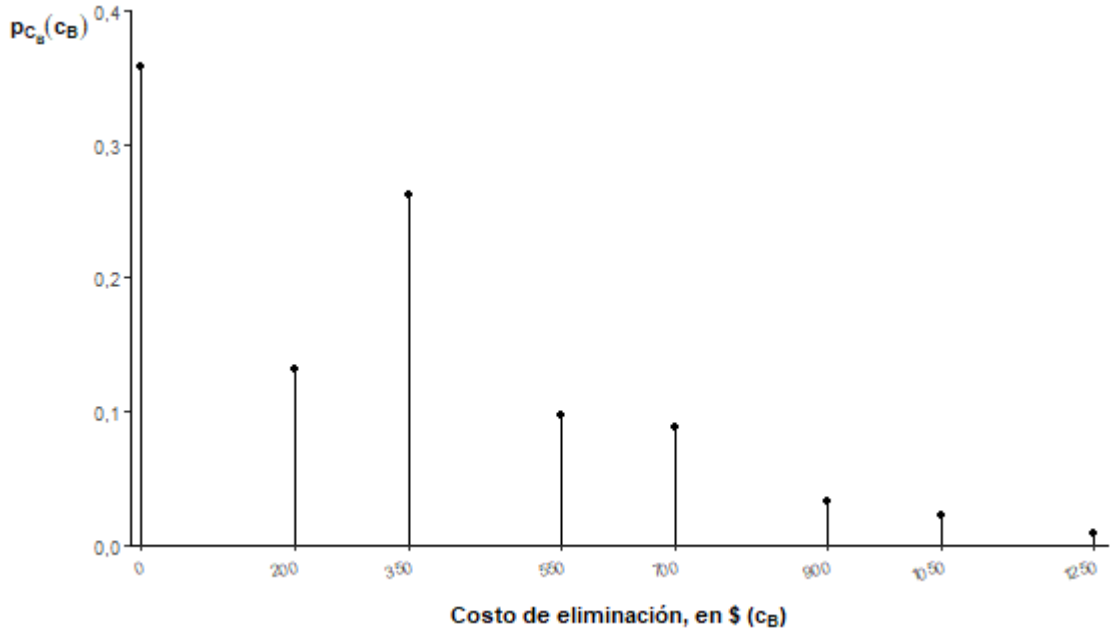


Figura 6.11 *Distribución de probabilidades de C_B, costo de eliminar los poros y las manchas de cada barra*

Aplicando los conceptos vistos en el Capítulo 3 se pueden obtener, entre otras, las siguientes medidas:

$E(C_B) = 295,5;$

$V(C_B) = 85536,75.$

El costo de eliminar los poros y las manchas de las 10 barras que componen cada lote, C_L, se puede definir de la siguiente manera:

$C_L = C_{B1} + C_{B2} + \cdots + C_{B10}.$

En este caso, las variables C_{Bi} están igualmente distribuidas y el modelo que describe su comportamiento en probabilidad es el que se presenta en la Tabla 6.8 y en la Figura 6.11. Si los lotes se van armando a medida que las barras se producen, se puede suponer razonablemente que las cantidades de defectos de las barras de cada lote (y por consiguiente, los costos de eliminarlos) son independientes entre sí.

Con esta información se puede obtener, analíticamente, el valor de los siguientes parámetros:

$E(C_L) = E(\sum_{i=1}^{10} C_{Bi}) = \sum_{i=1}^{10} E(C_{Bi}) = 10E(C_B) = 10 \cdot 295,5 = 2955;$

$V(C_L) = V(\sum_{i=1}^{10} C_{Bi}) = \sum_{i=1}^{10} V(C_{Bi}) = 10V(C_B) = 10 \cdot 85536,75 = 855367,5;$

$D(C_L) = \sqrt{855367,5} = 924,86.$

Sin embargo, como se observa en la Figura 6.11, el comportamiento de la variable C_B no es simétrico y el número de variables que se suman no es lo suficientemente grande como para poder aplicar el Teorema Central del Límite. Se puede recurrir entonces a la simulación.

Conocido el comportamiento del costo de eliminar los poros y las manchas de cada barra, C_B, descripto en la Tabla 6.8, se simulan 35000 valores para cada una de las 10 variables asociadas a las barras de cada lote y se obtienen 35000 valores de C_L. En la Figura 6.12 se presenta el histograma obtenido a partir de los valores simulados.

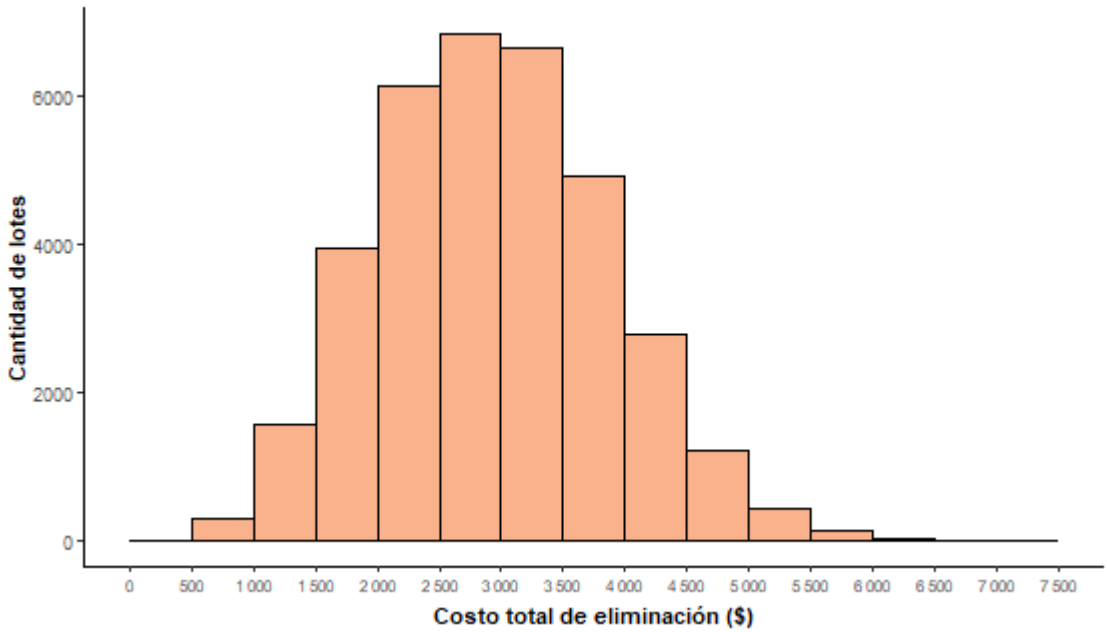


Figura 6.12 *Histograma obtenido a partir de 35000 valores de la suma de los costos de eliminar los poros y las manchas de las 10 barras de un lote, suponiendo el modelo descripto en la Tabla 6.8 para cada barra*

Tabla 6.9 *Medidas de resumen obtenidas a partir de 35000 valores de la suma de los costos de eliminar los poros y las manchas de las 10 barras de un lote, suponiendo el modelo descripto en la Tabla 6.8 para cada barra*

Medida	Valor (en $)
Mínimo	0,000
Máximo	7150,000
Promedio	2961,463
Cuartil 1	2300,000
Mediana	2900,000
Cuartil 3	3550,000
Desvío estándar	929,923

Los valores del promedio y del desvío estándar del costo de limpiar las 10 barras de cada lote obtenidos por simulación son similares a los obtenidos analíticamente. La simulación permite obtener, además, de manera aproximada, los valores de otros parámetros, como algunos de los que se presentan en la tabla anterior y tener una idea de la forma de la distribución de probabilidades de la nueva variable. Se podrían haber obtenido, también, algunas probabilidades de interés.



Con lo visto hasta ahora se pueden resolver las Actividades 14 a 16, Sección 6.6.

Sea $(X_1, \ldots, X_k)$ un vector de variables aleatorias independientes e idénticamente distribuidas, todas con la misma distribución que una variable aleatoria X.

En general, sumar estas variables no es equivalente a multiplicar una de ellas por el factor k.

En otras palabras, las variables

$$T = \sum_{i=1}^{k} X_i \quad \text{y} \quad M = kX$$

no coinciden en distribución.

Considere el siguiente ejemplo, que ilustra lo afirmado en el párrafo anterior.

Ejemplo 6.10 En un taller se realizan reparaciones de equipos de medición. Cuando cada equipo llega, un técnico especializado lo revisa, identifica la cantidad de problemas que deben ser reparados y elabora el presupuesto por las reparaciones. Por experiencia se conoce que la cantidad de problemas que presentan los equipos (X) se comporta según la siguiente distribución:

Cant. de problemas (X)	$P(X = x) = p_X(x)$
1	0,88
2	0,09
3	0,03
Total	**1**

Para esta distribución de probabilidades se obtiene que $E(X) = 1,15$, $V(X) = 0,1875$ y $D(X) = 0,43$.

La población asociada a la variable X está formada por todos los equipos que llegan al taller para ser reparados. A partir de la distribución se observa que la cantidad de problemas que estos equipos presentan oscila entre 1 y 3 y que predominan los equipos con un solo problema para reparar. En promedio, los equipos presentan 1,15 problemas y el desvío estándar vale 0,43 problemas.

1. Por cada problema detectado en el equipo de medición se cobran 2 unidades monetarias [6] en concepto de elaboración de presupuesto. Interesa el monto total que se cobra al dueño de cada equipo por ese concepto (M).
   La población asociada a la variable M está formada por todos los equipos que llegan al taller para ser reparados. Dado que la cantidad de problemas puede tomar los valores 1, 2 o 3 y que por cada problema se cobran 2 unidades monetarias, la nueva variable puede tomar los valores 2, 4 o 6 con las mismas probabilidades. La distribución de probabilidades de esta variable es la siguiente:

Monto (M)	$P(M = m) = p_M(m)$
2	0,88
4	0,09
6	0,03
Total	**1**

Para esta distribución de probabilidades se obtiene que $E(M) = 2,30$, $V(M) = 0,75$ y $D(M) = 0,86$.

El monto que se debe pagar en concepto de presupuesto, para cada equipo, oscila entre 2 y 6 unidades monetarias y que predominan los equipos para los cuales se abonan 2 unidades monetarias. En promedio, se paga un monto de 2,30 unidades monetarias por equipo y el desvío estándar vale 0,86 unidades monetarias.

2. Suponga ahora que se seleccionan aleatoriamente 2 equipos entre los que deben ser reparados y se observa el número total de problemas en ambos. Surge el vector aleatorio (X_1, X_2).

   Por la selección al azar, es razonable pensar que las cantidades de problemas en esos equipos, X_1 y X_2, son independientes entre sí. Además, ambas variables están igualmente distribuidas y la distribución de probabilidades de cada una coincide con la de la variable X.

   Interesa conocer el comportamiento del número total de problemas en los dos equipos seleccionados, T, la cual puede definirse de la siguiente manera: $T = X_1 + X_2$.

   La población asociada a la variable T está formada por todos los conjuntos de dos equipos que llegan al taller para ser reparados. Dado que la cantidad de problemas puede tomar los valores 1, 2 o 3 y que se consideran dos equipos, la nueva variable puede tomar los valores 2, 3, 4, 5 o 6. La distribución de probabilidades de esta variable es la siguiente:

Total de problemas (T)	$P(T = t) = p_T(t)$
2 (cada equipo presenta un problema)	0,7744
3 (un equipo presenta un problema y el otro, dos)	0,1584
4 (cada equipo presenta dos problemas o bien, uno presenta uno y el otro, tres)	0,0609
5 (un equipo presenta dos problemas y el otro, tres)	0,0054
6 (cada equipo presenta tres problemas)	0,0009
Total	**1**

   La distribución de probabilidades de la variable T se construyó aplicando propiedades relacionadas con la probabilidad de la intersección de sucesos independientes y con la unión de sucesos excluyentes, presentadas en el Capítulo 4.

   Para esta distribución se obtiene que $E(T) = 2,30$, $V(T) = 0,375$ y $D(T) = 0,612$.

   La cantidad de problemas en ambos equipos oscila entre 2 y 6 y que predominan los pares de equipos para los cuales se presentan 2 problemas en total. En promedio, para cada par de equipos se encuentran 2,3 problemas y el desvío estándar vale 0,612 problemas.



$T = X_1 + X_2$ y $M = 2X$, ¿representan la misma variable?

Como se puede observar, T y M no representan la misma variable aleatoria. En efecto, T corresponde a la suma de dos variables aleatorias independientes e idénticamente distribuidas con la misma ley que X, mientras que M es una transformación lineal de una única variable

aleatoria. En consecuencia, aunque ambas pueden tener la misma esperanza, presentan distinta variabilidad y, por lo tanto, distribuciones diferentes.

Los valores de E, V y D para cada una de las variables consideradas en este ejemplo se obtuvieron aplicando las expresiones correspodientes, presentadas en el Capítulo 3.
Para el caso de la variable M también se podrían haber obtenido aplicando las propiedades de la esperanza y la variancia de una transformación lineal, vistas en el Capítulo 3.
Para el caso de la variable T, también se podrían haber obtenido aplicando las propiedades de la esperanza y de la variancia de una suma de variables independientes e igualmente distribuidas, vistas en la Sección 6.3.1.

[6] La expresión "unidades monetarias" se utiliza únicamente con fines didácticos.

Con lo visto hasta ahora se pueden resolver las Actividades 17 a 19, Sección 6.6.

6.4 Introducción al estudio de las mediciones y sus errores

En el Capítulo 1 se habló de la importancia de estudiar los procedimientos de medición que se van a utilizar para medir las variables de interés en la resolución del problema de naturaleza estadística que se aborde. También se planteó que el estudio de las propiedades de estos procedimientos (como la exactitud y la precisión, entre otras) también constituye un problema de naturaleza estadística.

En el Capítulo 3 se definió a la variable "resultado de una medición" (X), asociada a la población formada por todas las mediciones que se pueden realizar de una misma magnitud δ, con el procedimiento de interés y bajo las mismas condiciones.

En el resultado de una medición pueden estar presentes el error sistemático (o sesgo) y el error aleatorio. El primero representa la parte del error que es igual para cada medición y su valor se representa con la letra k; el segundo varía entre mediciones, es decir, es una variable aleatoria que se simboliza con ε.

La variable aleatoria X se puede expresar, entonces, de la siguiente manera: $X = \delta + k + \varepsilon$

Ahora bien, ¿cómo se distribuye la variable ε? ,¿y la variable X?

El error aleatorio de una medición (ε) se puede pensar como la suma de un gran número de causas independientes aproximadamente de la misma magnitud y cuyos efectos parciales tienen una distribución cualquiera. Si esto es razonable, en virtud de Teorema Central del Límite, su distribución de probabilidades se puede aproximar a la distribución Normal. Luego $\varepsilon \sim N(0\ ,\ \sigma)$.

Recuerde que, en el largo plazo se compensan los errores por defecto y los errores por exceso y por lo tanto es razonable pensar que $E(\varepsilon) = 0$.

El resultado de una medición, (X) se puede expresar en función de la variable ε, concretamente X se obtiene sumando una constante ($\delta + k$) a dicha variable. Por ese motivo, se puede afirmar que X también se distribuye según el modelo Normal, con los siguientes parámetros:

$E(X) = E(\delta + k + \varepsilon) = \delta + k + E(\varepsilon) = \delta + k$ (ya que $E(\varepsilon) = 0$)

$V(X) = V(\delta + k + \varepsilon) = V(\varepsilon) = \sigma^2$

$D(X) = D(\varepsilon) = \sigma$

En síntesis, en virtud del Teorema Central del Límite, $X \sim N(\delta + k, \sigma)$.

Si el procedimiento es exacto, es decir, si $k = 0$, $X \sim N(\delta, \sigma)$.



Problema 7 (pág. 8): *Una empresa desea evaluar el desempeño de un multímetro digital para medir tensión en corriente continua. Para ello lo envía a un laboratorio especializado en el que van a realizar mediciones repetidas de una tensión conocida, con el mencionado equipo.*
En este contexto, ***la población está definida por todas las mediciones que se pueden realizar con cierto procedimiento (que incluye el uso del multímetro), de una tensión conocida*** δ***, en condiciones similares*** *y* ***la variable de interés*** X ***es el resultado de una medición de la tensión*** δ ***con ese procedimiento****.*
Considerando los posibles errores de este procedimiento de medición de distancias, la variable X *se puede expresar de la siguiente manera:* $X = \delta + k + \varepsilon$*, de modo que:*
Como el error aleatorio de una medición de la tensión (ε*) se puede pensar como la suma de un gran número de causas independientes aproximadamente de la misma magnitud y cuyos efectos parciales tienen una distribución cualquiera, en virtud de Teorema Central del Límite, su distribución de probabilidades se puede aproximar a la distribución normal. Luego* $\varepsilon \sim N(0, \sigma)$ *y, consecuentemente,* $X = \delta + k + \varepsilon$ *se comporta también normalmente, con los siguientes parámetros:*

$E(X) = \delta + k$ *y* $D(X) = D(\varepsilon) = \sigma_\varepsilon$.

Es decir, $X \sim N(\delta + k, \sigma)$.

6.5 Síntesis

En muchos problemas de Ingeniería las variables de interés se pueden expresar como función de dos o más variables aleatorias, es decir, de un vector aleatorio. Para poder tomar decisiones que involucren a estas variables es necesario determinar su distribución de probabilidades y los parámetros correspondientes.

Existen propiedades que permiten obtener la esperanza y la variancia para estas nuevas variables y otras, como las propiedades reproductivas y el Teorema Central del Límite, que permiten definir su distribución de probabilidades. También está la posibilidad de utilizar la simulación para obtener algunas probabilidades y otros parámetros de interés.

En particular, el Teorema Central del Límite permite aproximar en algunos casos, la distribución del error aleatorio de las mediciones, ε, al modelo Normal, lo cual habilita a utilizar este modelo para describir el comportamiento del resultado de cada medición, X.

Entre las propiedades presentadas en este capítulo se destacan las siguientes:

- Si $G = a_1 Y_1 + a_2 Y_2 + \cdots + a_k Y_k$ con $a_i \in \mathbb{R}$, para $i = 1, 2, \ldots, k$, se demuestra que

$E(G) = a_1 E(Y_1) + a_2 E(Y_2) + \cdots + a_k E(Y_k)$.

- Si, además, las variables $Y_1, Y_2, \ldots, Y_k$ son independientes, se demuestra que $V(G) = a_1^2 V(Y_1) + a_2^2 V(Y_2) + \cdots + a_k^2 V(Y_k)$ y, por lo tanto, $D(G) = \sqrt{a_1^2 V(Y_1) + a_2^2 V(Y_2) + \cdots + a_k^2 V(Y_k)}$.
- Si las variables $Y_1, Y_2, \ldots, Y_k$ son independientes y normales, es decir, $Y_i \sim N(E(Y_i) = \mu_i, D(Y_i) = \sigma_i)$ para todo $i = 1, \ldots, k$, entonces cualquier combinación lineal $G = a_1 Y_1 + a_2 Y_2 + \cdots + a_k Y_k$ también tiene distribución Normal, por la propiedad reproductiva de la distribución Normal: $G \sim N\left(a_1 \mu_1 + a_2 \mu_2 + \cdots + a_k \mu_k,\ \sqrt{a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \cdots + a_k^2 \sigma_k^2}\right)$.
- Se puede plantear la propiedad reproductiva para variables definidas como sumas de variables independientes con distribuciones Poisson o Binomial, entre otras.
- Si G se define como una suma de variables aleatorias independientes, $G = Y_1 + Y_2 + \cdots + Y_k$, no necesariamente idénticamente distribuidas, con $E(Y_i) = \mu_i$ y $V(Y_i) = \sigma_i^2$ para $i = 1, \ldots, k$, donde las medias y varianzas son finitas, entonces, bajo las condiciones generales del Teorema Central del Límite y para k suficientemente grande, la distribución de G puede aproximarse mediante una distribución Normal: $G \sim N\left(\mu_1 + \mu_2 + \cdots + \mu_k,\ \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_k^2}\right)$.
- La simulación permite obtener de manera aproximada la distribución de probabilidades de una variable G definida como función de las variables del vector aleatorio $(Y_1, Y_2, \ldots, Y_k)$, cuando se dispone de un modelo probabilístico para sus componentes.

6.6 Actividades propuestas

1. Se ensamblará un circuito eléctrico de tal manera que cuando una componente falle comience a funcionar la siguiente, y así sucesivamente con 40 componentes. Sean Y: duración del circuito, en horas y X_i: duración de la componente i-ésima, en horas.
   a) Exprese a Y en función de las variables X_i
   b) ¿Qué variable debería definir para saber si la primera componente que se conecta en el sistema dura más que la segunda?
   c) ¿Qué variable debería definir si interesa la duración del circuito en minutos?

2. Para las siguientes variables aleatorias, defina un vector aleatorio asociado y expréselas como función de las variables de dicho vector.
   a) Espesor, en mm, de piezas metálicas fabricadas superponiendo 15 láminas.
   b) Peso, en kg, de cajas que contienen 20 botellas de leche.
   c) Cantidad de quejas que se reciben mensualmente en una empresa de servicios.
   d) Consumo eléctrico, en kW, de una región determinada, en períodos de una hora.
   e) Cantidad de imperfecciones en tramos de 100 m de alambre.

3. Suponga que respecto de las componentes que se conectan para formar el circuito mencionado en la Actividad 1 se conoce que $E(X_i) = 150$ h y $V(X_i) = 200$ h^2 $\forall i$ y se puede pensar también que las duraciones de las componentes son independientes entre sí.
   Indique cuánto valen la esperanza y la variancia de las variables definidas en dicha actividad.

4. La duración de cierto componente (X), en horas, sigue una ley Normal con parámetros $\mu = 25$ y $\sigma = 0,6$. Se seleccionan 4 componentes al azar, los cuales funcionan independientemente.
   a) ¿Cuál es la probabilidad de que los 4 componentes seleccionados duren más de 25 h cada uno?
   b) Suponga que los 4 componentes se conectan de tal manera que cuando el primero deja de funcionar, se activa automáticamente el segundo y así sucesivamente. ¿Cuál es la probabilidad de que los 4 componentes duren en total más de 100 h?
   c) Explique por qué las probabilidades obtenidas en los ítems anteriores no coinciden. Justifique por qué la probabilidad obtenida en el ítem *b* es mayor.

5. En la empresa metalúrgica productora de las barras de acero se realiza un control del proceso a partir de grupos de 10 unidades (Sección 6.3, página 301). En cada una de las barras seleccionadas se mide la longitud, en milímetros (Y) y luego se obtiene la longitud promedio del grupo (M). Por la forma en que se seleccionan las barras de cada grupo es razonable pensar que las longitudes de las mismas son independientes entre sí. Suponga que Y se comporta según el modelo Normal con parámetros $\mu = 250$ y $\sigma = 0,2$.
   a) Defina el vector aleatorio asociado a cada grupo y exprese a la longitud promedio en función de las variables que lo integran.
   b) Indique cómo se comporta la longitud promedio de grupos de 10 barras. Justifique su respuesta.
   c) De acuerdo al criterio que se describe en relación al control del proceso en la página mencionada, ¿en qué proporción de las veces se decide revisar las longitudes de todas las barras producidas en la última hora?

6. Una máquina embotelladora automática está programada para inyectar un cierto volumen de líquido (Y_V), en cm^3. Se puede considerar que el contenido vertido está distribuido normalmente con media 988 y desvío estándar 5, mientras que el volumen de las botellas (Y_B), en cm^3, se distribuye normalmente con media 1000 y desvío 4.
   Por el costo del líquido que se envasa, se pretende que no se produzcan derrames, aunque se admite que esto ocurra como máximo el 1 % de las veces. Si este porcentaje se supera, se llevará a cabo un reajuste en la máquina embotelladora.
   ¿Qué recomendaría usted a la empresa? Justifique exhaustivamente.

7. Una solución utilizada como refrigerante para unas maquinarias especiales se envasa en tambores. Antes del envasado, la solución se encuentra en una tina que contiene 2400 L y desde ahí se van cargando los tambores. La carga de la tina a los tambores se hace por un sistema automático que envía un volumen (V), en litros, cuyo comportamiento es Normal, con promedio de 29,6 y desvío estándar de 0,5.
   Suponga que durante una jornada de trabajo se planean llenar 80 tambores. ¿Considera que alcanza con el contenido de la tina para esta tarea? Justifique.

8. Un cierto tipo de mezcla de concreto se comercializa en bolsas que están etiquetadas con un peso de 20 kg; aunque, en realidad, el peso (en kilogramos) de las mismas es una variable aleatoria distribuida según el modelo $N(20; 0,04)$. Las bolsas se comercializan en pallets de 10 unidades. Para un cliente en particular se va a realizar un envío de 25 pallets. Suponga que interesa que las bolsas no pesen más de 20,04 kg.
   a) ¿Cuál es la probabilidad de que un pallet haya más de 3 bolsas con un peso superior a 20,04 kg? Defina la población y la variable de interés, así como el modelo correspondiente. Justifique.
   b) ¿Cuál es la probabilidad de que en un envío de 25 pallets haya más de 50 bolsas con un peso superior a 20,04 kg? Definida la población como los infinitos envíos de 25 pallets, escriba el vector aleatorio que informa la cantidad de bolsas con peso superior a 20,04 kg en cada pallet del envío y defina a la nueva variable en función de las variables del vector. Indique la distribución de probabilidades de la misma. Justifique.
   c) Interprete las probabilidades obtenidas en los ítems anteriores como frecuencias relativas poblacionales.

9. Una compañía alquila y repara fotocopiadoras de alto rendimiento, de uso en empresas y oficinas. Para un modelo de fotocopiadora en particular se conoce que las fallas se producen según un modelo de Poisson, a razón de 0,045 fallas por día. A comienzos de este año, la empresa tiene alquiladas 32 máquinas de ese tipo en la zona y ha contratado a una técnica especializada en repararlas, que por experiencia puede atender como máximo 2 fallas por día. ¿En qué proporción de las jornadas la técnica no podrá atender las fallas de las máquinas a cargo de la empresa? ¿Recomendaría la contratación de más personal para esta tarea? Justifique.

10. Reconsidere las actividades 1 y 3, referidas a circuitos eléctricos ensamblados con 40 componentes.
    a) Proponga un modelo aproximado para la duración de los mismos (Y, medida en horas).
    b) Indique cuánto vale la proporción de circuitos que superan las 6200 h.

11. En la empresa productora de barras hacen un tratamiento para eliminar los poros y las manchas

que estas poseen, antes de comercializarlas. Interesa estudiar el costo de eliminación de estos dos tipos de defectos en cada barra (Sección 6.3, página 300).

a) Dada la variable C_B, costo (en %) de eliminar los poros y las manchas de una barra, observe su distribución de probabilidades presentada en el Ejemplo 6.9 (Tabla 6.8, Figura 6.11, página 322).
   1) Defina la población asociada a esta variable.
   2) Indique, a modo de ejemplo, cómo se obtuvo el valor $0,0972$ asociado a $c_B = \$550$.
   3) Obtenga e interprete, en contexto, dos medidas de localización y dos de variabilidad.

b) Suponga ahora que un cliente compra las barras en lotes de 50 unidades y que para que la venta sea rentable, el costo total de eliminación de los poros y las manchas de las barras de cada lote no debe superar los $ 18000. ¿Considera que la venta de los lotes resulta rentable? Justifique.

12. Reconsidere la situación del servidor neumático incorporado en sembradoras, que distribuye semillas en surcos (Problema 3, Sección 1.4, página 7). Suponga que la distancia (en centímetros) entre una semilla y la siguiente cuando se utiliza dicho servidor se comporta según el modelo Exponencial con parámetro $\alpha = 0,018$/cm. Se cuenta con 100 semillas, ¿considera que estas alcanzan para sembrar un surco de 52 metros de longitud? Justifique (suponga para este caso que no se desperdicia ninguna semilla en el proceso de sembrado).

13. Una empresa fabrica componentes electrónicas cuya duración (en horas) sigue una ley Exponencial. Se conoce además que el 90 % de las componentes no supera las 80 h de uso.

a) ¿Cuál es la duración promedio de las componentes? ¿Y su desvío?

b) Un usuario conecta estas componentes en un sistema de tal manera que cuando una falla automáticamente se activa la siguiente y así sucesivamente. ¿Qué cantidad mínima de componentes debe conectar si necesita que el sistema realice una tarea continua durante más de 1600 h? (Suponga que este usuario pretende que como mínimo el 99 % de estos sistemas superen las 1600 horas).

   Sugerencia: Comience suponiendo un cierto número de componentes, k, por ejemplo $k = 60$, y verifique si con ese número se cumple con lo pretendido. Si se cumple, pruebe con un número menor y si no se cumple, con un número mayor, hasta encontrar el número pedido que está entre 60 y 70.

14. Reconsidere el Ejemplo 6.6, referido a la fabricación de una máquina en 10 etapas (página 316). Utilice herramientas de simulación con `R` y responda a las siguientes consignas:

a) ¿Cuánto vale la probabilidad de que el tiempo total de fabricación (T) sea superior a 36 h?

b) Indique cuánto valen el tiempo promedio de fabricación de la máquina, así como la mediana y los cuartiles.

c) Mencione el valor de la desviación estándar del tiempo total de fabricación.

d) Construya un gráfico con los datos simulados y comente sobre la distribución de T.

e) Compare los valores obtenidos a partir de la simulación con los obtenidos analíticamente.

15. Reconsidere la Actividad 11, relativa al costo de eliminar los poros y las manchas de las 50 barras de cada lote. Utilice herramientas de simulación con `R` para obtener, de manera aproximada, la distribución de dicha variable, así como la probabilidad de que el costo supere $18000 y otros parámetros de interés. Compare los resultados con los de la actividad mencionada.

16. En una empresa interesa decidir la cantidad de artículos de cierto tipo que se debería tener en stock. Por experiencia se conoce que el número de unidades demandadas por día de ese tipo de artículo puede ser 0, 1 o 2 con igual probabilidad. Al momento del estudio, la empresa tiene en depósito 100 unidades de este tipo de artículo y se desea saber si son suficientes para hacer frente a la demanda de los próximos 90 días hábiles. ¿Qué aconsejaría usted? Resuelva el problema aplicando herramientas de simulación.

17. Un fabricante de láminas metálicas de trabajo pesado para camiones comienza obteniendo las piezas fundidas y luego las somete a un proceso de recubrimiento. Ambos procesos se realizan independientemente uno del otro. Una característica de interés es el espesor de las láminas, evaluado al final de ambos procesos. Dicho espesor total se denota por X_T y resulta de la suma de dos variables: el espesor a la salida del proceso de fundición (X_F) y el espesor aportado por el recubrimiento (X_R). Todos los espesores se miden en pulgadas.
    Las especificaciones definidas para el espesor total son: $(0,9 \pm 0,06)$ y el fabricante pretende que como mínimo el 99,73 % de las piezas terminadas las cumplan.
    a) Suponga que es razonable pensar que esta variable se comporta según el modelo Normal. ¿Qué valores deben asumir la media y el desvío estándar para cumplir con las pretensiones del fabricante? Justifique su respuesta.
    b) Por experiencia se conoce que el espesor de las láminas a la salida del horno de fundición (X_F), se comporta según el modelo Normal, con promedio 0,8 y desvío estándar 0,015. Si el espesor del recubrimiento (X_R) es una variable que también se puede considerar distribuida normalmente, ¿qué valores deben tener los parámetros de esta última variable para que se cumplan las pretensiones del fabricante? Justifique. **Sugerencia:** Dado que $X_T = X_F + X_R$, exprese a $E(X_T)$ y a $V(X_T)$ en función de las esperanzas y las variancias de ambas variables y obtenga así los valores pedidos.

18. Se define una variable $T = \sum_{i=1}^{k} X_i$, donde las variables X_i son independientes y están igualmente distribuidas, con la misma distribución que una variable aleatoria X y otra variable $M = kX$.
    a) Obtenga $E(T)$ y $E(M)$.
    b) Obtenga $V(T)$ y $V(M)$.
    c) ¿En qué casos se puede pensar que cada una de estas variables se comporta según el modelo Normal?
    d) En función de lo analizado en los ítems anteriores, ¿puede afirmar que T y M representan a la misma variable? Justifique su respuesta.
    e) Dé un ejemplo de variables aleatorias que puedan definirse como las variables consideradas en esta actividad.

19. Indique si las siguientes afirmaciones son verdaderas o falsas. Justifique su respuesta, especialmente en el caso que esta sea falsa.
    a) La variancia de la suma de k variables aleatorias independientes es la suma de sus variancias.
    b) Siempre se necesita sumar como mínimo 30 variables aleatorias independientes para poder aplicar el Teorema Central del Límite y aproximar la distribución de dicha suma a la Normal.
    c) La esperanza de una resta de variables aleatorias independientes es la resta de sus esperanzas.
    d) La suma de k variables aleatorias independientes siempre se puede aproximar a la distribu-

ción Normal.

e) La suma de 15 variables aleatorias independientes, cada una con distribución Uniforme, se puede aproximar a la distribución Normal por el Teorema Central del Límite.

f) La variancia de la resta de dos variables aleatorias independientes es la resta de sus variancias.

6.7 Estadística con R

En el Capítulo 5 se presentaron diversas funciones que permiten trabajar con distribuciones de probabilidades, facilitando el cálculo de probabilidades y la búsqueda de valores de cierta variable según acumulen o anti-acumulen cierta probabilidad. Para las distribuciones *conocidas* que R incluye en el paquete `stats`, basta incluir `d`, `p` o `q` antes del nombre de la distribución para realizar diferentes cálculos. Así, por ejemplo, `pnorm` se utiliza para calcular la probabilidad acumulada hasta un valor específico bajo una distribución Normal con ciertos parámetros.

En forma similar, precediendo el nombre de la distribución con la letra `r`, es posible simular o extraer valores de una población hipotética donde la variable sigue determinada distribución. Al igual que lo aprendido en la Sección 5.6, aquí cada distribución de probabilidades requiere la especificación de los valores de sus parámetros. A continuación se presentan algunos ejemplos:

Distribución	Ejemplo	Resultado
Normal	`rnorm(n = 1000, mean = 10, sd = 1)`	Genera 1000 valores a partir de una distribución Normal con media 10 y desvío estándar 1.
Uniforme	`runif(n = 500, min = 0, max = 20)`	Genera 500 valores a partir de una distribución Uniforme entre 0 y 20.
Binomial	`rbinom(n = 2000, size = 10, prob = 0.3)`	Genera 2000 valores de una variable con distribución Binomial con $n = 10$ y $\pi = 0,3$.
Poisson	`rpois(n = 300, lambda = 2)`	Genera 300 valores a partir de una distribución Poisson de parámetro $\lambda = 2$.

A partir de los valores simulados para una o más variables, es posible generar los valores de una nueva variable que sea función de ella/s y analizarlos a fin de conocer la distribución de probabilidades de dicha nueva variable.

Suponga que cierta variable aleatoria Y puede definirse como la suma de una variable $X_1 \sim N(10; 0,2)$ y $X_2 \sim N(15; 0,3)$, donde X_1 y X_2 son variables independientes entre sí. Para conocer el comportamiento de Y se pueden generar valores aleatorios para X_1 y X_2 y luego sumarlos para generar los valores de Y. El código a aplicar es el siguiente, donde la primera línea permite seleccionar una semilla de arranque del proceso aleatorio para que los resultados sean reproducibles:

```
set.seed(124)
X1 <- rnorm(n=30000, mean=10, sd=0.2)
X2 <- rnorm(n=30000, mean=15, sd=0.3)
Y <- X1 + X2
```

Otra posible situación podría ser que otra variable Y sea la suma de 100 variables aleatorias independientes, $X_1, X_2, \ldots, X_{100}$, todas con distribución Normal estándar. En este caso, es dificultoso generar las 100 variables de la manera anterior, definiendo cada una de ellas. Entonces, es conveniente emplear un proceso iterativo. A continuación, se conforma una matriz en la que cada columna corresponde a

una variable X_j, con $j = 1,2,\cdot,100$ y en cada repetición del proceso iterativo, se generan 25000 valores para cada X_j. Así, en el primer paso se generan 25000 valores para X_1, en el segundo paso se generan 25000 valores para X_2 y así sucesivamente, hasta completar los valores de X_{100}. Una vez completada la matriz, para obtener los 25000 valores de Y, se suman los valores de cada fila de la matriz generada anteriormente.

```
set.seed(233)
x <- matrix(nrow=25000, ncol=100)

for (j in 1:100) {
    x[,j] <- rnorm(n=25000)
}
y <- rowSums(x)
```

En el siguiente ejemplo, se muestra un problema que requiere la aplicación de simulación para estudiar el comportamiento de ciertas variables, y el posterior análisis del comportamiento de los valores generados.

Ejemplo 6.11 Un laboratorio de Argentina se encuentra interesado en convertirse en líder en el servicio de impresión 3D de piezas para la industria manufacturera. Uno de los productos que ofrece es el separador cilíndrico, un accesorio comúnmente utilizado en el empalme de varillas roscadas. Una empresa de la región requiere separadores cilíndricos de 6 mm de diámetro y 35 mm de largo. El laboratorio se convertirá en su proveedor, enviando estos separadores en cajas de 100 unidades. Para el diseño de la caja, es importante conocer el volumen requerido por las 100 unidades a soportar. Se conoce que el diámetro de los separadores cilíndricos (D) sigue una distribución Normal con promedio 6 mm y desvío estándar 0,02 mm, mientras que el largo de las piezas (L) tiene una distribución Uniforme entre 34,9 y 35,1 mm. Al laboratorio le interesa saber:

1. ¿Qué distribución de probabilidades tiene el volumen de los separadores cilíndricos que requiere la empresa?
2. ¿Qué proporción de los separadores cilíndricos tienen un volumen superior a 1000 mm^3?
3. ¿Qué distribución de probabilidades tiene el volumen total de 100 separadores cilíndricos a envasar en cada caja?
4. Si la caja tiene un volumen de 99100 mm^3, ¿qué proporción de veces 100 separadores cilíndricos no entrarán en la caja?

Resolución:

1. Sea V: volumen del separador cilíndrico, en mm^3. Se conoce que $V = \pi\,(D/2)^2\,L$.
   Para estudiar la distribución de V, se simulan 20000 valores de D y de L, de manera independiente, a partir de sus distribuciones y luego se obtienen los correspondientes valores de V. Al inicio del código, se fija una semilla aleatoria a fin de garantizar la reproducibilidad de los resultados. Los valores obtenidos para la variable V son almacenados en un conjunto de datos denominado *vol*, en una columna de nombre *v*, para su posterior uso:

```
set.seed(276)
d <- rnorm(n=20000, mean=6, sd=0.02)
l <- runif(n=20000, min=34.9, max=35.1)

v <- pi*l*(d/2)^2
vol <- data.frame(v)
```

Una vez obtenidos los 20000 valores simulados del volumen de los separadores cilíndricos, se pueden calcular las correspondientes medidas de resumen:

```
library(tidyverse)
Medidas <- summarise(vol,
                     Mínimo=min(v),
                     Máximo=max(v),
                     Media=mean(v),
                     Cuartil1=quantile(v,probs=0.25),
                     Mediana=median(v),
                     Cuartil3=quantile(v,probs=0.75),
                     Varianza=var(v),
                     Desvío=sd(v))
Medidas <- t(round(Medidas,3))
```

Las mismas resultan:

Tabla 6.10 *Medidas de resumen obtenidas a partir de 20000 valores del volumen de los separadores cilíndricos, suponiendo los modelos* $N(6; 0,02)$ *para el diámetro y* $U(34,9; 35,1)$ *para el largo*

Medida	Valor (en mm^3)
Mínimo	962,621
Máximo	1015,783
Promedio	989,543
Cuartil 1	984,923
Mediana	989,597
Cuartil 3	994,142
Desvío estándar	6,805

Los resultados que se presentan a continuación deben interpretarse como aproximaciones, dado que se obtuvieron a partir de un estudio de simulación basado en 20000 observaciones y bajo el supuesto de validez de los modelos propuestos.

A partir de los resultados de la simulación, y asumiendo que los modelos propuestos para el diámetro y el largo son adecuados, puede afirmarse que el volumen de los separadores cilíndricos fabricados por el laboratorio oscila entre 962,6 mm^3 y 1015,8 mm^3, con un promedio de 989,5 mm^3 y un desvío estándar de 6,8 mm^3. La mitad de las piezas tienen volumen de hasta 989,6 mm^3.

Con los valores simulados, se puede construir un histograma para observar la forma de la distribución de probabilidades del volumen de los separadores cilíndricos:

```
min <- 962.6
max <- 1015.8
intervalos <- 15

#Se genera una secuencia con los valores limites de los intervalos
#partiendo desde min, hasta max
cortes <- seq(min, max, length.out=intervalos+1)

#Se dividen los datos los volúmenes de los cilindros en intervalos
#abiertos por izquierda y cerrados por derecha
division <- cut(as.double(vol$v), cortes)

#Histograma
unit <- quote(bold(mm^3))

ggplot(vol, aes(v)) +
  geom_histogram(aes(y=..count..),
                 color="black", fill="#f9b28c", breaks=cortes) +
  scale_x_continuous(breaks=cortes,
                     labels = scales::label_number(accuracy = 0.1,
                     decimal.mark = ','))+
  #Configuraciones generales
  #Nombres de los ejes
  labs(x = bquote(bold("Volumen de los separadores
  (en ") ~ .(unit) ~ bold(")")), y = "Cantidad de separadores") +
  #Configuraciones de formato
  #Estilo
  theme_classic()+
  scale_y_continuous(expand=c(0,0))+
  #Fuente para los ejes
  theme(axis.text.x =  element_text(colour="black",
                                    size = 7, angle = 25, hjust = 1),
        axis.title.x = element_text(face="bold", colour="black",
                                    size = 12),
        axis.title.y = element_text(face="bold", colour="black",
                                    size = 12))
```

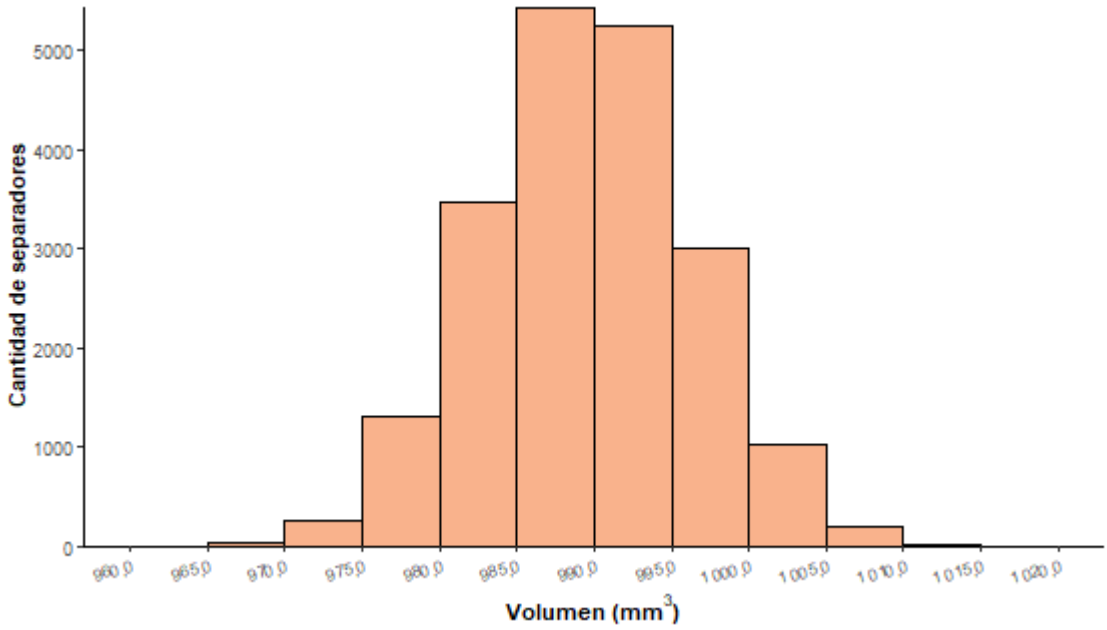


Figura 6.13 *Histograma obtenido a partir de 20000 valores del volumen de los separadores cilíndricos suponiendo los modelos* $N(6; 0,02)$ *para el diámetro y* $U(34,9; 35,1)$ *para el largo*

Se observa que la distribución del volumen de los separadores sigue una distribución campanular y simétrica.

2. Para conocer la proporción de separadores cilíndricos que tienen un volumen superior a 1000 mm^3, se puede calcular la frecuencia relativa de unidades que cumplen esta condición dentro de los valores simulados. Una forma de hacerlo es:

```
prop <- nrow(subset(vol, v > 1000))/nrow(vol)
prop
```

Resulta que la proporción de separadores cilíndricos que tienen un volumen superior a 1000 mm^3 es 0,0623. Es decir, aproximadamente 6 de cada 100 piezas superan dicho volumen.

3. Para encontrar la distribución de probabilidades del volumen total de 100 separadores cilíndricos a envasar en cada caja, se deben generar numerosos valores para la variable T, volumen total de 100 separadores cilíndricos, en mm^3. Para esto, se deben simular en forma independiente los volúmenes de los 100 separadores, de modo que al sumarlos permitan obtener un valor para la variable T. Al repetir este procedimiento 20000 veces, se obtendrán entonces 20000 volúmenes totales.

```
set.seed(123)
vol_caja <- 0
for (i in 1:20000) {
  d <- rnorm(n=100, mean=6, sd=0.02)
  l <- runif(n=100, min=34.9, max=35.1)
  v <- pi*l*(d/2)^2
  vol_total <- sum(v)
  vol_caja <- c(vol_caja, vol_total)
```

```
}
vol_caja <- data.frame(vol_caja[-1])
colnames(vol_caja) <- "Volumen"
```

De esta forma, en el conjunto de datos *vol_caja*, en la columna *Volumen*, quedan almacenados los 20000 valores generados para el volumen total de 100 piezas. Ahora podemos calcular sus medidas de resumen:

```
Medidas <- summarise(vol_caja,
                     Mínimo=min(Volumen),
                     Máximo=max(Volumen),
                     Media=mean(Volumen),
                     Cuartil1=quantile(Volumen,probs=0.25),
                     Mediana=median(Volumen),
                     Cuartil3=quantile(Volumen,probs=0.75),
                     Varianza=var(Volumen),
                     Desvío=sd(Volumen))
Medidas <- t(round(Medidas,3))
```

Tabla 6.11 *Medidas de resumen obtenidas a partir de 20000 valores del volumen total de 100 separadores cilíndricos, suponiendo los modelos* $N(6; 0,02)$ *para el diámetro y* $U(34,9; 35,1)$ *para el largo de cada separador*

Medida	Valor (en mm^3)
Mínimo	98715,009
Máximo	99227,164
Promedio	98960,617
Cuartil 1	98914,989
Mediana	98960,436
Cuartil 3	99006,517
Desvío estándar	67,760

Observe que el volumen total de 100 separadores cilíndricos fabricados por el laboratorio oscila aproximadamente entre 98715 mm^3 y 99227 mm^3, con un promedio aproximado de 98960,6 mm^3 y una desviación estándar de 67,8 mm^3. La mitad de las piezas tienen volumen de hasta 98960,4 mm^3.

Con los valores simulados, se puede construir un histograma para observar la forma de la distribución de probabilidades del volumen total de 100 separadores cilíndricos:

```
min <- 98715
max <- 99228
intervalos <- 15

cortes <- seq(min, max, length.out=intervalos+1)

division <- cut(as.double(vol_caja$Volumen), cortes)
```

```
#Histograma
unit <- quote(bold(mm^3))

ggplot(vol_caja, aes(Volumen)) +
  geom_histogram(aes(y=..count..),
                 color="black", fill="#f9b28c", breaks=cortes) +
  scale_x_continuous(breaks=cortes,
                     labels = scales::label_number(accuracy = 0.1,
                     decimal.mark = ','))+
  #Configuraciones generales
  #Nombres de los ejes
  labs(x = bquote(bold("Volumen total de los 100
  separadores de cada caja(en ") ~ .(unit) ~ bold(")")),
  y = "Cantidad") +
  #Configuraciones de formato
  #Estilo
  theme_classic()+
  scale_y_continuous(expand=c(0,0))+
  #Fuente para los ejes
  theme(axis.text.x =  element_text(colour="black",
                                    size = 7, angle = 25, hjust = 1),
        axis.title.x = element_text(face="bold", colour="black",
                                    size = 12),
        axis.title.y = element_text(face="bold", colour="black",
                                    size = 12))
```

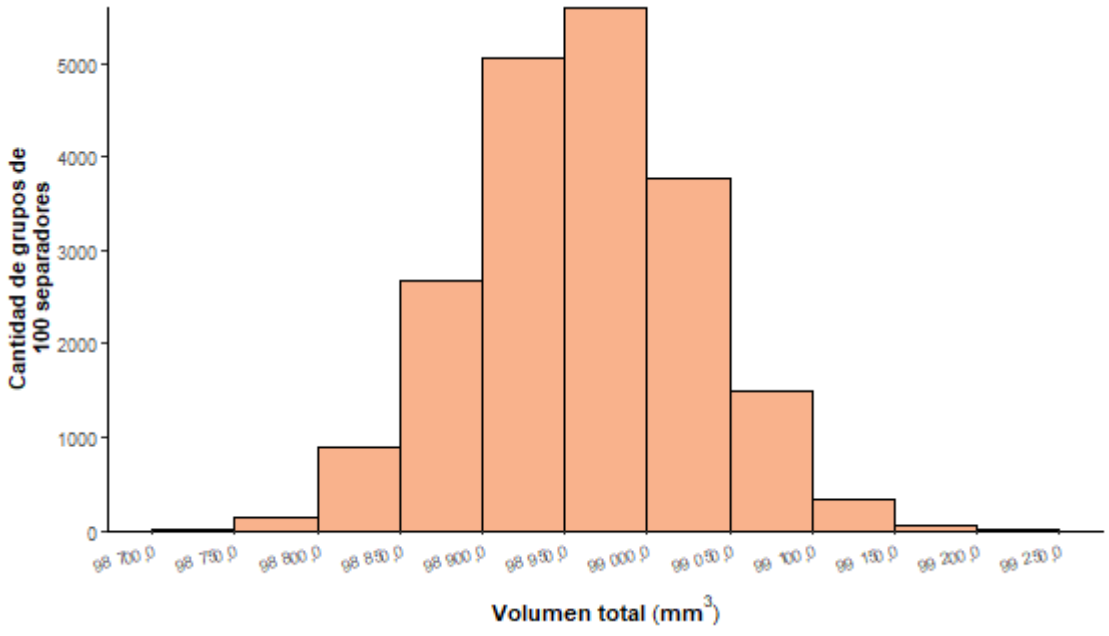


Figura 6.14 *Histograma obtenido a partir de 20000 valores del volumen total de 100 separadores cilíndricos, suponiendo los modelos* $N(6; 0,02)$ *para el diámetro y* $U(34,9; 35,1)$ *para el largo de cada separador*

Se observa que la distribución del volumen total de 100 separadores sigue una distribución campanular y simétrica.

4. Para conocer la proporción de veces que 100 separadores cilíndricos no entrarán en una caja con un volumen de 99100 mm^3, se puede hacer:

```
prop <- nrow(subset(vol_caja, Volumen > 99100))/nrow(vol_caja)
prop
```

Resulta que la proporción de veces que 100 separadores cilíndricos no entrarán en una caja con un volumen de 99100 mm^3 es 0,019.

6.8 Apéndice

6.8.1 Distribuciones de probabilidades conjunta, marginales y condicionales

En esta primera sección del Apéndice se definen y ejemplifican las distribuciones de probabilidades conjunta, marginales y condicionales para vectores bivariados discretos y continuos.

1. Vectores bivariados discretos

La distribución de probabilidades puntual conjunta del vector aleatorio discreto (Y_1, Y_2), $p_{Y_1Y_2}(y_1, y_2)$, es una función que describe el comportamiento de ambas variables simultáneamente. Esta distribución está definida para los pares (y_1, y_2) pertenecientes al recorrido del vector y representa la probabilidad de que la variable Y_1 asuma el valor y_1 y que la variable Y_2 asuma el valor y_2.

$p_{Y_1Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2) \ \ \forall (y_1, y_2) \in R_{(Y_1,Y_2)}$ satisface las siguientes condiciones:

- $p_{Y_1Y_2}(y_1, y_2) \geq 0 \ \ \forall (y_1, y_2) \in R_{(Y_1,Y_2)}$;
- $\sum_{y_1 \in R_{Y_1}} \sum_{y_2 \in R_{Y_2}} p_{Y_1Y_2}(y_1, y_2) = 1$.

A modo de ejemplo se presenta la distribución de probabilidades puntual conjunta del vector bivariado correspondiente al primer vector aleatorio (Sección 6.2).

En este caso, interesa estudiar el comportamiento de la cantidad de poros (W_1) y de la cantidad de manchas (W_2) en forma simultánea. En cada barra del proceso se observan ambas variables. La población está conformada por todas las barras de cierto tipo que produce la empresa metalúrgica. El vector aleatorio es (W_1, W_2).

La distribución de probabilidades conjunta del vector (W_1, W_2) se presenta en la Tabla 6.1 (Sección 6.2) y se repite en este Apéndice, como Tabla 6.12.

Tabla 6.12 *Distribución de probabilidades conjunta de la cantidad de poros (W_1) y de la cantidad de manchas (W_2) por barra*

W_2	W_1				Total
	0	1	2	3	
0	0,3577	0,2628	0,0876	0,0219	**0,73**
1	**0,1323**	0,0972	0,0324	0,0081	0,27
Total	0,49	**0,36**	0,12	0,03	1

Dada la distribución conjunta de las variables de un vector aleatorio bivariado discreto, se puede obtener la distribución de probabilidades puntual para cada una de las variables del vector, también llamada distribución marginal: $p_{Y_i}(y_i) = P(Y_i = y_i) \ \ \forall y_i \in R_{Y_i}$

La distribución marginal o función de probabilidad puntual de cada una de las variables del vector se define de la siguiente manera:

- $p_{Y_1}(y_1) = P(Y_1 = y_1) = \sum_{y_2 \in R_{Y_2}} p(y_1, y_2)$
- $p_{Y_2}(y_2) = P(Y_2 = y_2) = \sum_{y_1 \in R_{Y_1}} p(y_1, y_2)$

Las funciones de probabilidad puntual marginal para cada una de las variables del vector verifican las mismas propiedades enunciadas en la Sección 3.2.2 para las funciones de probabilidad puntual para una única variable aleatoria discreta.

Las distribuciones marginales correspondientes se presentan en las Tablas 6.13 y 6.14 respectivamente.

Tabla 6.13 *Distribución de probabilidades de la cantidad de poros por barra (W_1)*

w_{1i}	$p(w_{1i})$
0	**0,49**
1	0,36
2	0,12
3	0,03
Total	**1**

Tabla 6.14 *Distribución de probabilidades de la cantidad de manchas por barra (W_2)*

w_{2i}	$p(w_{2i})$
0	0,73
1	**0,27**
Total	**1**

A modo de ejemplo, $P(W_1 = 0) = p_{W_1}(0) = 0,49$.

Pensando en una gran cantidad de barras producidas, aproximadamente el 49 % de ellas no presenta poros.

Si se selecciona al azar una barra del proceso, la chance de que la misma no presente poros es 0,49.

Análogamente, $P(W_2 = 1) = p_{W_2}(1) = 0,27$.

Pensando en una gran cantidad de barras producidas, aproximadamente el 27 % de ellas presenta una sola mancha.

Si se selecciona al azar una barra del proceso, la chance de que la misma presente una sola mancha es 0,27.

Observe que la distribución marginal de la cantidad de poros (W_1) (Tabla 6.13) es la que se encuentra en la última fila de la tabla donde se presenta la distribución conjunta (Tabla 6.12) y que la distribución marginal de la cantidad de manchas (W_2) (Tabla 6.14) es la que se encuentra en la última columna. De ese hecho deriva el nombre de "marginales" para estas distribuciones. En la Figura 6.15 se presentan gráficamente las distribuciones marginales correspondientes. Posteriormente, se calculan algunos parámetros de interés.

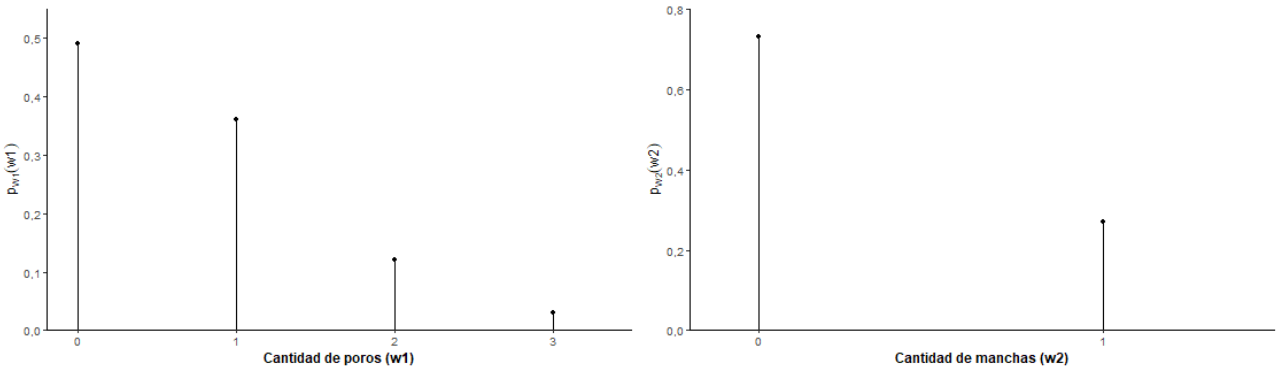


Figura 6.15 *Distribuciones de probabilidades marginales para la cantidad de poros (W_1) y la cantidad de manchas (W_2) por barra*

- $E(W_1) = 0 \cdot 0,49 + 1 \cdot 0,36 + 2 \cdot 0,12 + 3 \cdot 0,03 = 0,69$
- $V(W_1) = [E(W_1^2) - (E(W_1))^2] = (0^2 \cdot 0,49 + 1^2 \cdot 0,36 + 2^2 \cdot 0,12 + 3^2 \cdot 0,03) - (0,69)^2 = 0,6339$
- $D(W_1) = \sqrt{0,6339} = 0,796$

- $E(W_2) = 0 \cdot 0,73 + 1 \cdot 0,27 = 0,27$
- $V(W_2) = [E(W_2^2) - (E(W_2))^2] = (0^2 \cdot 0,73 + 1^2 \cdot 0,27) - (0,27)^2 = 0,1971$
- $D(W_2) = \sqrt{0,1971} = 0,4439$

A partir de la distribución de probabilidades conjunta del vector bivariado discreto, también se pueden definir las distribuciones condicionales, que describen el comportamiento en probabilidad de una de las variables del vector, suponiendo que la otra variable asume un valor dado.

El concepto de probabilidad condicional presentado en el Capítulo 4, para el caso de sucesos, se aplica ahora para obtener la distribución de probabilidades de la variable Y_1, condicionada a que la variable Y_2 toma un cierto valor y_2 y también para obtener la distribución de probabilidad de la variable Y_2, condicionada a que la variable Y_1 toma un cierto valor y_1.

La **distribución condicional** se define de la siguiente manera:

- Distribución de probabilidades de Y_1 condicionada a que Y_2 toma el valor y_2:
  $P(Y_1|Y_2 = y_2) = \frac{P(Y_1, Y_2 = y_2)}{P(Y_2 = y_2)}, \;\; P(Y_2 = y_2) > 0 \; \forall y_1 \in R_{Y_1}$
- Distribución de probabilidades de Y_2 condicionada a que Y_1 toma el valor y_1:
  $P(Y_2|Y_1 = y_1) = \frac{P(Y_1, Y_2 = y_2)}{P(Y_1 = y_1)}, \;\; P(Y_1 = y_1) > 0 \; \forall y_2 \in R_{Y_2}$

Observe que para obtener las distribuciones condicionales, se requiere información de la distribución conjunta y de la distribución marginal.

En este caso en el que se consideran vectores bivariados, las distribuciones de probabilidades condicionales se definen para una de las variables dado cierto valor de la restante, de modo que se trata de distribuciones univariadas. Por ese motivo, verifican las mismas propiedades enunciadas en la Sección 3.2.2 para las funciones de probabilidad puntual.

Suponga que interesa conocer cómo se distribuye la cantidad de poros (W_1) cuando la barra no tiene manchas ($W_2 = 0$), es decir, la distribución de probabilidad condicional $P(W_1/W_2 = 0)$.

Esta función se presenta en la primera fila de la Tabla 6.15 y su expresión es $P(W_1|W_2=0) = P(W_1 = w_1,\ W_2 = 0)/P(W_2 = 0)$.

En la segunda fila, se presenta la distribución de la cantidad de poros cuando la barra tiene una sola mancha, $P(W_1|W_2=1) = P(W_1 = w_1,\ W_2 = 1)/P(W_2 = 1)$.

Tabla 6.15 *Distribuciones de probabilidades de la cantidad de poros (W_1) condicionadas a la cantidad de manchas (W_2) por barra*

W_2	W_1				Total
	0	1	2	3	
0	$0{,}3577/_{0{,}73} = 0{,}49$	$0{,}2628/_{0{,}73} = 0{,}36$	$0{,}0876/_{0{,}73} = 0{,}12$	$\mathbf{0{,}0219/_{0{,}73} = 0{,}03}$	1
1	$\mathbf{0{,}1323/_{0{,}27} = 0{,}49}$	$0{,}0972/_{0{,}27} = 0{,}36$	$0{,}0324/_{0{,}27} = 0{,}12$	$0{,}0081/_{0{,}27} = 0{,}03$	1

$P(W_1 = 3|W_2 = 0) = P(W_1 = 3,\ W_2 = 0)/P(W_2 = 0) = \frac{0{,}0219}{0{,}73} = 0{,}03$.

Pensando en un gran número de barras sin manchas, aproximadamente el 3 % de las mismas presenta 3 poros. Si se selecciona una barra sin manchas, la chance de que la misma presente 3 poros es 0,03.

Análogamente:

$P(W_1 = 0|W_2 = 1) = P(W_1 = 0,\ W_2 = 1)/P(W_2 = 1) = \frac{0{,}1323}{0{,}27} = 0{,}49$.

Pensado en un gran número de barras con una sola mancha, aproximadamente el 49 % de las mismas no presenta poros. Si se selecciona una barra con una sola mancha, la chance de que la misma no presente poros es 0,49.

Las dos distribuciones de probabilidades condicionadas presentadas en la Tabla 6.15 se representan gráficamente en la Figura 6.16. Posteriormente, se calculan las correspondientes esperanzas, variancias y desviaciones estándar.

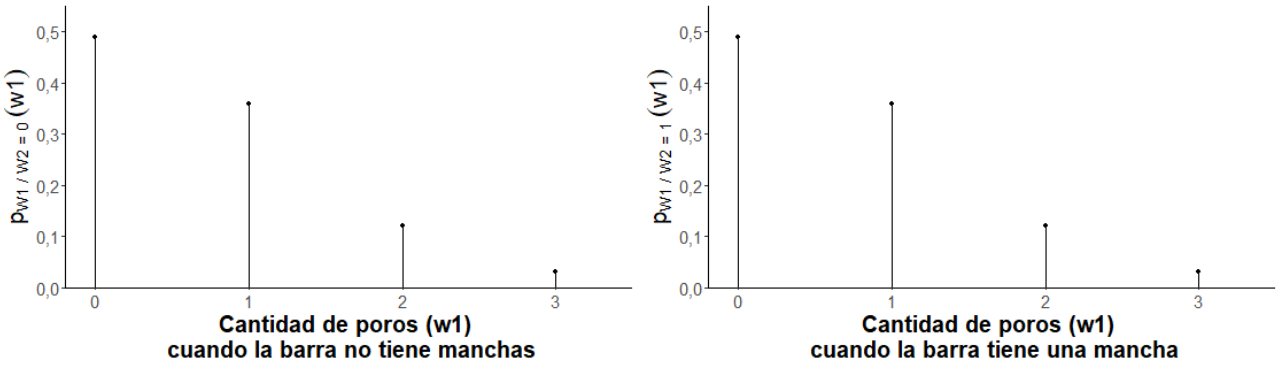


Figura 6.16 *Distribuciones de probabilidades condicionadas para la cantidad de poros (W_1) según la cantidad de manchas (W_2) por barra*

- $E(W_1|W_2=0) = 0\cdot 0{,}49 + 1\cdot 0{,}36 + 2\cdot 0{,}12 + 3\cdot 0{,}03 = 0{,}69$
- $V(W_1|W_2=0) = [E(W_1^2|W_2=0) - (E(W_1|W_2=0))^2] = (0^2\cdot 0{,}49 + 1^2\cdot 0{,}36 + 2^2\cdot 0{,}12 + 3^2\cdot 0{,}03) - (0{,}69)^2 = 0{,}6339$
- $D(W_1|W_2=0) = \sqrt{0{,}6339} = 0{,}796$

- $E(W_1|W_2=1) = 0\cdot 0{,}49 + 1\cdot 0{,}36 + 2\cdot 0{,}12 + 3\cdot 0{,}03 = 0{,}69$

- $V(W_1|W_2=1) = [E(W_1^2|W_2=1) - (E(W_1|W_2=1))^2] = (0^2 \cdot 0,49 + 1^2 \cdot 0,36 + 2^2 \cdot 0,12 + 3^2 \cdot 0,03) - (0,69)^2 = 0,6339$
- $D(W_1|W_2=1) = \sqrt{0,6339} = 0,796$

Observe que las distribuciones de probabilidades de la cantidad de poros para las barras sin manchas $(W_1|W_2=0)$ y para las barras con una mancha $(W_1|W_2=1)$ coinciden entre sí y también coinciden con la distribución marginal del número de poros (W_1) (Tabla 6.13, Figura 6.15). Es decir que, para cualquier número de manchas, la distribución de probabilidades de la cantidad de poros es la misma.

Suponga ahora que la distribución de probabilidades conjunta del vector (W_1, W_2) es la que se presenta en la Tabla 6.16.

Tabla 6.16 *Nueva distribución de probabilidades conjunta de la cantidad de poros (W_1) y de la cantidad de manchas (W_2) por barra*

W_2	W_1				Total
	0	**1**	**2**	**3**	
0	0,47	0,20	0,05	0,01	**0,73**
1	**0,02**	0,16	0,07	0,02	0,27
Total	0,49	**0,36**	0,12	0,03	1

Las distribuciones marginales de las variables W_1 y W_2 son las que se presentan en las Tablas 6.13 y 6.14, respectivamente. En la Tabla 6.17 se presentan las distribuciones de W_1 condicionadas a los diferentes valores de W_2.

Tabla 6.17 *Nuevas distribuciones de probabilidades de la cantidad de poros (W_1) condicionadas a la cantidad de manchas (W_2) por barra*

W_2	W_1				Total
	0	**1**	**2**	**3**	
0	0,644	0,274	0,068	0,014	**1**
1	**0,074**	0,593	0,259	0,074	1

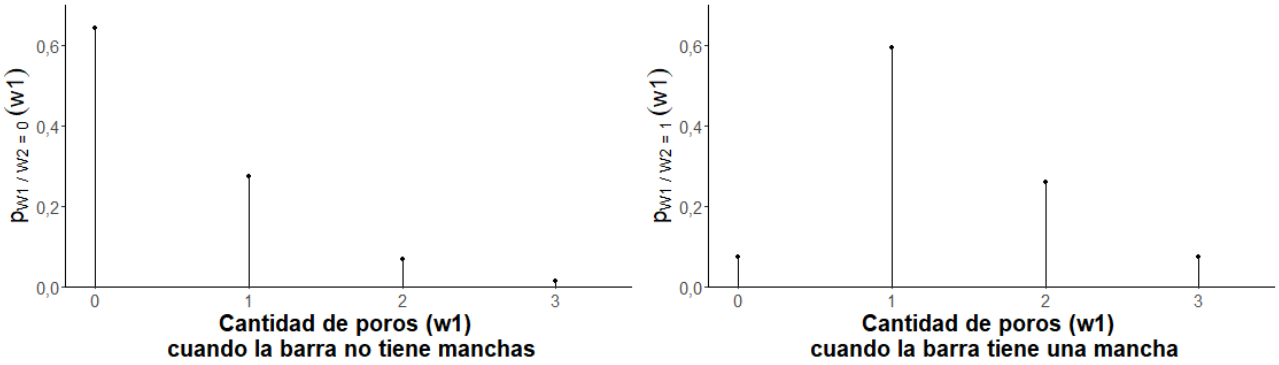


Figura 6.17 *Nuevas distribuciones de probabilidades condicionadas para la cantidad de poros (W_1) según la cantidad de manchas (W_2) por barra*

Observe que si la distribución de probabilidades conjunta de las variables del vector es la que se presenta en la Tabla 6.16, las distribuciones de probabilidades de la cantidad de poros para barras con diferentes cantidades de manchas $(W_1|W_2)$ no coinciden entre sí ni tampoco con la distribución marginal

de la cantidad de poros (W_1) (Tabla 6.17, Figura 6.17). Entre las barras sin manchas predominan claramente las que no tienen poros mientras que entre las barras con una mancha predominan las barras con un poro. La cantidad promedio de poros es mayor cuando hay una mancha que cuando no la hay (1,33 vs 0,452) y la desviación estándar también (0,7198 vs 0,683). En otras palabras, el comportamiento de la cantidad de poros varía con la cantidad de manchas.

En cambio si la distribución de probabilidades conjunta de las variables del vector es la que se presenta en la Tabla 6.12, se observa que el comportamiento de la cantidad de poros por barra (W_1) no varía con la cantidad de manchas de las mismas (W_2).

Esto está asociado con los conceptos de variables relacionadas e independientes, respectivamente, que se formalizan más adelante.

2. Vectores bivariados continuos

La función de densidad de probabilidad conjunta del vector aleatorio continuo (Y_1, Y_2), $f_{Y_1Y_2}(y_1, y_2)$, es una función que describe el comportamiento de ambas variables simultáneamente. Esta función está definida para los pares (y_1, y_2) pertenecientes al recorrido del vector y como en el caso univariado, no brinda directamente las probabilidades: es necesario integrarla para ello.

La función de densidad conjunta $f_{Y_1Y_2}(y_1, y_2)$ verifica las siguientes propiedades:

- $f_{Y_1Y_2}(y_1, y_2) \geq 0 \ \forall (y_1, y_2) \in R_{(Y_1,Y_2)}$
- $\int_{y_1 \in R_{Y_1}} \int_{y_2 \in R_{Y_2}} f_{Y_1Y_2}(y_1, y_2)\, dy_1\, dy_2 = 1$
- $P(a < Y_1 < b,\ c < Y_2 < d) = \int_a^b \int_c^d f_{Y_1Y_2}(y_1, y_2)\, dy_1\, dy_2$

A modo de ejemplo se presenta la función de densidad de probabilidad conjunta del vector bivariado correspondiente al segundo vector aleatorio (Sección 6.2).

En cada barra del proceso se observa el diámetro de la sección transversal (X) y la longitud (Y), ambas en mm.

La población está compuesta por todas las barras de cierto tipo que se producen en la empresa metalúrgica.

El vector aleatorio es (X, Y).

Suponga que se conoce que la función de densidad de probabilidad conjunta para el vector aleatorio es la siguiente:

$$f_{XY}(x,y) = \begin{cases} \frac{1}{5000} & si \qquad 0 \leq x \leq 10;\ 0 \leq y \leq 500 \\ 0 & \text{en otro caso} \end{cases}$$

Para obtener la probabilidad de que las variables del vector aleatorio tomen valores en cierta región, se debe integrar la función de densidad de probabilidad conjunta en dicha región.

Por ejemplo, la probabilidad de que el diámetro de la sección transversal de las barras esté entre 8 y 9 mm y que su longitud sea mayor que 250 mm se obtiene de la siguiente manera:

$P(8 < X < 9, 250 < Y < 500) = \int_8^9 \int_{250}^{500} \frac{1}{5000}\, dx\, dy = 0,05$

Pensando en un gran número de barras, aproximadamente en el 5 % de las mismas el diámetro de la sección transversal está entre 8 y 9 mm y su longitud está entre 250 y 500 mm.

Si se selecciona una barra al azar, la chance de que el diámetro de la sección transversal esté entre 8 y 9 mm y su longitud está entre 250 y 300 mm vale 0,05.

Análogamente al caso discreto se pueden obtener las funciones de densidad de probabilidad marginales para cada una de las variables del vector.

La distribución marginal de cada una de las variables del vector se define de la siguiente manera:

- $f_X(x) = \int_{y \in R_Y} f_{XY}(x,y)\, dy$
- $f_Y(y) = \int_{x \in R_X} f_{XY}(x,y)\, dx$

Las funciones de densidad de probabilidad marginales para cada una de las variables del vector verifican las mismas propiedades enunciadas en la Sección 3.2.1 para las funciones de densidad de probabilidad para una única variable aleatoria continua.

Dada la función de densidad de probabilidad conjunta para el vector aleatorio (Diámetro de la sección transversal, Longitud), $f_{XY}(x,y)$, la función de densidad de cada una de las variables del vector se obtiene integrando la función de densidad conjunta respecto de la restante.

Las dos funciones de densidad de probabilidad marginales en este caso son las siguientes:

$f_X(x) = \int_0^{500} \frac{1}{5000}\, dy = \frac{500}{5000} = \frac{1}{10}\ ;\ 0 < x < 10;$

$f_Y(y) = \int_0^{10} \frac{1}{5000}\, dx = \frac{10}{5000} = \frac{1}{500}\ ;\ 0 < y < 500.$

En la Figura 6.18 se representan las dos distribuciones marginales.

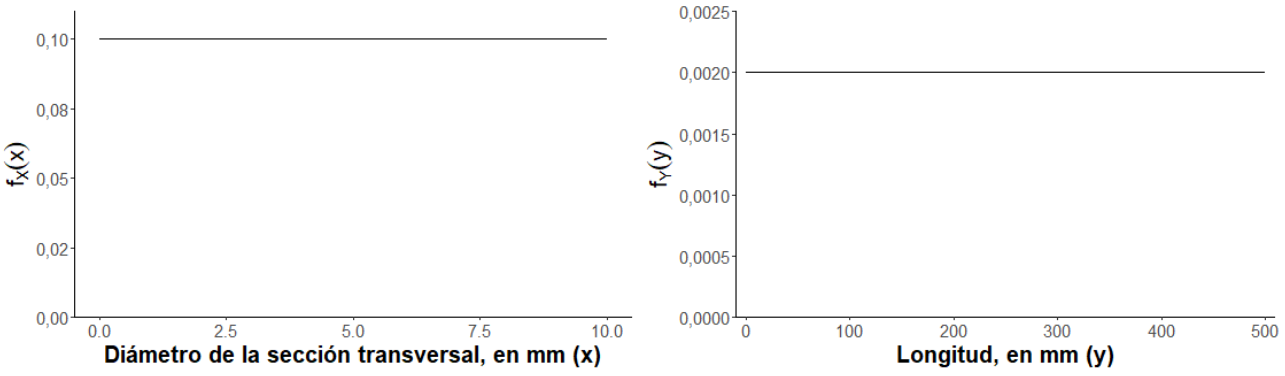


Figura 6.18 *Distribuciones de probabilidad marginales para el diámetro de la sección transversal y la longitud de las barras.*

Las funciones de densidad de probabilidad condicionales de la variable Y_1, condicionada a que la variable Y_2 toma un cierto valor y_2 y de la variable Y_2, condicionada a que la variable Y_1 toma un cierto valor y_1 se definen de la siguiente manera:

Las distribuciones condicionales se definen de la siguiente manera:

- Distribución de probabilidades de Y_1 condicionada a que Y_2 toma el valor y_2:
  $f_{Y_1}(Y_1|Y_2 = y_2) = \frac{f_{Y_1}(y_1, Y_2=y_2)}{f_{Y_2}(y_2)}$; $f_{Y_2}(y_2) > 0$.
- Distribución de probabilidades de Y_2 condicionada a que Y_1 toma el valor y_1:
  $f_{Y_2}(Y_2|Y_1 = y_1) = \frac{f_{Y_2}(Y_1=y_1, y_2)}{f_{Y_1}(y_1)}$; $f_{Y_1}(y_1) > 0$.

En este caso en el que se consideran vectores bivariados, las distribuciones de probabilidad condicionales se definen para una de las variables dado cierto valor de la restante, de modo que se trata de distribuciones univariadas. Por ese motivo, verifican las mismas propiedades enunciadas en la Sección 3.2.1 para las funciones de densidad de probabilidad para una única variable aleatoria continua.

Las funciones de densidad marginales para las variables del vector aleatorio (Diámetro de la sección transversal; Longitud) son:

$$f_X(x|Y = y) = \frac{f_X(x, Y=y)}{f_Y(y)} = \frac{\frac{1}{5000}}{\frac{1}{500}} = \frac{1}{10};\ 0 < x < 10;$$

$$f_Y(y|X = x) = \frac{f_Y(y, X=x)}{f_X(x)} = \frac{\frac{1}{5000}}{\frac{1}{10}} = \frac{1}{500};\ 0 < y < 500.$$

En este ejemplo, entonces, cualquiera sea el valor de X, la función de densidad de Y vale $\frac{1}{500}$ y cualquiera sea el valor de Y, la función de densidad de X vale $\frac{1}{10}$ En otras palabras,

$$f_X(x|Y = y) = 0,1 \ \ \forall y,$$

$$f_Y(y|X = x) = 0,002 \ \ \forall x$$

Los gráficos de las dos funciones de densidad de probabilidad coinciden entonces con los presentados en la Figura 6.18.

Observe que las funciones de densidad de probabilidad de X para cualquier valor de Y coinciden entre sí y coinciden también con la función de densidad marginal de X y lo mismo puede decirse de las funciones de densidad de probabilidad de Y para cualquier valor de X. En otras palabras, el comportamiento del área de la sección transversal no depende de la longitud de las barras y viceversa. Esto está asociado con el concepto de independencia de variables que se formaliza en la segunda sección de este Apéndice.

6.8.2 Variables relacionadas y variables independientes. Covariancia y correlación

En esta segunda sección del Apéndice se define y ejemplifica el concepto de independencia de variables. En el caso de variables relacionadas, se presentan dos medidas que cuantifican la relación entre ambas como la covariancia y el coeficiente de correlación.

Cuando se consideran vectores aleatorios, una cuestión importante es conocer si las variables del vector tienen alguna relación entre sí:

- si las variables no tienen ninguna relación, se dice que son independientes.
- si las variables tienen alguna relación, puede interesar conocer de qué tipo es (lineal, cuadrática, etc.) así como su sentido e intensidad.

En el caso de relaciones lineales entre dos variables del vector, se puede medir la intensidad de esa relación a través de la covariancia y el coeficiente de correlación.

1. Variables independientes entre sí

Las definiciones de independencia entre dos sucesos[7], pueden aplicarse de manera análoga a dos variables aleatorias discretas o continuas.

Dadas dos variables aleatorias discretas Y_1 e Y_2, estas resultan independientes si y solo si:

- $P(Y_1 = y_1 | Y_2 = y_2) = P(Y_1 = y_1) \ \ \forall (y_1, y_2) \in R_{(Y_1,Y_2)}$
- $P(Y_2 = y_2 | Y_1 = y_1) = P(Y_2 = y_2) \ \forall (y_1, y_2) \in R_{(Y_1,Y_2)}$
- $P(Y_1 = y_1,\ Y_2 = y_2) = P(Y_1 = y_1)\,P(Y_2 = y_2) \ \ \forall (y_1, y_2) \in R_{(Y_1,Y_2)}$

Dadas dos variables aleatorias continuas Y_1 e Y_2, estas resultan independientes si y solo si:

- $f_{Y_1}(y_1 | Y_2 = y_2) = f_{Y_1}(y_1) \ \ \forall (y_1, y_2) \in R_{(Y_1,Y_2)}$
- $f_{Y_2}(y_2 | Y_1 = y_1) = f_{Y_2}(y_2) \ \ \forall (y_1, y_2) \in R_{(Y_1,Y_2)}$
- $f_{Y_1Y_2}(y_1,\ y_2) = f_{Y_1}(y_1)\,f_{Y_2}(y_2) \ \ \forall (y_1, y_2) \in R_{(Y_1,Y_2)}$

Observe que para concluir que dos variables son independientes, las igualdades deben verificarse para todos los pares $(y_1, y_2) \in R_{(Y_1,Y_2)}$, de modo que si para un par ordenado cualquiera no se verifica, es suficiente para afirmar que las variables no son independientes.

Cuando se refiere a sucesos, el hecho de que dos sucesos A y B sean independientes indica que la probabilidad de ocurrencia del uno de los sucesos no depende de la ocurrencia del otro. Llevado a variables aleatorias, el concepto de independencia significa que el comportamiento en probabilidad de una de ellas no depende del valor que tome la otra.

En el caso del número de poros y del número de manchas, si se supone que la distribución conjunta es la que se presenta en la Tabla 6.12, se verifica la definición de independencia; lo mismo que para el caso del diámetro de la sección transversal de las barras y su longitud. En cambio, si la distribución conjunta es la que se presenta en la Tabla 6.16, se puede decir que las variables están relacionadas.

[7] Dados dos sucesos A y B, estos son independientes si y solo si:
$P(A|B) = P(A)$
$P(B|A) = P(B)$
$P(A \cap B) = P(A)\,P(B)$
Estas definiciones son equivalentes entre sí y puede usarse cualquiera de ellas para evaluar el cumplimiento de esta propiedad.

2. Variables relacionadas entre sí

Cuando no se verifica la independencia, se dice que las variables están relacionadas entre sí, es decir, que el comportamiento de una de ellas depende del valor que toma la otra.

Ahora bien, ¿de qué tipo es esa relación? ¿qué tan intensa es?

La relación entre variables puede ser de tipo lineal, cuadrática, logarítmica, etc. Cuando la relación es de tipo lineal, se definen dos medidas que dan cuenta del sentido y la intensidad de la relación que se utilizan frecuentemente y que se denominan: covariancia y coeficiente de correlación.

Recuerde que la variancia de una variable aleatoria se define como el promedio o esperanza de los desvíos cuadrados entre cada valor de la variable y la media. Su expresión, presentada en el Capítulo 3, es la siguiente:

$V(Y) = E(Y-\mu_Y)^2 = E(Y^2) - [E(Y)]^2$.

La covariancia considera la variación conjunta de dos variables Y_1 e Y_2, cada una respecto de su promedio. Análogamente se expresa de la siguiente manera:

$Cov(Y_1, Y_2) = E[(Y_1-\mu_{Y_1})(Y_2-\mu_{Y_2})] = E(Y_1 Y_2) - E(Y_1)E(Y_2)$.

Las expresiones de la covariancia para el caso de variables discretas y continuas se presentan a continuación.

$Cov(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2) =$
$\sum_{y_1 \in R_{Y_1}} \sum_{y_2 \in R_{Y_2}} y_1 y_2 \, p_{Y_1 Y_2}(y_1,\, y_2) - \left[\sum_{y_1 \in R_{Y_1}} y_1 \, p_{Y_1}(y_1)\right] \left[\sum_{y_2 \in R_{Y_2}} y_2 \, p_{Y_2}(y_2)\right]$.

$Cov(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2) =$
$\int_{y_1 \in R_{Y_1}} \int_{y_2 \in R_{Y_2}} y_1 y_2 \, f_{Y_1 Y_2}(y_1,\, y_2) dy_1 dy_2 - \left[\int_{y_1 \in R_{Y_1}} y_1 \, f_{Y_1}(y_1) dy_1\right] \left[\int_{y_2 \in R_{Y_2}} y_2 \, f_{Y_2}(y_2) dy_2\right]$.

Determinar la intensidad de la relación lineal a partir del valor de la covariancia es una tarea difícil ya que esta medida puede tomar cualquier valor real, dependiendo del recorrido de las variables involucradas.

Se define entonces otra medida, denominada **coeficiente de correlación lineal** que toma el valor de la covariancia pero lo divide por el producto de los desvíos estándar de las dos variables. Este coeficiente puede asumir valores en el intervalo $[-1, 1]$ e indica que la intensidad de la relación lineal es alta a medida que se acerca a 1 (en valor absoluto) y baja a medida que se acerca a 0.

La expresión del coeficiente de correlación para vectores bivariados discretos o continuos:

$\rho_{Y_1 Y_2} = \frac{Cov(Y_1, Y_2)}{D(Y_1) D(Y_2)}$.

Respecto al sentido de la relación lineal, tanto la covariancia como el coeficiente de correlación toman valores positivos o negativos según la relación sea directa o inversa.

Cuando dos variables aleatorias son independientes, la covariancia, que mide la variación conjunta de dos variables relacionadas linealmente, vale 0, lo mismo que el coeficiente de correlación. Pero que estos coeficientes tomen el valor 0 no implica necesariamente que ambas variables sean independientes entre sí, ya que ambos también pueden tomar valor 0 si las variables están relacionadas pero no de manera lineal.

Y_1 e Y_2 independientes entre sí $\Rightarrow Cov(Y_1, Y_2) = \rho = 0$.

$Cov(Y_1, Y_2) = \rho = 0 \not\Rightarrow Y_1$ e Y_2 independientes entre sí.

Considerando el caso del número de poros y del número de manchas, cuya distribución de probabilidades conjunta está descripta por la Tabla 6.16, en primer lugar se concluyó que ambas no eran independientes entre sí (Tabla 6.17).
Con la información de las distribuciones marginales de ambas variables (Tablas 6.13 y 6.14) se obtiene que:
$E(W_1) = 0,69$ poros, $D(W_1) = 0,796$ poros, $E(W_2) = 0,27$ manchas y $D(W_2) = 0,4439$ manchas.

A partir de los datos de la Tabla 6.16, se obtiene:

$E(W_1 W_2) = 0 \cdot 0 \cdot 0,47 + 0 \cdot 1 \cdot 0,02 + 1 \cdot 0 \cdot 0,20 + 1 \cdot 1 \cdot 0,16 + 2 \cdot 0 \cdot 0,05 + 2 \cdot 1 \cdot 0,07 + 3 \cdot 0 \cdot 0,01 + 3 \cdot 1 \cdot 0,02 = 0,36$.

Observe que cada término se asocia a un par ordenado y constituye el producto del valor de cada una de las dos variables y de la probabilidad conjunta. Por ejemplo, para el par $(2,1)$ que se refiere a barras con dos poros y una mancha, el producto es $2 \cdot 1 \cdot 0,07$.

Con esta información se calculan la covariancia y el coeficiente de correlación entre ambas variables:

$Cov(W_1, W_2) = E(W_1 W_2) - E(W_1) E(W_2) = 0,36$ - 0,69 · 0,27 = 0,1737;

$\rho_{W_1 W_2} = \frac{Cov(W_1, W_2)}{D(W_1) D(W_2)} = \frac{0,1737}{0,7960,4439} = 0,4916.$

En este caso, la covariancia y por ende la correlación son positivas, lo que indica que ambas variables están relacionadas de manera directa: valores bajos de la cantidad de poros se asocian frecuentemente con valores bajos de la cantidad de manchas y valores altos de la cantidad de poros se asocian frecuentemente con valores altos de la cantidad de manchas. El valor del coeficiente de correlación es cercano a 0,50, de modo que puede hablarse de una relación lineal de intensidad intermedia.

6.8.3 Expresiones para la esperanza y la variancia de una combinación lineal de variables aleatorias

En esta tercera sección del Apéndice se deducen las expresiones relativas a la esperanza y la variancia de una combinación lineal de variables aleatorias, mencionadas en la Sección 6.3.1.

Esto se realiza suponiendo que G es una combinación lineal de las variables de un vector aleatorio bivariado continuo. Para el caso discreto se deben reemplazar las funciones de densidad de probabilidad (conjuntas o marginales) por funciones de probabilidad puntuales y las integrales, por sumatorias.

1. Esperanza de una combinación lineal de variables aleatorias

Si $G = a_1 Y_1 + a_2 Y_2$, con $a_i \in \mathbb{R}$, es decir, G es una combinación lineal de variables del vector aleatorio, puede demostrarse que $E(G) = a_1 E(Y_1) + a_2 E(Y_2)$.

$E(G) = E(a_1 Y_1 + a_2 Y_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(a_1 Y_1 + a_2 Y_2) f(y_1, y_2) dy_1 dy_2$

$= a_1 \int_{-\infty}^{\infty} y_1 \left[\int_{-\infty}^{\infty} f(y_1, y_2) dy_2\right] dy_1 + a_2 \int_{-\infty}^{\infty} y_2 \left[\int_{-\infty}^{\infty} f(y_1, y_2) dy_1\right] dy_2$ [8] [9] [10]

$= a_1 \int_{-\infty}^{\infty} y_1 f(y_1) dy_1 + a_2 \int_{-\infty}^{\infty} y_2 f(y_2) dy_2$

$= a_1 E(Y_1) + a_2 E(Y_2)$

2. Variancia de una combinación lineal de variables aleatorias

Esto vale tanto para variables aleatorias continuas como discretas.

$V(G) = V(a_1 Y_1 + a_2 Y_2) = E[(a_1 Y_1 + a_2 Y_2) - (a_1 E(Y_1) + a_2 E(Y_2))]^2$ [11]

$= E[(a_1 Y_1 - a_1 E(Y_1)) + (a_2 Y_2 - a_2 E(Y_2))]^2$

$= E(a_1 Y_1 - a_1 E(Y_1))^2 + E(a_2 Y_2 - a_2 E(Y_2))^2 + 2E(a_1 Y_1 - a_1 E(Y_1)) E(a_2 Y_2 - a_2 E(Y_2))$

$= a_1^2 E[(Y_1 - E(Y_1))]^2 + a_2^2 E[(Y_2 - E(Y_2))]^2 + 2 a_1 a_2 E[(Y_1 - E(Y_1)(Y_2 - E(Y_2))]$

$= a_1^2 V(Y_1) + a_2^2 V(Y_2) + 2 a_1 a_2 Cov(Y_1, Y_2)$

Si las variables del vector son independientes entre sí, su covariancia es nula, y la expresión de la variancia de la combinación lineal se reduce a la siguiente:

$V(G) = a_1^2 V(Y_1) + a_2^2 V(Y_2)$.

6.8.4 Propiedad reproductiva de la suma de variables aleatorias independientes para los modelos Binomial y Poisson

Como se manifestó en la Sección 6.3.2 (página 305) para algunas distribuciones de probabilidades es válida la siguiente propiedad, que recibe el nombre de **"propiedad reproductiva"**: "si dos (o más) variables aleatorias independientes distribuidas según cierto modelo, se suman, la nueva variable aleatoria se distribuye según el mismo modelo que el de los sumandos".

En esta sección se presenta la propiedad reproductiva para el caso de los modelos Binomial y Poisson, en la Tabla 6.18. Note que en el modelo Binomial vale para sumas de variables siempre que la probabilidad de éxito de las mismas no varíe.

[8] $f_{Y_1 Y_2}(y_1, y_2)$ es la función de densidad conjunta de las variables del vector.
[9] Observe que la expresión en el primer [] no es otra cosa que la función de densidad marginal de la variable Y_1.
[10] Observe que la expresión en el segundo [] no es otra cosa que la función de densidad marginal de la variable Y_2
[11] Recuerde que, dada una variable aleatoria Y,$V(Y) = E[Y - E(Y)]^2$

Tabla 6.18 *Propiedad reproductiva para la variable G, suma de k variables aleatorias independientes entre sí*

Distribución de probabilidades para las variables Y_i	**Distribución de probabilidades para la variable G, función de variables aleatorias independientes entre sí**
$Y_i \sim Bi(n_i, \pi)$ (*)	Si $G = Y_1 + Y_2 + \cdots + Y_k$, entonces $G \sim Bi(\sum_{i=1}^{k} n_i, \pi)$ por propiedad reproductiva de la distribución Binomial. Si las k variables tienen la misma distribución, es decir, si $Y_i \sim Bi(n, \pi)\ \forall i$, entonces $G \sim Bi(n\,k, \pi)$
$Y_i \sim Po(\alpha_i)$	Si $G = Y_1 + Y_2 + \cdots + Y_k$, entonces $G \sim Po(\alpha = \sum_{i=1}^{k} \alpha_i)$ por propiedad reproductiva de la distribución de Poisson. Si las k variables tienen la misma distribución, es decir si $Y_i \sim Po(\alpha)\ \forall i$, entonces $G \sim Po(k\,\alpha)$

(*) Observe que π no lleva subíndice porque se requiere que la probabilidad de éxito sea la misma para todas las variables consideradas para aplicar la Propiedad Reproductiva de la Distribución Binomial.

A continuación se presentan algunos ejemplos en los que se aplica la propiedad reproductiva de la distribución Binomial o Poisson, para obtener la distribución de probabilidades de la nueva variable.

Ejemplo 6.12 En una empresa producen envases plásticos para la industria alimenticia. Para un tipo de envase particular, interesa controlar si cada uno presenta o no imperfecciones. Por experiencia se conoce que aproximadamente el 2 % de los envases de ese tipo presentan imperfecciones. Cada día se revisan 50 envases y si se encuentran más de 2 envases con imperfecciones, se considera realizar acciones correctivas en el proceso de producción.

¿Qué chances hay de que deban realizarse dichas acciones a partir del control de un día en particular?

Si se piensa en la evaluación de cada envase, la variable (X_i) lo clasifica según tenga o no imperfecciones asumiendo el valor 1 o 0 respectivamente. Se comporta según el modelo Bernoulli con parámetro $\pi = 0,02$ o según el modelo Binomial, con parámetros $n = 1$ y $\pi = 0,02$.

Pero en realidad, interesa el número de envases con imperfecciones entre los 50 revisados en cada día (Y).

Claramente se puede ver que $Y = X_1 + X_2 + X_3 + X_4 + X_5 + \cdots + X_{50}$.

En esta situación las 50 variables que se suman están igualmente distribuidas y además, es razonable pensar que son independientes entre sí. Entonces, se puede aplicar la propiedad reproductiva del modelo Binomial (Tabla 6.18) para afirmar que Y también se comporta según el modelo Binomial, con parámetros $n = n_1 + n_2 + n_3 + n_4 + n_5 + \cdots + n_{50} = 50 \cdot 1 = 50$ y $\pi = 0,02$.

En resumen:

Sobre las variables del vector aleatorio	Sobre la nueva variable, expresada como función de las variables del vector
Para el envase i, con $i = 1,2,\ldots,50$ X_i: condición del envase (en cuanto a si tiene o no imperfecciones). X_i toma el valor 1 si el envase tiene imperfecciones y 0 si no tiene. Población asociada a X_i: todos los envases. $X_i \sim Be(\pi = 0,02)$ o también $X_i \sim Bi(n_i = 1; \pi = 0,02)$	Para cada control diario (o para cada día) $Y = X_1 + X_2 + \cdots + X_{50}$ indica la cantidad de envases que presentan imperfecciones entre los 50 analizados en un control diario. Población asociada a la variable Y: todos conjuntos de 50 envases (o bien, todos los controles o todos los días, ya que se analizan 50 envases en cada control diario). $Y \sim Bi(n = 50; \pi = 0,02)$

Observe que al definir a la variable Y como la cantidad de envases con imperfecciones entre 50, esta ya puede considerarse distribuida según el modelo Binomial con parámetros $n = 50$ y $\pi = 0,02$, sin necesidad de plantearla como una suma.

Conocido el modelo de la nueva variable, se pueden obtener probabilidades de interés y otros parámetros. En este caso interesa saber la probabilidad de que a partir del control de un día cualquiera se deban realizar acciones correctivas, es decir, interesa obtener la probabilidad de que en el control de un día cualquiera se encuentren más de 2 envases con imperfecciones entre los 50 revisados:

$P(Y > 2) = 0,0784.$

Pensando en un gran número de controles diarios, aproximadamente en el 8 % de ellos se van a encontrar más de 2 envases con imperfecciones en su parte externa, entre los 50 revisados y por lo tanto se van a realizar acciones correctivas en el proceso de producción. O también, si se selecciona un día al azar, la probabilidad o chance de que en el control de ese día se encuentren más de 2 envases con imperfecciones en su parte externa entre los 50 revisados es aproximadamente 0,08.

$E(Y) = 50 \cdot 0,02 = 1$

$V(Y) = 50 \cdot 0,02 \cdot 0,98 = 0,98$

$D(Y) = \sqrt{0,98} = 0,9899$

Pensando en un gran número de controles diarios, en promedio se encuentra 1 envase con imperfecciones y la desviación estándar es aproximadamente 0,98 envases con imperfecciones.

En la Figura 6.19 se presenta la distribución de probabilidades del número de envases con imperfecciones encontrados por control, $Bi(50; 0,02)$, y se señala la probabilidad pedida.

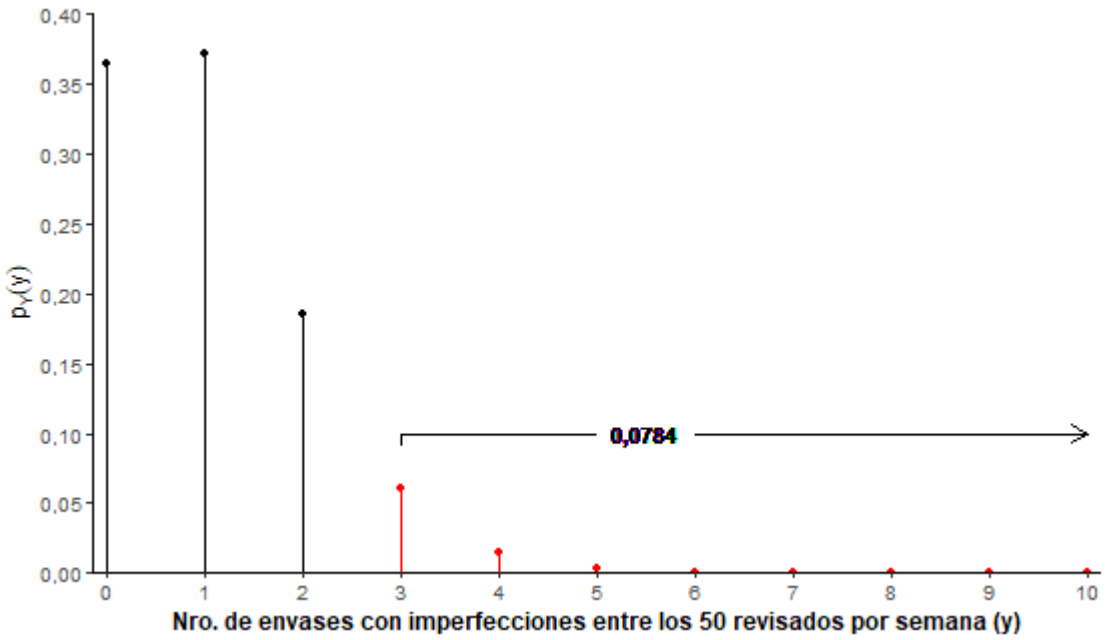


Figura 6.19 *Función de probabilidad puntual de Y, número de envases con imperfecciones entre los 50 revisados por día. La suma de las alturas de los bastones en naranja representa* $P(Y > 2)$

Observe que una variable Y que se comporta según el modelo Binomial de parámetros n y π, se puede definir como la suma de n variables independientes Bernoulli de parámetro π.

Ejemplo 6.13 Una empresa fabrica cable para instalaciones aéreas de telefonía. Entre otras características de interés, en la empresa evalúan la presencia de defectos en la superficie de la aislación. Los defectos pueden ser grietas, agujeros, burbujas y partículas extrañas, los cuales se presentan de manera independiente en los cables. Por experiencia se conoce que la cantidad de defectos de cada tipo en rollos de 100 m de cable sigue una ley de Poisson. La empresa está interesada en vender este tipo de cable para una obra de cableado telefónico de gran envergadura. En relación a los defectos en la superficie de aislación, los compradores admiten como máximo un total de 5 (sin importar de qué tipo) en rollos de 100 m. ¿Qué proporción de los rollos podrán ser comercializados para la obra de cableado? ¿Cuál es el número promedio de defectos cada 100 m de cable? ¿Y la variancia?

En esta situación, el número total de defectos en los rollos de 100 m de cable G_{100}, se puede expresar como la suma del total de grietas, agujeros, burbujas y partículas, las cuales son variables independientes entre sí. $G_{100} = Y_{100,1} + Y_{100,2} + Y_{100,3} + Y_{100,4}$

Como las variables del vector se distribuyen según el modelo Poisson, se puede aplicar la propiedad reproductiva de dicho modelo (Tabla 6.18) para afirmar que G_{100} también se distribuye según la misma distribución con parámetro $\alpha = \alpha_1 + \cdots + \alpha_4 = 2 + 1 + 1,2 + 1,5 = 5,7$.

En resumen:

Sobre las variables del vector aleatorio	Sobre la nueva variable, expresada como función de las variables del vector
Para cada rollo de 100 m $Y_{100,1}$: número de grietas de un rollo. $Y_{100,2}$: número de agujeros de un rollo. $Y_{100,3}$: número de burbujas de un rollo. $Y_{100,4}$: número de partículas de un rollo. Población asociada a cada una de estas variables: todos los rollos de 100 m de cable. $Y_{100,1} \sim Po(2)$ $Y_{100,2} \sim Po(1)$ $Y_{100,3} \sim Po(1,2)$ $Y_{100,4} \sim Po(1,5)$	Para cada rollo de 100 m $G_{100} = Y_{100,1} + Y_{100,2} + Y_{100,3} + Y_{100,4}$ representa el número total de defectos cada 100 m. Población asociada a la variable G_{100}: todos los rollos de 100 m de cable. $G_{100} \sim Po(5,7)$

Conocida la distribución de probabilidades del número total de defectos por rollo de 100 m, G_{100}, resta ahora obtener la proporción de rollos que podrán ser comercializados para la obra de cableado, es decir la probabilidad de que el número total de defectos sea a lo sumo 5.

$P(G_{100} \leq 5) = 0,495$

Pensando en un gran número de rollos de 100 m de cable, en el 49,5 % de dichos rollos se van a encontrar a lo sumo 5 defectos en la superficie de aislación. O también, si se selecciona un rollo de 100 m de cable al azar, la probabilidad o chance de que en dicho rollo se encuentren a lo sumo 5 defectos en la superficie de aislación es aproximadamente 0,495. En números redondos, se puede afirmar que solo la mitad de los rollos de 100 metros de cable podrán ser comercializados por tener a lo sumo 5 defectos en su superficie.

$E(G_{100}) = 5,7$

$V(G_{100}) = 5,7$

$D(G_{100}) = \sqrt{5,7} = 2,387$

Pensando en un gran número de rollos de 100 metros de cable, estos presentan en promedio 5,7 defectos y la desviación estándar es de 2,387 defectos. En la Figura 6.20 se presenta la distribución de probabilidades de la variable G_{100}, es decir, la distribución $Po(5,7)$ y se señala la probabilidad de interés.

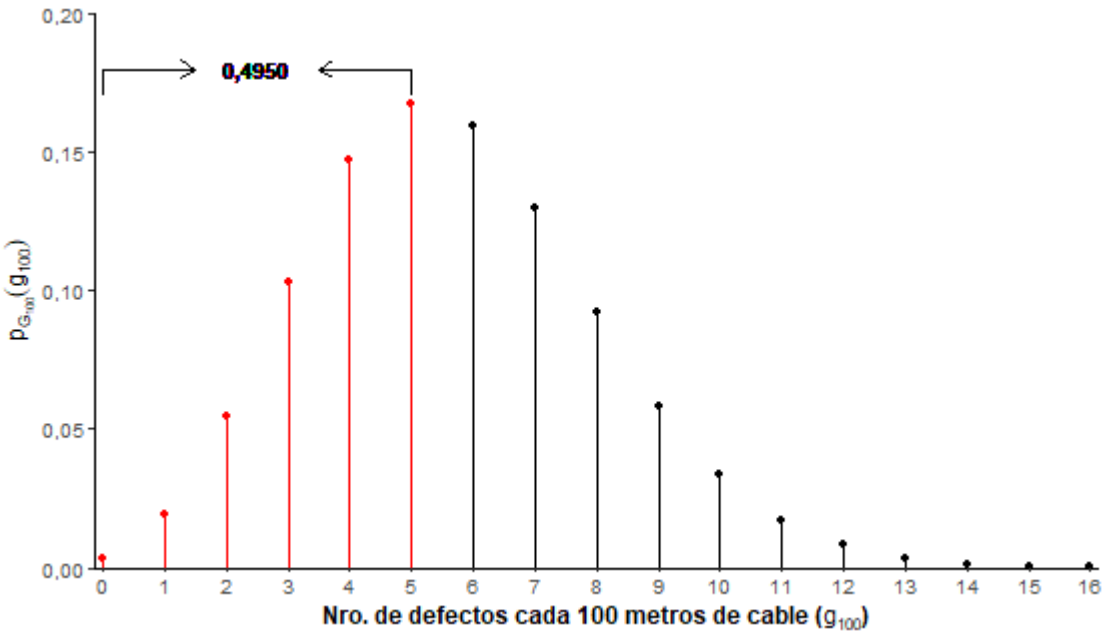


Figura 6.20 *Función de probabilidad puntual de G_{100}, número de defectos cada 100 m de cable. La suma de las alturas de los bastones en naranja representa $P(G_{100} \leq 5)$*



Con lo visto en esta sección, se pueden resolver las Actividades 8 y 9, Sección 6.6.



7. Muestras aleatorias simples y distribuciones muestrales

7.1 Introducción

En el Capítulo 1 se describe el ciclo PPDAC para la resolución de un problema de naturaleza estadística. En la etapa de Planificación del estudio estadístico se decide el tipo de estudio que se va a realizar y, en particular, si se trabajará con una muestra o con toda la población. En caso de optar por una muestra, en la etapa de Análisis de los datos surge la necesidad de realizar análisis inferencial, el cual consiste en la utilización de herramientas que permiten obtener conclusiones sobre la población, a partir de la información de la muestra seleccionada (Sección 1.5.4). Estas herramientas permiten obtener conclusiones con riesgos (de concluir erróneamente) controlados, lo cual se basa en resultados de la Teoría de la Probabilidad y en el estudio de las distribuciones de variables aleatorias, desarrollados en los Capítulos 3 a 6.

En un análisis inferencial, las conclusiones que se obtienen están basadas en la información que brinda una muestra, por ejemplo, en algún caso podrían basarse en el valor obtenido de un estadístico. Como la muestra consiste en un conjunto de unidades seleccionadas de la población, un concepto fundamental que subyace en este contexto es la aleatoriedad que existe al momento de tomar la muestra. Es decir, los datos que se obtienen dependen de cuáles son las unidades de la población seleccionadas y, por ende, los valores de los estadísticos y las conclusiones obtenidas a continuación, también. En este capítulo se estudia lo relativo a cómo puede variar el valor de un estadístico según la muestra que se seleccione de una determinada población.

Los objetivos son:

- Definir el concepto de muestra aleatoria simple.
- Definir los estadísticos que se utilizan como estimadores de los parámetros μ, σ^2 y π.
- Obtener la distribución de probabilidades de dichos estimadores.
- Mencionar algunas propiedades de los estimadores estudiados.

7.2 Muestra aleatoria simple

Un **muestreo probabilístico** es aquel que asigna a cada unidad elemental de la población una cierta probabilidad de ser seleccionada en la muestra. Esto ocurre cuando los elementos que pertenecen a la muestra se seleccionan de manera completamente aleatoria. En algunos casos esta probabilidad es la misma para todas las unidades y en otros más complejos, puede ser diferente. Pero de una forma u otra, las probabilidades asignadas se tienen en cuenta al momento de hacer inferencia para que las conclusiones no sean sesgadas por la propia selección de las unidades.

Existen muchas maneras de obtener muestras probabilísticas; pero en este libro solo se considera el **muestreo aleatorio simple**, el cual refiere a que todas las muestras posibles de n unidades tienen la misma probabilidad de ser seleccionadas.

En una **muestra no probabilística**, en algún momento del proceso de selección de las unidades se deja de lado la aleatoriedad. Podría ser, por ejemplo, que se seleccione un conjunto de n unidades porque son las más cómodas de tomar. Si estas, además, tienen características muy diferentes a otras unidades de la población, la inferencia que se realice resultará sesgada por cómo se eligió la muestra, es decir, ocurre un sesgo de selección.

Se presentan dos ejemplos en los que al tomar la muestra ocurre un sesgo de selección y, por lo tanto, las muestras obtenidas no serían aleatorias simples. Son dos casos donde la forma de tomar la muestra es diferente.

1. Llega un camión con láminas de madera a una empresa y se decide evaluar, mediante una muestra, la calidad de las mismas en función del número de imperfecciones en su superficie. Para ello, se seleccionan 15 láminas. Las láminas presentan dimensiones de 3 m $\times$ 2 m y un espesor de 1 cm y se encuentran apiladas en grupos de 250 unidades. Por razones de comodidad, se eligen las láminas ubicadas en la parte superior.
   Durante el largo viaje, es posible que la presión ejercida sobre las láminas ubicadas en la parte inferior haya provocado que cualquier partícula interpuesta deje marcas en su superficie, efecto que no necesariamente se presenta en las láminas superiores. Por otro lado, las inclemencias del tiempo (sol, lluvia) podrían haber afectado en mayor medida a las láminas situadas en la parte superior de la pila. Ante esta situación, cabe preguntarse si existe un sesgo de selección. La respuesta depende de si la variable de interés está o no relacionada con la ubicación de las láminas dentro de la pila. En este caso, dicha relación es plausible, ya que el número de imperfecciones podría depender de la posición de la lámina. Por lo tanto, la muestra obtenida es sesgada y no permite extender de manera confiable las conclusiones al conjunto total de láminas transportadas en el camión.
   Note que si la variable de interés fuese la longitud de las láminas, no hay por qué pensar que esta esté relacionada con la ubicación. En ese caso, seleccionar las 15 láminas superiores podría interpretarse como una muestra aleatoria del lote.
2. Se quiere conocer la opinión de los clientes de una empresa sobre un servicio de atención telefónica. Para ello se realiza una encuesta voluntaria al finalizar la llamada.
   En este contexto, es probable que entre las personas insatisfechas con el servicio —y, por lo tanto, más propensas a emitir una opinión negativa— exista una mayor predisposición a responder la encuesta y expresar su disconformidad, en comparación con aquellas que no lo están. Esto

indica que la variable de interés está relacionada con la decisión de participar en la encuesta, es decir, con la posibilidad de formar parte de la muestra. En consecuencia, se produce un sesgo de selección y la muestra no puede considerarse una muestra aleatoria simple.

Suponga que, para una población determinada, asociada a una variable aleatoria Y, interesa conocer los valores de algunos parámetros y para ello se va a seleccionar una muestra aleatoria simple de n unidades de dicha población.

Antes de seleccionar cada unidad de la muestra, surge la variable aleatoria asociada al valor de Y que se observará en dicha unidad. En este sentido se define Y_i como el valor que toma la variable Y en la i-ésima unidad de la muestra. Este procedimiento se repite para todas las unidades de la muestra, dando lugar al vector aleatorio $(Y_1, Y_2, \ldots, Y_n)$.

Por la forma en que se definen las variables Y_i, se puede pensar a cada una de ellas como una replicación de Y, es decir, obtener la muestra es realizar n observaciones de la variable Y y entonces, es lógico suponer que la distribución de probabilidades de cada Y_i es la misma que la de Y.

Además, por el hecho de que la elección de cada unidad es aleatoria y resulta independiente de la de las restantes, se deduce que las variables Y_i son independientes entre sí. Formalizando esta idea:

Una **muestra aleatoria simple de tamaño** n (M.A.S.), es un vector de n variables aleatorias Y_i, independientes e idénticamente distribuidas a la variable definida en la población de la que se toma la muestra (Y):

M.A.S: $(Y_1, Y_2, \ldots, Y_n)$, donde:

- Y_i es independiente de Y_j $\forall i \neq j$
- $f_{Y_i}(y) = f_Y(y)$ $\forall i$.

El concepto de muestra aleatoria simple como vector de variables aleatorias aparece antes de seleccionar la muestra; pero una vez que las n unidades se seleccionaron y se les midió la variable, se convierte en un vector de n observaciones o números $(y_1, y_2, \ldots, y_n)$, al que se puede llamar **muestra observada**.

7.3 Parámetros y estadísticos

Un **parámetro** es una magnitud que describe algún aspecto de la población en estudio. Por ejemplo, la proporción de unidades de la población que cumple cierta condición, la media poblacional y el desvío estándar poblacional son parámetros. En muchas situaciones reales no se puede estudiar a la población completa (ya sea porque la población se considera infinita, porque las pruebas son destructivas, por los costos, etc), entonces el valor del parámetro es desconocido y debe ser estimado a partir de la información de una muestra.

Un **estadístico** es una medida que resume la información de la muestra. Si se lo utiliza para estimar un parámetro desconocido, se lo conoce con el nombre de **estimador**. La media muestral, la variancia muestral y la proporción muestral son ejemplos de estadísticos que pueden ser estimadores.

Una vez definida una población, el valor del parámetro es fijo, mientras que el valor del estimador es función de los valores de la muestra seleccionada y por lo tanto puede variar de una muestra a otra. Generalmente, la definición del estimador se inspira en la del parámetro que este pretende estimar, pero cada uno en el conjunto de unidades correspondiente. Por ejemplo, un parámetro puede ser $E(Y)$ (promedio poblacional) y su estimador puede ser $\bar{Y}$ (promedio muestral). Sin embargo, como el valor del estimador depende de la muestra seleccionada, no siempre coincidirá con el del parámetro. Esto significa que, cuando se utiliza información muestral, se puede estar cometiendo algún error al estimar a un parámetro a partir del valor obtenido de un estadístico.

En este capítulo se estudia cómo varían los valores de los estadísticos (o estimadores) pensando en todas las muestras que se podrían seleccionar de una población, es decir, formalmente se estudiarán las distribuciones de probabilidad de los estimadores. En este libro se consideran los parámetros μ, σ^2 y π y sus estimadores media muestral, variancia muestral y frecuencia relativa muestral, respectivamente.

7.4 Distribuciones muestrales

1. Antes de seleccionar la muestra

Los estadísticos son funciones del vector aleatorio $(Y_1, Y_2, \ldots, Y_n)$ y, por lo tanto son también variables aleatorias. En este sentido, cada estadístico está definido sobre el conjunto de todas las posibles muestras de tamaño n que pueden extraerse de la población original.

Por ejemplo, antes de tomar una muestra, se puede definir el estadístico $\bar{Y}$: promedio muestral, el cual se obtendrá a partir de los n valores de la muestra que se vaya a seleccionar.

El estadístico es una variable aleatoria que toma distintos valores según la muestra seleccionada. La población asociada a esta variable está constituida por el conjunto de todas las posibles muestras de tamaño n que pueden obtenerse de la población original, siendo cada una de ellas una unidad elemental.

La **distribución muestral** de un estadístico es su distribución de probabilidades para muestras del mismo tamaño.

2. Después de seleccionar la muestra

Una vez seleccionada la muestra, se obtiene el valor observado del estadístico, el cual es un número que puede servir para dar una estimación del valor del parámetro y realizar la inferencia.



¿Para qué sirve conocer la distribución muestral del estadístico?

Al utilizar un valor observado del estadístico como estimador de un parámetro, puede existir una diferencia entre ambos valores, es decir, el valor del estadístico obtenido a partir de la muestra puede no coincidir con el del parámetro. Esta diferencia se denomina **error de estimación**. El conocimiento de la distribución de probabilidades del estadístico permite establecer una cota para dicho error, tal que la probabilidad de no excederla sea igual a un valor prefijado. Estos temas se abordan con más detalle en el Capítulo 8.

7.4.1 Distribución de probabilidades del promedio o media muestral



Problema 1 (pág. 7): *Considere la situación de la empresa metalúrgica de la provincia de Santa Fe, interesada en estudiar el comportamiento de la longitud de un cierto tipo de barra de acero. Se define en este caso:*

- *Población: todas las barras de acero que se producen en la metalúrgica.*
- *Variable aleatoria: Y: longitud de una barra (mm).*

En alguna determinada situación podría interesar realizar inferencia sobre la media poblacional a partir de una muestra de tamaño n. Un estimador de la media poblacional es la media muestral.
Con el objetivo de aproximarse a la distribución de probabilidades del estimador, se realiza un primer análisis basado en la media muestral. Para ello, se simula la extracción de múltiples muestras de un tamaño fijo, observando los valores que toma este estadístico en cada caso, con el fin de estudiar empíricamente su comportamiento.

- *Población: todas las muestras de n barras de acero.*
- *Variable aleatoria: $\bar{Y}$: longitud promedio de una muestra de n barras (mm).*

Se comienza suponiendo que la longitud de las barras (Y) sigue una distribución Normal con media 250 mm y desvío $0,2$ mm. Luego se obtiene una gran cantidad de muestras de tamaño n (fijo) y para cada muestra se obtiene la media muestral. Finalmente se construye la distribución de frecuencias relativas de la media muestral.

Como esta distribución de frecuencias se construye a partir de una gran cantidad de valores del estadístico, ofrece una buena aproximación de la distribución de probabilidades de la media muestral, tanto en su forma como en sus parámetros.

En primer lugar, se simula una gran cantidad de muestras de tamaño $n = 10$. En la Tabla 7.1 se presentan, a modo de ejemplo, los valores de la media muestral obtenidos para las primeras 14 muestras simuladas. Como se puede observar, la media muestral varía de muestra a muestra y con la distribución de frecuencias obtenida se puede tener una primera idea de cómo lo hace.

Tabla 7.1 *Valores de la longitud promedio muestral para algunas muestras posibles, de tamaño 10, extraídas de la población*

Muestra N°	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Promedio muestral	249,987	250,197	250,061	249,944	250,056	249,979	250,024
Muestra N°	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Promedio muestral	250,033	249,988	249,936	249,994	250,092	249,927	250,047
...	...	...	...	...	...	...	...

En la Figura 7.1 se presentan la distribución de probabilidades de la variable Y: longitud de una barra, y un histograma construido con los promedios de todas las muestras de tamaño 10 seleccionadas. Observe que la distribución de frecuencias de la media muestral tiene un promedio muy próximo al de la media poblacional de la variable Y (250 mm) y un desvío estándar menor al de Y (0,2 mm). Por ejemplo, una observación de Y podría tomar un valor menor a 249,6 mm con una cierta probabilidad, pero para el promedio de 10 observaciones, es casi nula la chance de que eso suceda.

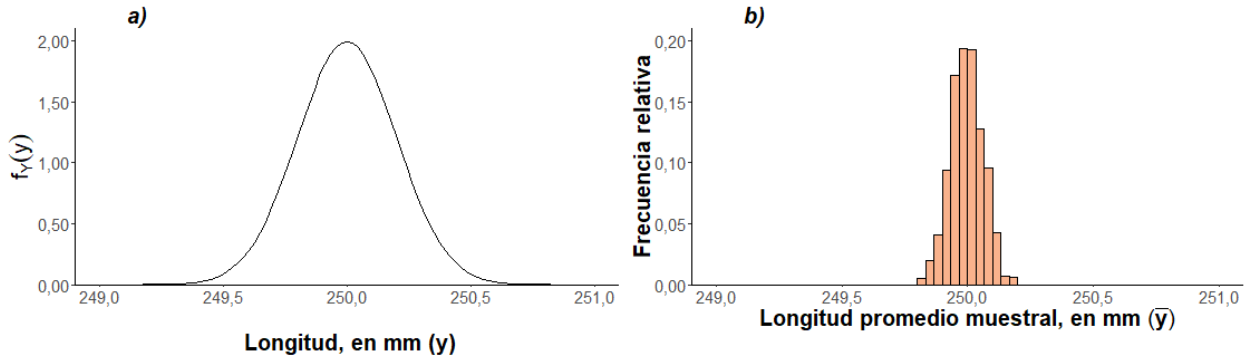


Figura 7.1 *a) Función de densidad de probabilidad de Y, longitud de las barras ; b) Histograma construido con 1000 promedios de muestras de tamaño 10*

De la población de barras se seleccionan ahora muestras de tamaño n = 20, n = 50 y n = 100 y para cada muestra se obtiene el valor de la longitud promedio. También interesa estudiar cómo será la distribución de probabilidades de la media muestral $\bar{Y}$.
Los histogramas construidos con los promedios de todas las muestras de tamaño 10, 20, 50 y 100 respectivamente se presentan en la Figura 7.2. Luego, en la Tabla 7.2 se presentan los valores del promedio y del desvío estándar correspondientes a cada histograma.

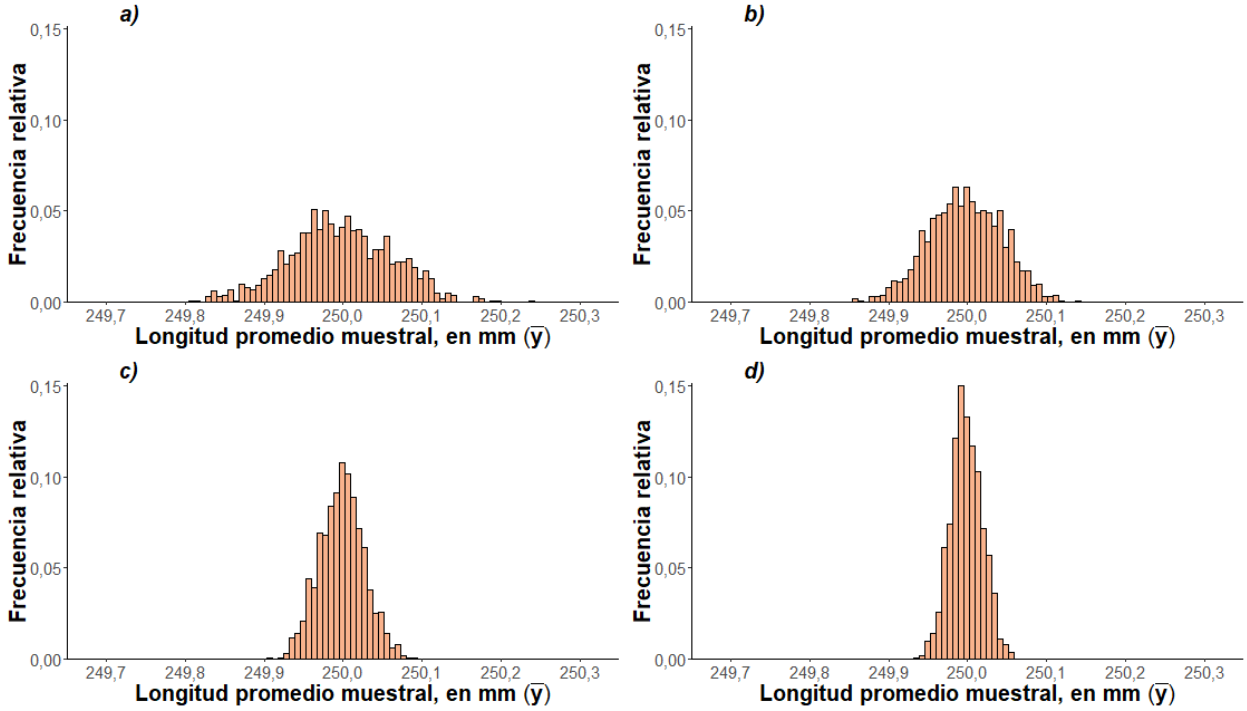


Figura 7.2 *Histogramas construidos con una gran cantidad de promedios de muestras de diferentes tamaños. a) n=10, b) n=20, c) n=50 y d) n=100*

Tabla 7.2 *Promedio y desvío estándar de una gran cantidad de medias muestrales, según el tamaño de la muestra*

Tamaño de la muestra	Promedio	Desvío estándar
n = 10	249,9963	0,0653
n = 20	249,9972	0,0461
n = 50	249,9985	0,0283
n = 100	249,9985	0,0201

En los histogramas anteriores se puede observar que la distribución de frecuencias del promedio muestral para los distintos tamaños de muestra considerados sigue una forma campanular, centrada aproximadamente en el valor del promedio poblacional (250 mm). Además, en la Tabla 7.2 se observa que a medida que el tamaño de muestra aumenta, la variabilidad de dichos promedios muestrales disminuye. Se dice, entonces, que los valores del promedio muestral tienden a concentrarse alrededor del promedio poblacional a medida que el tamaño de la muestra se incrementa.

Todo lo analizado hasta ahora se puede formalizar utilizando las propiedades y teoremas vistos en el Capítulo 6, como se muestra a continuación:

- En una población se define una variable cuantitativa Y con $E(Y) = \mu$ y $D(Y) = \sigma$.
- Se obtienen muestras aleatorias de tamaño n.
- El parámetro de interés es el promedio poblacional, μ.
- Antes de la selección de la muestra, se define el vector aleatorio $(Y_1, Y_2, \ldots, Y_n)$ y, como función de las variables Y_i, se define el estadístico $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$.
- Utilizando las propiedades de la suma de variables aleatorias que se muestran en el Capítulo 6, se pueden expresar los parámetros de la media muestral $\bar{Y}$ en función de los de la variable de interés Y, como se muestra a continuación (ver detalle en Apéndice, Sección 7.8.1):
  - $E(\bar{Y}) = E(Y) = \mu$,
  - $V(\bar{Y}) = \frac{V(Y)}{n} = \frac{\sigma^2}{n}$.
- En cuanto a la distribución de probabilidades de $\bar{Y}$, el Teorema Central del Límite permite afirmar que, para valores de n suficientemente grandes [1], la variable $\bar{Y}$ tiene distribución aproximadamente Normal. Además, por la propiedad reproductiva, si Y tiene distribución Normal, entonces $\bar{Y}$ también tiene distribución Normal para cualquier tamaño de muestra n.
- En síntesis, bajo las condiciones mencionadas, $\bar{Y} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.



Problema 1 (pág. 7): *Retomando la situación de las longitudes de las barras, se supone que esta variable sigue una ley Normal con media 250 mm y desvío 0,2 mm. Se seleccionan muestras de tamaño n = 10 e interesa estudiar el comportamiento en probabilidad de la longitud promedio de las barras de la muestra. Entonces:*

- *Población: todas las muestras de 10 barras.*
- *Variable aleatoria: $\bar{Y}$: longitud promedio de las barras.*

Por ser $\bar{Y}$ una combinación lineal de variables con distribución Normal, $\bar{Y}$ también se distribuye normalmente (Propiedad reproductiva de la distribución Normal), con los siguientes parámetros:

[1] Todo lo relacionado al valor de n suficientemente grande se discute en el Capítulo 6

- $E(\bar{Y}) = E(Y) = 250\ mm$
- $V(\bar{Y}) = \frac{V(Y)}{n} = \frac{0{,}04}{10} = 0{,}004\ mm^2$
- $D(\bar{Y}) = 0{,}0632\ mm$

En síntesis, $\bar{Y} \sim N(250; 0{,}0632)$.

Conocida la distribución de la variable $\bar{Y}$*, se pueden obtener diferentes probabilidades. Por ejemplo, la probabilidad de que la longitud promedio de una muestra aleatoria de 10 barras sea inferior a 249,9 mm, es* $P(\bar{Y} < 249{,}9) = 0{,}0569$

Esta probabilidad, que se presenta en la Figura 7.3, puede interpretarse de las siguientes maneras:

- *para un número grande de muestras de tamaño n (n = 10 barras seleccionadas al azar), aproximadamente el 5,7 % de ellas presenta una longitud promedio por barra inferior a 249,9 mm.*
- *si se selecciona al azar una muestra de n = 10 barras, hay una chance de 0,057 de que la longitud promedio por barra, para dicha muestra, sea inferior a 249,9 mm.*

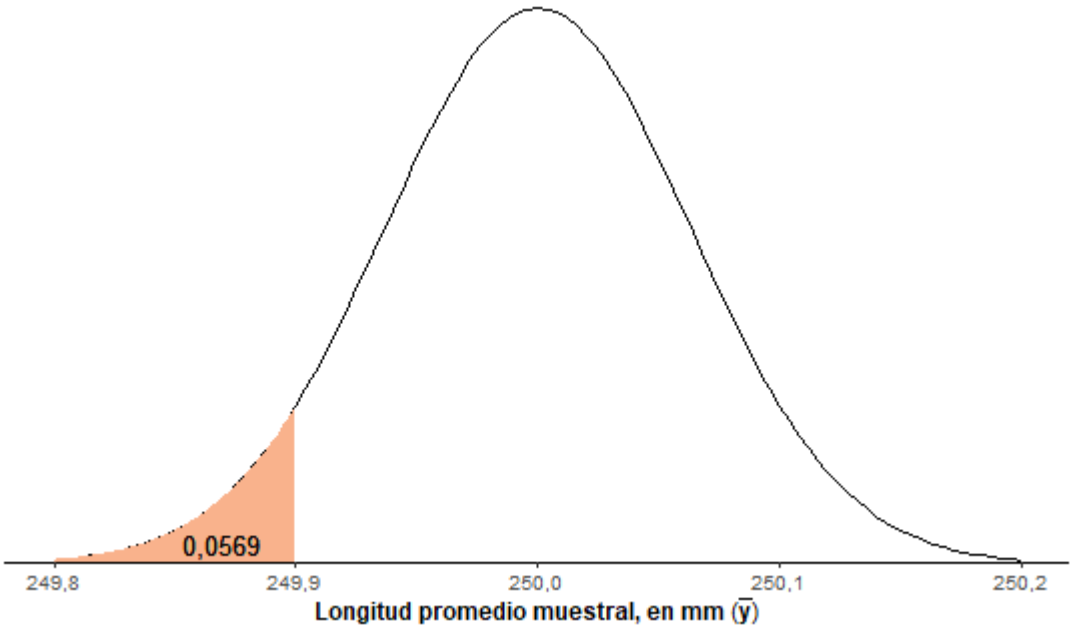


Figura 7.3 *Distribución muestral de la longitud promedio muestral,* $\bar{Y}$*, para muestras de tamaño* $n = 10$*. El área de la región en naranja indica la probabilidad de que* $\bar{Y}$ *tome valores menores que 249,9 mm*

También, a partir de la distribución de $\bar{Y}$ *se puede plantear un intervalo centrado en la media poblacional, en el cual se encuentre, por ejemplo, el 95 % de las medias muestrales (pensando en una gran cantidad de valores posibles de la media muestral):* $P(250 - e < \bar{Y} < 250 + e) = P(\frac{(250-e)-250}{0{,}0632} < Z < \frac{(250+e)-250}{0{,}0632}) = 0{,}95$.

Luego se obtiene que $\frac{e}{0{,}0632} = z_{0{,}975} = 1{,}96$*, que es el percentil de orden 0,975 de la distribución Normal estándar).*
Entonces, $e = 1{,}96 \cdot 0{,}0632 = 0{,}1239$*. El intervalo obtenido,* $(249{,}876; 250{,}124)$*, se presenta en la Figura 7.4.*

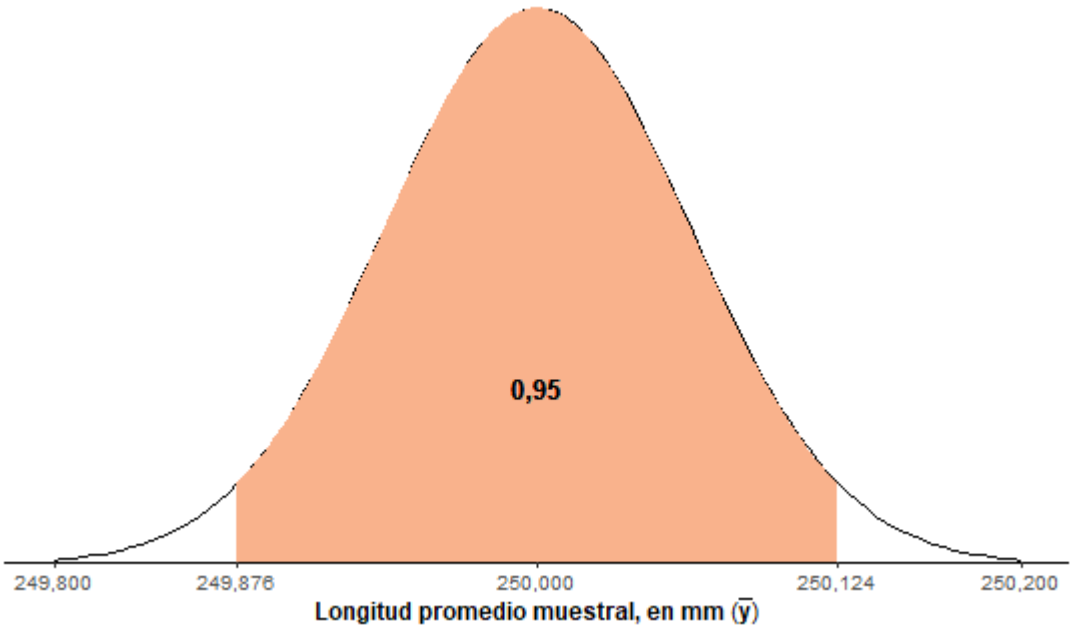


Figura 7.4 *Distribución muestral de la longitud promedio muestral, $\bar{Y}$, para muestras de tamaño n = 10. El área de la región en naranja indica la probabilidad de que $\bar{Y}$ tome valores en el intervalo (249,876 ; 250,124) mm*

Se observa en la expresión anterior que el valor de e (semiampitud del intervalo obtenido) es igual al desvío estándar de la variable media muestral ($\bar{Y}_{10}$) multiplicado por una constante que es el percentil 0,975 de la distribución Normal estándar.

A partir del resultado obtenido, se pueden hacer las siguientes interpretaciones:

- *Para un número grande de muestras de tamaño n ($n = 10$ barras seleccionadas al azar), aproximadamente el 95 % de ellas presenta una longitud media por pieza entre 249,876 mm y 250,124 mm.*
- *Si se selecciona al azar una muestra de $n = 10$ barras, hay una chance de 0,95 de que la longitud promedio por barra, para dicha muestra, esté entre 249,876 mm y 250,124 mm.*
- *Para un número grande de muestras de tamaño n ($n = 10$ barras seleccionadas al azar), aproximadamente el 95 % de ellas presenta una longitud media que se aleja como máximo en 0,1239 del promedio poblacional.*
- *Si se selecciona al azar una muestra de $n = 10$ barras, hay una chance de 0,95 de que la longitud promedio, para dicha muestra, se aleje del promedio poblacional como máximo en 0,1239.*

En estas últimas dos interpretaciones queda de manifiesto la gran utilidad que tiene conocer la distribución del estimador para poder acotar el error de estimación al realizar inferencia, tema que se profundizará en los capítulos siguientes.



Con lo visto hasta ahora, se pueden resolver las Actividades 1 a 7, Sección 7.6.

7.4.2 Distribución de probabilidades de la frecuencia relativa de la muestra

En esta sección se presenta la distribución muestral del estadístico **frecuencia relativa de la muestra** o **proporción muestral**[2].

En la población se define una variable de Bernoulli (Y) que divide a las unidades elementales en dos grupos: las que cumplen una cierta condición y las que no. La variable se define de la siguiente manera:

$$Y = \begin{cases} 1 & \text{si la unidad cumple la condición} \\ 0 & \text{si no.} \end{cases}$$

El parámetro de interés es la proporción, π, de unidades que cumplen la condición en la población, es decir, $\pi = P(Y = 1)$.

Antes de la selección de la muestra, se define el vector aleatorio $(Y_1, Y_2, \ldots, Y_n)$, donde cada variable Y_i está definida para la unidad i-ésima de la muestra del mismo modo que la variable Y. Luego se definen los estadísticos C: cantidad de unidades de la muestra que cumplen la condición y fr: proporción de unidades de la muestra que cumplen la condición, como función de las variables Y_i.

- $C = \sum_{i=1}^{n} Y_i$
- $fr = \frac{C}{n}$

El estadístico C tiene distribución Binomial de parámetros n y π, dado que se cumplen los supuestos correspondientes.

Como $fr = \frac{C}{n}$, fácilmente se puede deducir su distribución. La misma no es Binomial, pero es similar, ya que lo que cambia con respecto a la distribución de C es que los valores de la variable varían entre 0 y 1 en lugar de hacerlo entre 0 y n (es como si se realizara un cambio de escala). Por ejemplo, para $\pi = 0,1$ y $n = 10$, las distribuciones de C y fr se presentan en la Figura 7.5.

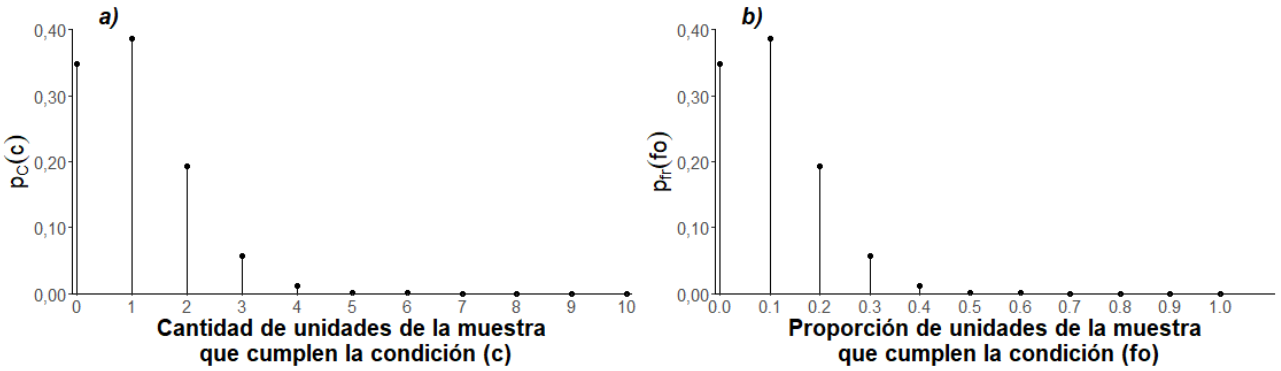


Figura 7.5 *Distribución muestral de la cantidad de unidades de la muestra que cumplen la condición, C (a) y proporción de unidades de la muestra que cumplen la condición, fr, para muestras de tamaño n = 10 (b)*

Dado que $C \sim Bi(n, \pi)$, se puede afirmar que:

- $E(C) = n\pi$,
- $D(C) = \sqrt{n\,\pi\,(1-\pi)}$.

[2]Se trata de la frecuencia relativa o proporción de unidades de la muestra que cumplen cierta condición

Dada la relación entre C y fr y las propiedades de la esperanza y la variancia de una transformación lineal vistas en el Capítulo 3, Sección 3.3.3, se puede afirmar que:

- $E(fr) = \pi$,
- $D(fr) = \sqrt{\frac{\pi(1-\pi)}{n}}$.

Estos resultados se desarrollan en el Apéndice, Sección 7.8.2. Por ejemplo, si $\pi = 0,1$ y $n = 10$, entonces:

- $E(C) = 1$ y $D(C) = \sqrt{0,90} = 0,9486$,
- $E(fr) = 0,10$ y $D(fr) = \sqrt{0,009} = 0,09486$.



Problema 1 (pág. 7): *La empresa automotriz está dispuesta a comprar las barras siempre y cuando sus longitudes se encuentren dentro del intervalo (250 +/- 0,6) mm. En la empresa metalúrgica que fabrica las barras se cree que el 2,5 % de las mismas no cumple dichas especificaciones.*

- *Población: todas las barras.*
- *Condición de interés: la barra no cumple con las especificaciones.*
- *Variable: condición de una barra en relación a las especificaciones (Y).*
  Y toma el valor 1 si la barra no cumple las especificaciones y 0 si las cumple.

Se van a tomar muestras aleatorias de n = 500 barras. En cada barra se va a verificar el cumplimiento de las especificaciones y se va a obtener luego la cantidad y la proporción de barras que no cumplen las especificaciones.

Se quiere obtener la distribución de la proporción muestral (antes de tomar la muestra) de barras que no cumplen con las especificaciones.

- *Población: todas las muestras de tamaño $n = 500$.*
- *Variables aleatorias:*
  - *fr: proporción de barras que no cumplen las especificaciones en una muestra de tamaño $n = 500$ barras.*
  - *C: cantidad de barras que no cumplen las especificaciones en una muestra de tamaño $n = 500$ barras. $C \sim Bi(500; 0,025)$*

Por ejemplo, la probabilidad de que en una muestra de tamaño $n = 500$ se obtenga por lo menos un 3 % de barras que no cumplan las especificaciones es:

$P(fr \geq 0,03) = P(C \geq 15) = 0,2731$.

Esta probabilidad, se puede interpretar de las siguientes maneras:

- *Para un número grande de muestras de tamaño $n = 500$ barras seleccionadas al azar, aproximadamente el 27,31 % de ellas presenta una proporción de barras que no cumplen las especificaciones mayor o igual a 0,03.*
- *Si se selecciona al azar una muestra de $n = 500$ barras, la chance de que el porcentaje de barras que no cumplen las especificaciones para dicha muestra sea superior al 3 % es 0,2731.*

La distribución de la cantidad de barras que no cumplen las especificaciones, entre 500 (C) y la probabilidad obtenida se presentan en la Figura 7.6.

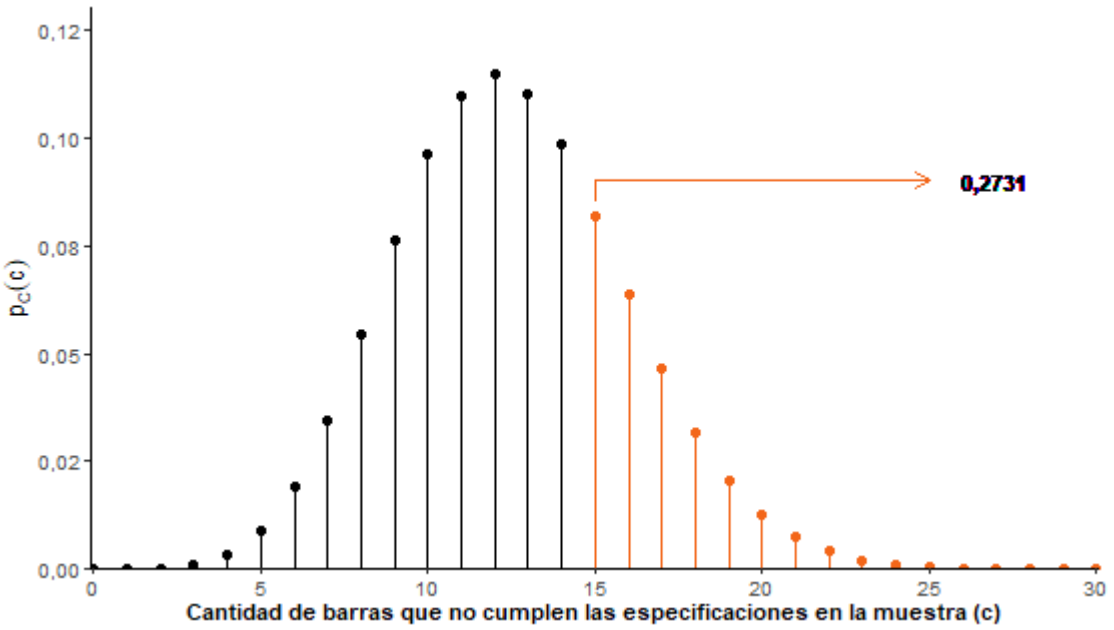


Figura 7.6 *Distribución muestral de la cantidad de barras que no cumplen las especificaciones, C, para muestras de tamaño n = 500. La suma de las alturas de los bastones en naranja indica la probabilidad de que la frecuencia relativa muestral, fr, tome valores mayores o iguales que 0,03 ($P(fr \geq 0,03) = P(C \geq 15) = 0,2731$)*

Como se mencionó en párrafos anteriores, C, que es el numerador de fr, se puede expresar como la suma de n variables independientes e idénticamente distribuidas $Be(\pi)$.
Entonces, **si n es considerablemente grande se puede aplicar el Teorema Central del Límite para aproximar la distribución de fr a una distribución Normal**.

$$fr = \frac{C}{n} \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}})$$

Se sugiere, además de un tamaño de muestra grande, que $n\pi \geq 5$ y $n(1-\pi) \geq 5$, ya que cuando π o $(1-\pi)$ presentan valores muy pequeños, la distribución Bernoulli resulta fuertemente asimétrica y la aproximación basada en el Teorema Central del Límite puede ser poco precisa.

Problema 1 (pág. 7): *Se retoma la situación planteada en relación a la empresa metalúrgica que fabrica las barras, en la cual se cree que el 2,5% de las mismas no cumple con las especificaciones en relación a su longitud.*
Se van a tomar muestras aleatorias de n = 500 barras. En cada barra se va a verificar el cumplimiento de las especificaciones y se va a obtener luego la cantidad y la proporción de barras que no cumplen las especificaciones en la muestra, fr
Dado que el tamaño de muestra es grande (y además $n\pi = 500 \cdot 0,025 \geq 5$ y $n(1-\pi) = 500 \cdot 0,975 \geq 5$), el modelo Normal constituye una buena aproximación en este caso.

$$fr \sim N(0,025; \sqrt{\frac{0,025 \cdot 0,975}{500}}),$$

$$fr \sim N(0,025; 0,007).$$

Observe que la distribución Binomial constituye un modelo exacto para C e indirectamente para fr, que es válido para cualquier tamaño de muestra y para cualquier valor de π; mientras que el modelo Normal no siempre constituye una buena aproximación.

Con lo visto hasta ahora, se pueden resolver las Actividades 8 a 10, Sección 7.6.

7.4.3 Distribución de probabilidades de la variancia muestral

En esta sección se presenta la distribución de probabilidades del estadístico **variancia muestral**. Se parte de una situación similar a cuando se estima una media poblacional, es decir, en una población se define una variable cuantitativa Y, pero ahora el parámetro de interés es la variancia σ^2.

Antes de la selección de la muestra, se define el vector aleatorio $(Y_1, Y_2, \ldots, Y_n)$ y, como función de las variables Y_i, se define el estadístico $S^2 = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n-1}$.

Se puede demostrar que:

- $E(S^2) = \sigma^2$,
- $V(S^2) = \frac{2\sigma^4}{n-1}$.

En cuanto al modelo de distribución de este estadístico, cuando la población de la que se obtuvo la muestra es Normal, se utiliza un modelo que no se presentó en el Capitulo 6, denominado Chi-cuadrado.

Se puede afirmar que una función del estadístico se comporta según este modelo, como se presenta a continuación: $\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ (Ver Apéndice, Sección 7.8.3).

Problema 1 (pág. 7): *Retomando la situación de las longitudes de las barras, se supone que la longitud de las barras sigue una ley Normal con media 250 mm y desvío estándar 0,2 mm. Se seleccionan muestras de tamaño n = 10 e interesa estudiar el comportamiento en probabilidad de la variancia muestral de las longitudes de las barras.*

- *Población: todas las muestras de 10 barras.*
- *Variable aleatoria: S^2: variancia de las longitudes de las 10 barras de una muestra.*

Dado que $Y \sim N(250; 0,2)$, se puede afirmar que el estadístico $\chi^2 = \frac{9S^2}{0,2^2} \sim \chi^2_9$.

Conocida la distribución del estadístico mencionado, se pueden obtener diferentes probabilidades. Por ejemplo, la probabilidad de que la variancia de las longitudes de una muestra aleatoria de 10 barras sea inferior a 0,035 mm^2 es:

$P(S^2 < 0,035) = P(\frac{9S^2}{0,2^2} < \frac{9 \cdot 0,035}{0,2^2}) = P(\chi^2 < 7,875) = 0,4532.$

Esta probabilidad, que se presenta en la Figura 7.7, puede interpretarse de las siguientes maneras:

- *Para un número grande de muestras de tamaño n (n = 10 barras seleccionadas al azar), aproximadamente en el 45,32 % de ellas la variancia muestral de las longitudes es inferior a 0,035 mm^2.*
- *Si se selecciona al azar una muestra de n = 10 barras, hay una chance de 0,4532 de que la variancia muestral de las longitudes, para dicha muestra, sea inferior a 0,035 mm^2.*

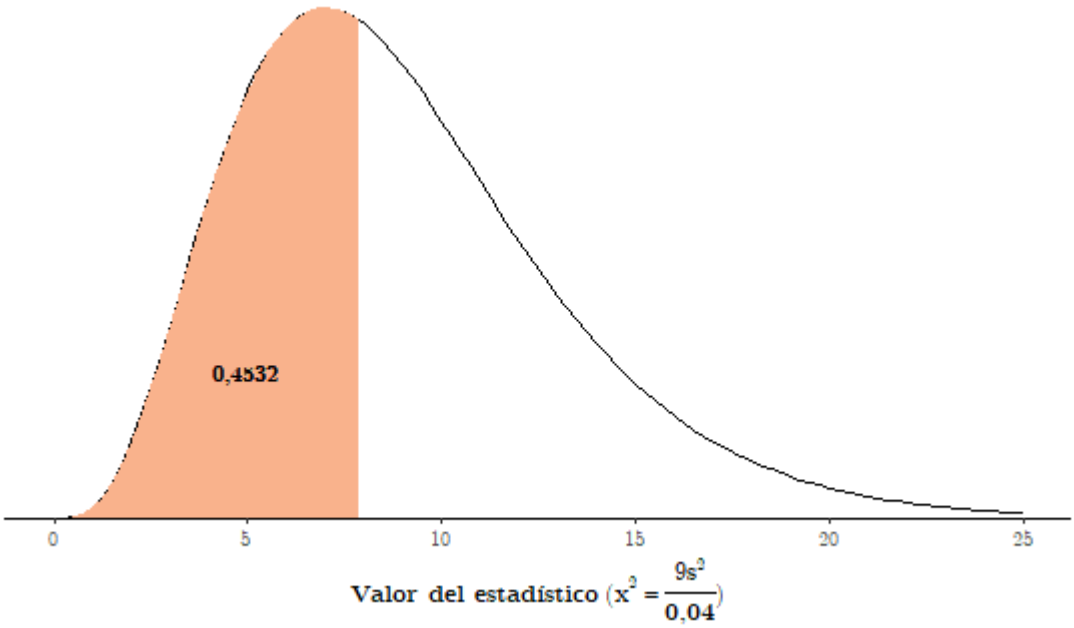


Figura 7.7 *Distribución muestral de* $\frac{9S^2}{0,2^2}$, χ^2 *(Distribución* χ^2_9*). El área de la región en naranja indica la probabilidad de que la variancia muestral tome un valor menor que 0,035 (*$P(S^2 < 0,035) = P(\chi^2 < 7,875) = 0,4532$*)*



Con lo visto hasta ahora, se puede resolver la Actividad 11, Sección 7.6.

Estadísticos pivote: definición y distribuciones

En inferencia estadística es especialmente útil trabajar con estadísticos cuya distribución no depende de los parámetros desconocidos de la población. Estos, llamados **estadísticos pivote**, se introducen y quedan definidos a continuación junto con sus distribuciones, para ser utilizados en los procedimientos inferenciales que se desarrollan en los capítulos posteriores.

- Estadístico $Z = \frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}}$, que es la estandarización de la variable $\bar{Y}$.
  Este estadístico se utiliza para realizar inferencias sobre la media poblacional (μ), cuando se conoce el valor del desvío estándar poblacional (σ).

- Estadístico $T = \frac{\bar{Y}-\mu}{\frac{S}{\sqrt{n}}}$, que es función de la media muestral $\bar{Y}$ y del desvío muestral S.
  Este estadístico se utiliza para realizar inferencias sobre la media (μ) de una población normal, cuando no se conoce el valor del desvío estándar poblacional (σ).

- Estadístico $Z = \frac{fr-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$, que es la estandarización de la variable fr.
  Este estadístico se utiliza para realizar inferencias sobre la proporción de éxitos en la población (π), con la aproximación Normal (no siempre es posible; se prefiere la distribución exacta).

- Estadístico $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$, que es función de la variancia muestral S^2.
  Este estadístico se utiliza para realizar inferencias sobre la variancia (σ^2) de una población normal.

En cuanto a su distribución, se puede afirmar que:

- Si $\bar{Y} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, entonces $Z = \frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$.
- Si $Y \sim N(\mu, \sigma)$, entonces:
  - $T = \frac{\bar{Y}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$ [3] (Apéndice, Sección 7.8.4).
  - $\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ [4] (Apéndice, Sección 7.8.4).
- Si n es suficientemente grande (y, además $n\pi \geq 5$ y $n(1-\pi) \geq 5$), entonces $Z = \frac{fr-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$.

7.4.4 Uso de simulación para la obtención de la distribución de probabilidades de un estimador

Como los estimadores son funciones de las observaciones de la muestra, se puede obtener su distribución mediante simulación siempre que se conozca la distribución de la población. Esto es útil cuando no se pueden aplicar los teoremas o propiedades mencionados en el Capítulo 6, como por ejemplo, cuando el estimador es una función no lineal o cuando se trabaja con muestras pequeñas de una población con distribución no Normal.

Con la ayuda de un software se pueden simular una gran cantidad de muestras de la población conocida y luego obtener el valor del estimador en cada una de ellas y así construir la distribución del mismo.



Problema 8 (pág. 8): *Una empresa se especializa en realizar ensayos de vida en dispositivos electrónicos. Para un dispositivo en particular, llevó adelante un estudio y registró la ocurrencia de fallas y el tiempo transcurrido entre cada una, en horas (Y). Suponga que la distribución de la variable Y es Exponencial con* $\alpha = 0,0028$.

Suponga además que se van a tomar muestras de $n = 8$ *de dichos dispositivos electrónicos, y se desea conocer la distribución de probabilidades del promedio muestral,* $\bar{Y}$.

Con el software `R` *se simulan 50000 muestras de tamaño* $n = 8$ *de la distribución* $Exp(0,0028)$ *y se calcula el promedio en cada una de las muestras. De esta manera, al contar con 50000 valores de esta distribución, puede considerarse que se obtiene por simulación una buena aproximación de la distribución de probabilidades del promedio de todas las muestras posibles de tamaño n = 8. Por lo tanto, se puede considerar a las medidas que se calculen como parámetros de dicha distribución y a las frecuencias relativas como probabilidades. En la Figura 7.8 se presenta un histograma construido con las medias aritméticas de las 50000 muestras simuladas* [5].

[3] t_{n-1} denota la distribución t de Student con $(n-1)$ grados de libertad.
[4] χ^2_{n-1} denota la distribución chi-cuadrado con $(n-1)$ grados de libertad.

El valor medio de las 50000 medias obtenidas resultó ser 356,064 unidades y la desviación estándar 126,454 unidades. A modo de ejemplo, se obtuvo también la proporción de muestras que arrojó un promedio superior a 350 unidades y la misma resultó 0,4718.

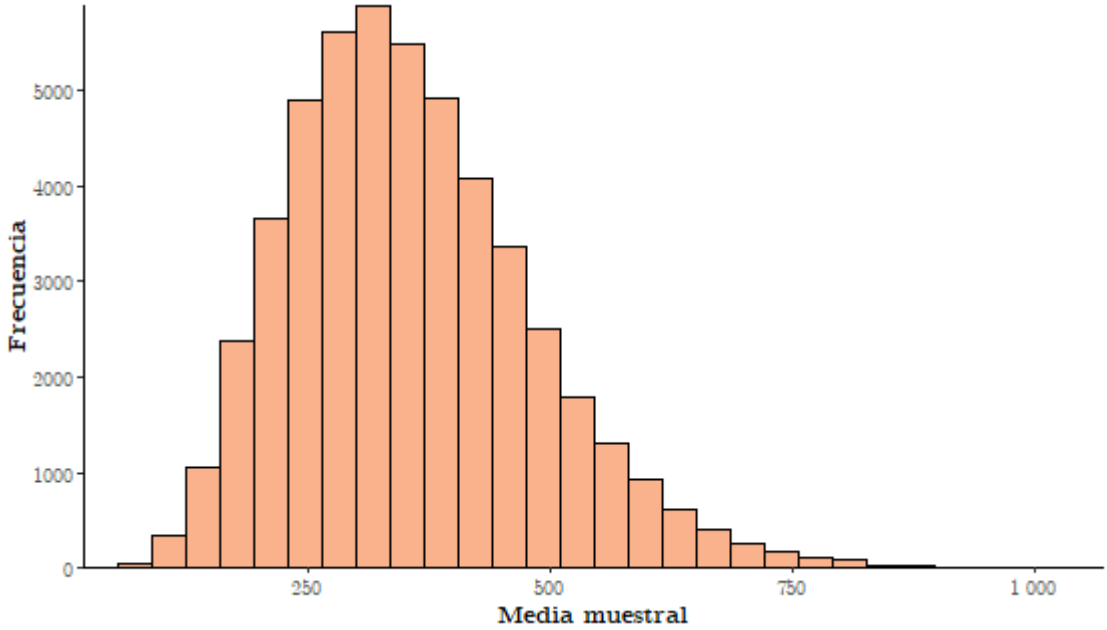


Figura 7.8 *Histograma construido con las medias aritméticas de 50000 muestras de tamaño* $n = 8$ *tomadas de una distribución* $Exp(0,028)$

Con lo visto hasta ahora, se puede resolver la Actividad 12, Sección 7.6.

7.4.5 Otros métodos alternativos para obtener la distribución de probabilidades del estimador

Para los casos en los que no se conoce la distribución de probabilidades de la variable de interés, existen métodos que utilizan la propia información obtenida de la muestra para obtener una distribución, que puede usarse como aproximación de la del estimador.

Uno de estos métodos es el denominado Bootstrap, surgido en la década de 1980, que se basa en algo similar a la simulación. Si bien su desarrollo es amplio, se puede resumir como idea central que luego de tomar una muestra relativamente grande, se considera a esta muestra como una población y se toma una gran cantidad de muestras (denominadas re-muestras) con reposición del mismo tamaño que la muestra original. Calculado el estimador en cada remuestra, se construye la distribución de frecuencias y a partir de la misma, se obtiene la distribución del estimador por simulación considerando que la distribución de frecuencias de la muestra observada es la distribución de probabilidad de la población. En este método se toma como única información a los propios datos obtenidos de la muestra, por eso, no es recomendable aplicar este método cuando se cuenta con una muestra chica.

[5] Los códigos empleados para realizar esta simulación se presentan en la Sección 7.7

7.4.6 Propiedades de los estimadores

En este libro se consideran los parámetros media poblacional (μ), variancia poblacional (σ^2) y proporción poblacional de unidades que cumplen cierta condición (π) y sus estimadores media muestral ($\bar{Y}$), variancia muestral (S^2) y frecuencia relativa muestral (fr).

En la Tabla 7.3 se presentan las expresiones la esperanza y el desvío estándar de cada uno de los estimadores mencionados (Ver Apéndice, Secciones 7.8.1, 7.8.2 y 7.8.3).

Tabla 7.3 *Esperanza y desvío estándar de los estimadores $\bar{Y}$, S^2 y fr*

Parámetro	Estimador	Esperanza del estimador	Desvío estándar del estimador
μ	$\bar{Y}$	$E(\bar{Y}) = \mu$	$D(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$
σ^2	S^2	$E(S^2) = \sigma^2$	$D(S^2) = \sqrt{\frac{2\sigma^4}{n-1}}$
π	fr	$E(fr) = \pi$	$D(fr) = \sqrt{\frac{\pi(1-\pi)}{n}}$

En la tabla se observa, para cada uno de los estimadores considerados, que el promedio de todos los posibles valores coincide con el valor del parámetro correspondiente. También se observa que el desvío estándar de cada estimador disminuye a medida que el tamaño de muestra aumenta. Esto indica que para un tamaño de muestra más grande hay mayor probabilidad de que el valor observado del estimador se acerque más al valor del parámetro. Estos dos resultados fundamentales son los que justifican la utilización de esos estadísticos como estimadores de los parámetros correspondientes.

En la Figura 7.9 se presenta, a modo de ejemplo, la distribución de probabilidades de la media muestral para diferentes tamaños muestrales. Se observa que esta distribución está centrada en la media poblacional y que, a medida que n aumenta, los valores de la media muestral se concentran más alrededor del valor del parámetro de interés ya que la variabilidad disminuye.

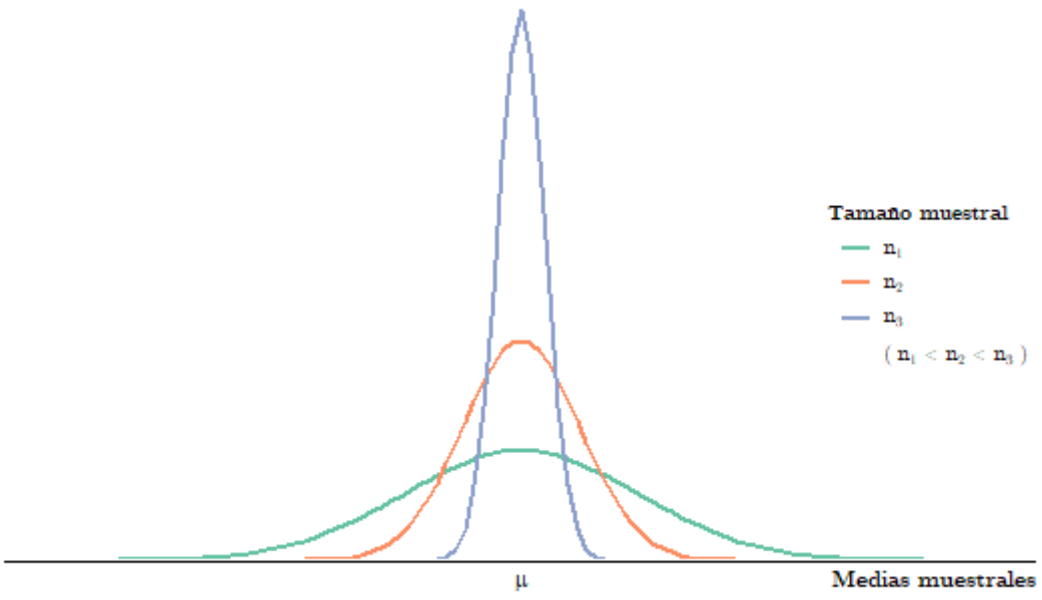


Figura 7.9 *Distribución de probabilidades de la media muestral, $\bar{Y}$, para diferentes tamaños muestrales*

La media muestral es, entonces, un estimador insesgado y consistente de la media poblacional y lo mismo puede afirmarse de los otros dos estimadores considerados en este texto, respecto de sus respectivos parámetros. Ver el Apéndice, Sección 7.8.5, para mayor claridad.



En el Capítulo 2 se introdujo la variancia muestral observada, s^2, calculada con denominador $(n1)$, incorporando así el concepto de **grados de libertad** (página 85).

En este capítulo se estudia el estadístico S^2, asociado al conjunto de todas las posibles muestras de tamaño n. [6]

La utilización del denominador $(n-1)$ en la definición de S^2 permite establecer que dicho estadístico es un estimador insesgado de la variancia poblacional σ^2.

7.5 Síntesis

Cuando en un problema estadístico se trabaja con información de una muestra para obtener conclusiones sobre parámetros de interés, en la etapa de Análisis de Datos se deben aplicar herramientas inferenciales.

Antes de seleccionar las unidades surge el concepto de muestra aleatoria como un vector aleatorio y se deben definir los estadísticos que se van a utilizar como estimadores de los parámetros de interés. Algunos ejemplos de estadísticos son la media muestral, la proporción de unidades de la muestra que cumplen cierta condición y la variancia muestral.

Dado que estos estadísticos se pueden definir como funciones de las variables aleatorias del vector, constituyen variables aleatorias en sí mismos y se pueden derivar analíticamente sus distribuciones de probabilidades, denominadas distribuciones muestrales, aplicando propiedades de funciones del vector aleatorio.

Para obtener estas distribuciones, además de los métodos analíticos, existen alternativas computacionalmente intensivas como la simulación, que puede aplicarse cuando se conoce la distribución de probabilidades de la variable de interés. También existen métodos basados en la información de la muestra observada, como el Bootstrap, que pueden utilizarse en prácticamente cualquier situación.

El conocimiento de la distribución muestral de un estadístico permite evaluar la precisión y confiabilidad de las inferencias, así como cuantificar la incertidumbre asociada, tanto en problemas de estimación como en pruebas de hipótesis.

[6] Recuerde que el estadístico S^2 es una variable aleatoria y está definido sobre el conjunto de todas las posibles muestras de tamaño n que pueden extraerse de la población original; mientras que s^2 es un número que representa al valor observado del estadístico para una muestra dada.

7.6 Actividades propuestas

1. Un analista toma una muestra de 100 piezas y encuentra que el diámetro promedio de dicha muestra es 2,15". Otra analista toma otra muestra de 100 piezas y encuentra que el diámetro promedio es 2,06".
   El primer analista afirma alguno de los dos cometió un error porque los promedios fueron diferentes. Usted, ¿qué opina? Explique.

2. Reconsidere el Problema 6 (Sección 1.4, página 7). Suponga que se conoce que el espesor promedio en un tramo de ruta de 5 km es 22 cm y el desvío estándar es 0,4 cm. A partir de muestras aleatorias de 50 puntos de dicho tramo, se define $\bar{Y}_{50}$ como el espesor promedio muestral.
   a) ¿Cómo se distribuye $\bar{Y}_{50}$? Justifique su respuesta.
   b) ¿Cuál es la probabilidad de que el espesor promedio de una muestra seleccionada al azar sea superior a 22,1 cm?
   c) Suponga que en vez de tomar muestras de $n = 50$ tramos, se consideran muestras de $n = 100$. Indique si las siguientes afirmaciones son verdaderas o falsas. Justifique.
      - El promedio de $\bar{Y}_{100}$ coincide con el promedio de $\bar{Y}_{50}$.
      - La desviación estándar de $\bar{Y}_{100}$ es mayor que la desviación estándar de $\bar{Y}_{50}$.
      - Los valores de $\bar{Y}_{100}$ están más concentrados alrededor de $\mu = 22$ cm.
   d) Grafique, de manera comparativa, las distribuciones de $\bar{Y}_{50}$ y de $\bar{Y}_{100}$

3. En una empresa producen piezas metálicas de cierto tipo. La longitud (en milímetros), Y, de dichas piezas se distribuye normalmente, con promedio 10 y desvío estándar 1. Periódicamente se seleccionan muestras aleatorias simples de n piezas y se define $\bar{Y}$, longitud promedio muestral. Sobre esta última variable se obtuvo la siguiente información con R:

```
pnorm(10.3, 10, 0.5)
```
0.726

```
qnorm(0.975, 10, 0.5)
```
10.98

```
pnorm(11,10,0.5, lower.tail = F)
```
0.023

```
qnorm(0.99, 10, 0.5, lower.tail = F)
```
8.84

   a) Indique cuánto vale el tamaño de las muestras, n. Justifique.
   b) ¿Cómo se distribuye $\bar{Y}$? ¿Por qué?
   c) Exprese formalmente las probabilidades obtenidas con `R`. Interprételas en términos del problema.
   d) Exprese formalmente a los percentiles obtenidos con `R`. Interprételos en términos del problema.

e) Indique si las siguientes afirmaciones son verdaderas o falsas. Justifique.
   1) $P(\mid \bar{Y} - 10 \mid < 2) \approx 0,95$
   2) Es prácticamente imposible obtener una longitud promedio menor o igual que 8 mm.

4. El tiempo (en minutos) que demanda el armado de cierto tipo de piezas (X) se distribuye normalmente con media 150 y desviación estándar 4. Se seleccionan muestras de n = 16 piezas e interesa el tiempo promedio de armado para cada muestra.
   a) ¿Cómo se distribuye la variable $\bar{X}_{16}$? Justifique.
   b) ¿Podría indicar, de manera aproximada, cuáles son los valores de $\bar{X}_{16}$ más alejados de $\mu = 150$ min? Justifique su respuesta y presente un esquema apropiado.
   c) Responda a la pregunta anterior suponiendo que ahora el tamaño de las muestras es de $n = 25$ piezas. ¿Qué cambios observa?
   d) Si se desconociera la distribución del tiempo de armado de este tipo de piezas, ¿en qué casos se puede utilizar en forma aproximada a la distribución Normal para modelar el comportamiento del tiempo de armado promedio de muestras de n piezas?

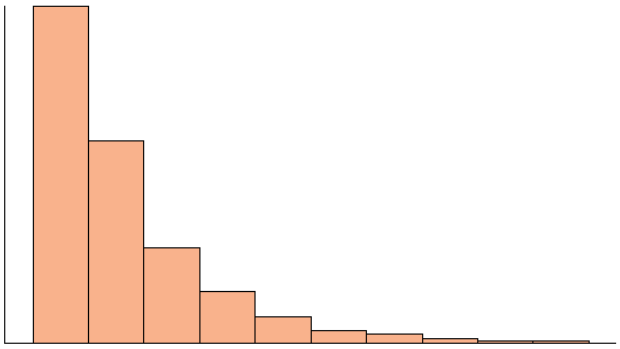
5. Un fabricante de tubos para cableado eléctrico afirma que el diámetro exterior (en pulgadas) de los mismos, X, tiene una media de 1,30, con un desvío de 0,031. Se toman muestras de 10 tubos.
   a) ¿Cuál es la distribución de probabilidades del estadístico "diámetro exterior promedio" (en ") para muestras de 10 tubos? Suponga que se conoce que $X \sim N$.
   b) ¿Cuál es la probabilidad de que el diámetro exterior promedio para una muestra de 10 tubos se encuentre, en valor absoluto, a menos de 0,02 unidades del valor 1,30"?
   c) Aproximadamente el 99 % de las muestras de 10 tubos arrojan un valor promedio que se encuentra en el intervalo (1,30 +/- a)". Obtenga el valor de a (exprese formalmente) y represente el intervalo obtenido en el gráfico de la distribución correspondiente.

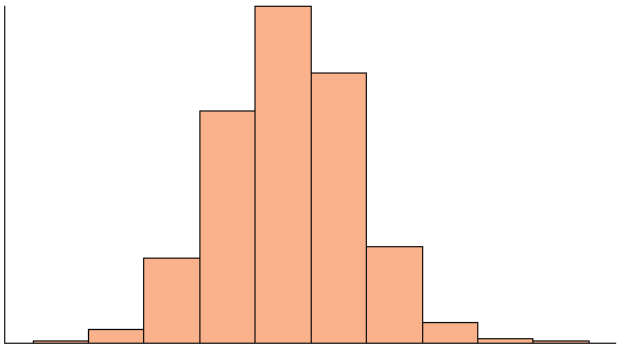
6. Considere una máquina expendedora de café. Para monitorear el desempeño de la misma se mide, entre otras variables, el volumen vertido por vaso, V. Una persona afirma que, si se toma una muestra suficientemente grande de vasos de café, esta variable puede considerarse distribuida normalmente, por el Teorema Central del Límite.
   a) ¿Cuál de las siguientes opciones es la correcta?
   A: La afirmación es correcta, porque el Teorema Central del Límite establece que toda variable aleatoria se distribuye aproximadamente Normal cuando el tamaño muestral es grande.
   B: La afirmación es incorrecta, porque el Teorema Central del Límite solo se aplica cuando la variable original V tiene distribución Normal.
   C: La afirmación es incorrecta, porque el Teorema Central del Límite no describe el comportamiento de una variable individual, sino el de la suma (o media) de las n variables aleatorias asociadas a la muestra.
   D: La afirmación es correcta siempre que el tamaño de la muestra sea suficientemente grande y no haya valores atípicos en los datos.
   b) Para la opción correcta, indique qué requerimientos deben verificarse para que la aplicación del teorema sea válida.
   c) Si se conociera que V se comporta normalmente, ¿es necesario aplicar el Teorema Central del Límite para afirmar que $\bar{V}$ también lo hace? Justifique.
   d) Sobre las distribuciones de las variables V y $\bar{V}$, indique si las siguientes afirmaciones son

verdaderas o falsas. Justifique.

1) Si $V \sim N$, entonces $\bar{V} \sim N$, para cualquier tamaño de muestra.
2) $E(V) = E(\bar{V})$.
3) Si V no se comporta según el modelo Normal, entonces $\bar{V}$ tampoco lo hace.
4) $D(\bar{V}) < D(V)$.
5) A medida que aumenta el tamaño de la muestra, la distribución de V se concentra más alrededor de su media.

7. Suponga que se toma una gran cantidad de muestras de un tamaño considerablemente grande de una población compuesta por piezas metálicas. En cada pieza se mide el tiempo que demanda su elaboración (Y), en minutos. Suponga que se conoce que $Y \sim Exp(\alpha)$.
   A- Para una de las muestras obtenidas, se construyó un histograma que describe el comportamiento del tiempo de elaboración registrado para las piezas de dicha muestra.
   B- Con la información del tiempo promedio de cada una de las muestras obtenidas, se construyó otro histograma que describe el comportamiento de esta variable.
   a) Para las situaciones A y B, indique variable y población.
   b) Indique cuál es el histograma correspondiente en cada caso. Justifique exhaustivamente.


Figura a)


Figura b)

8. Reconsidere nuevamente el Problema 6 (Sección 1.4, página 7). Suponga que se conoce que el 2 % de los puntos de un tramo de ruta de 5 km tienen una resistencia a la compresión menor que 30 MPa. Alguien afirma entonces que en una muestra de 100 puntos seleccionados al azar se encontrarán exactamente 2 con resistencia menor a 30 MPa. ¿Qué opina Ud. de dicha afirmación? Justifique su respuesta.

9. En una industria se realizan controles periódicos sobre la proporción de piezas que no cumplen con las especificaciones, en base a muestras de tamaño $n = 400$.
   a) Determine la distribución de probabilidades del número de piezas defectuosas en la muestra, suponiendo que el proceso presenta un 5 % de piezas fuera de especificaciones. Represéntela gráficamente.
   b) Calcule la probabilidad de que en una muestra se obtenga por lo menos un 6,4 % de piezas fuera de especificaciones. Señale esta probabilidad en el gráfico construido anteriormente.

10. Reconsidere la situación del Problema 4 (Sección 1.4, página 7), referido a la comercialización de productos en línea. La empresa envía un número muy grande de pedidos y se puede pensar que en el último tiempo el 3 % de los pedidos fueron devueltos (por cualquier motivo). Se van a tomar muestras aleatorias de n pedidos y en cada una se va a registrar la proporción de pedidos devueltos (fr).

Se obtuvo la siguiente información con R:

```
pbinom(0,50, 0.03)
0.218

pbinom(1,50, 0.03)
0.555

pbinom(2,50, 0.03)
0.811
```

a) Defina la población y la población estadística asociadas a la variable fr.

b) Indique cuál es el tamaño de las muestras de pedidos y luego informe cómo se distribuye el número de pedidos devueltos en muestras de tamaño n, C, suponiendo que la proporción de pedidos devueltos en la población es 0,03. Esquematice la distribución de probabilidades de C y de fr.

c) Utilice la información obtenida con R para indicar cuánto vale lo siguiente:

1) Proporción de muestras en las cuales la proporción de pedidos devueltos es mayor que 0,035.
2) Chance de que en una muestra de n pedidos se encuentre entre un 2 % y un 4,2 % (ambos incluidos) de pedidos devueltos.
3) $P(fr < 0,01)$.

Señale los valores obtenidos en el gráfico y dé otra interpretación de los mismos.

11. Un taller realiza una gran cantidad de reparaciones de diferentes tipos de máquinas. Para un tipo de máquina en particular, se conoce que el tiempo de reparación (en minutos), T, sigue una distribución Normal con media 60 y desvío estándar 4.
Suponga que se van a tomar muestras aleatorias de $n = 30$ reparaciones de las máquinas de ese tipo y que el estadístico de interés es la variancia muestral, S^2.

a) Defina el estadístico pivote, función de la variancia muestral e indique en qué casos este se distribuye χ^2.

b) Si considera razonable utilizar la distribution χ^2, utilícela para obtener la probabilidad de que S^2 sea mayor que 16 min^2.

c) Represente gráficamente el valor obtenido (bosquejo) e interprételo en términos del problema.

12. Un estudio de las características de confiabilidad de sistemas auxiliares que trabajan con diesel reveló que el tiempo (en horas) antes de que sea necesario dar mantenimiento correctivo a dichos sistemas, T, se comporta según el modelo exponencial con una media de 1700 h.

a) Utilizando R, simule 10000 muestras de tamaño $n = 10$, $n = 50$ y $n = 100$ y para cada una de las muestras simuladas, obtenga el valor del tiempo promedio. Construya un gráfico con los 10000 valores obtenidos para cada tamaño de muestra.

b) ¿Qué ocurre con la distribución del tiempo medio antes de que sea necesario dar mantenimiento, según el tamaño de la muestra? Comente.

c) Obtenga la proporción de muestras de tamaño $n = 50$ en las cuales el tiempo promedio es

mayor que 2500 h.

d) ¿Puede considerar al valor obtenido en el ítem anterior como una aproximación de $P(\bar{Y} > 2500)$? Explique.

e) Vuelva a realizar los ítems *b)*, *c)* y *d)* suponiendo $T \sim U(1000; 2400)$.

7.7 Estadística con R

El siguiente código permite simular la distribución de probabilidades del estadístico media muestral, a partir de una población con distribución Exponencial(0,0028), como se ejemplifica en la Sección 7.4.4.

Con la función `rexp(n, 0.0028)` se simula una muestra de tamaño n, de una distribución Exponencial con $\alpha = 0{,}0028$ y, luego, se calcula una medida de resumen en cada muestra. En este caso, la medida de resumen es la media de la muestra.

```
mediaMuestral=function(n){
  muestra=rexp(n,0.0028)
  media=mean(muestra)
  return(media)
}
```

Con este código se decide simular 50000 muestras, todas de tamaño $n = 8$. De cada muestra, en este caso se obtiene la media muestral. Los 50000 valores de la media muestral se almacenan en "Medias".

```
set.seed(222)
Medias=replicate(50000,mediaMuestral(8))
```

En este caso, a partir de tener los 50000 valores de la media muestral (uno por cada muestra simulada de tamaño $n = 8$), se obtienen algunas medidas de resumen como el promedio y la desviación estándar, y la frecuencia relativa, correspóndiente a la proporción de medias mayores que 350 unidades. Los valores de estas medidas constituyen buenas aproximaciones de los parámetros de la distribución de probabilidades de $\bar{Y}$ por haberse obtenido a partir de la simulación de una gran cantidad de muestras.

El valor almacenado como “propor” se puede considerar una buena aproximación de la probabilidad de que la media de una muestra seleccionada al azar de la población sea mayor que 350 unidades.

```
mean(Medias)
sd(Medias)
propor <-length(Medias[Medias>350])/length(Medias)
```

Con el siguiente código, se puede elaborar el histograma para representar la distribución muestral de las medias obtenidas:

```
Medias <- data.frame(Medias)

ggplot(Medias, aes(Medias)) +
  geom_histogram(aes(y=..count..),
                 color="grey", fill="#f9b28c") +
  #Configuraciones generales
```

```
#Nombres de los ejes
labs(x = "Media muestral", y = "Frecuencia") +
#Configuraciones de formato
#Estilo
theme_classic()+
scale_x_continuous(expand=c(0,0),
labels = scales::label_number(accuracy = 1,
scale_y_continuous(expand=c(0,0))+
theme(axis.title.x = element_text(face="bold", colour="black",
size = 11, vjust=2),
      axis.title.y = element_text(face="bold", colour="black",
      size = 11, angle = 90, vjust = 1))
```

7.8 Apéndice

7.8.1 Estadístico: Promedio o media muestral

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- $(Y_1, Y_2, \ldots, Y_n)$ es una muestra aleatoria simple de tamaño n, [7]
- $E(Y_i) = \mu \ \forall i$,
- $V(Y_i) = \sigma^2 \ \forall i$.

Luego, aplicando las propiedades de la esperanza y la variancia de una combinación lineal de variables independientes e idénticamente distribuidas vistas en el Capítulo 6, se tiene que:

$$E(\bar{Y}) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n} n E(Y) = E(Y) = \mu,$$

$$V(\bar{Y}) = V(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2} n V(Y) = \frac{V(Y)}{n} = \frac{\sigma^2}{n}.$$

7.8.2 Estadístico: Proporción o frecuencia relativa muestral [8]

$$fr = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- $(Y_1, Y_2, \ldots, Y_n)$ es una muestra aleatoria simple de tamaño n, [7]
- $Y_i = 1$ si la unidad cumple cierta condición, e $Y_i = 0$ en caso contrario.
- $Y_i \sim Be(\pi) \forall i$ y, por lo tanto:
  - $E(Y_i) = \pi \ \forall i$,
  - $V(Y_i) = \pi(1-\pi) \ \forall i$.

En la Sección 7.4.2, fr se expresa como el cociente $\frac{C}{n}$, donde C indica la cantidad de unidades de la muestra que cumplen con la condición de interés. En este caso, C se comporta según el modelo Binomial, es decir, $C \sim Bi(n, \pi)$ y, por lo tanto:

- $E(C) = n\pi$,
- $V(C) = n\pi(1-\pi)$.

[7] Recuerde que en una muestra aleatoria simple las variables Y_i están igualmente distribuidas y son independientes entre sí.

[8] Se refiere a la proporción de unidades de la muestra que cumplen cierta condición.

Dado que $fr = \frac{C}{n}$, entonces:

- $E(fr) = \frac{E(C)}{n} = \frac{n\pi}{n} = \pi$,
- $V(fr) = \frac{V(C)}{n^2} = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$.

También se pueden obtener $E(fr)$ y $V(fr)$ aplicando las propiedades de la esperanza y la variancia de una combinación lineal de variables independientes e idénticamente distribuidas $Be(\pi)$, vistas en el Capítulo 6. Se tiene que:

$$E(fr) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n} nE(Y) = E(Y) = \pi$$

$$V(fr) = V\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2} nV(Y) = \frac{V(Y)}{n} = \frac{\pi(1-\pi)}{n}$$

7.8.3 Estadístico: Variancia muestral

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$(Y_1, Y_2, \ldots, Y_n)$ es una muestra aleatoria simple de tamaño n [7].

Observe que, en este caso, el estadístico no es una función lineal de las variables del vector. A continuación se obtiene su distribución de probabilidades en el caso que Y se comporte según la distribución Normal.

Si $Y \sim N(\mu, \sigma)$, $Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$ y $Z^2 = \frac{(Y-\mu)^2}{\sigma^2} \sim \chi^2_1$.

A su vez, $\sum_{i=1}^{n} \frac{(Y-\mu)^2}{\sigma^2} \sim \chi^2_n$, por propiedad reproductiva de la distribución χ^2.

Se puede demostrar que si se reemplaza la media poblacional μ por la media muestral $\bar{Y}$, entonces $\sum_{i=1}^{n} \frac{(Y-\bar{Y})^2}{\sigma^2} \sim \chi^2_{n-1}$, ya que este estadístico se puede expresar como la suma de $(n-1)$ variables estándar Z, elevadas al cuadrado.

Multiplicando el numerador y el denominador de la expresión anterior por $(n-1)$, se llega al estadístico $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ que se comporta según el modelo χ^2_{n-1}.

Por propiedades de la distribución Gamma [9], de la cual la distribución chi-cuadrado es un caso particular, si una variable $\chi^2 \sim \chi^2_{n-1}$, entonces:

[9] La distribución Gamma no se aborda en este libro, pero se puede encontrar información sobre ella en cualquier texto de Estadística

- $E(\chi^2) = n-1$,
- $V(\chi^2) = 2(n-1)$.

En este caso, la variable con distribución chi-cuadrado es $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$. De ahí que:

$E(\frac{(n-1)S^2}{\sigma^2}) = n-1$,

$\frac{n-1}{\sigma^2}E(S^2) = n-1 \Rightarrow E(S^2) = \sigma^2$,

$V(\frac{(n-1)S^2}{\sigma^2}) = 2(n-1)$,

$\frac{(n-1)^2}{\sigma^4}V(S^2) = 2(n-1) \Rightarrow V(S^2) = \frac{2\sigma^4}{n-1}$.

7.8.4 Distribuciones t de Student y chi-cuadrado

- t_{n-1} denota la distribución t de Student con $(n-1)$ grados de libertad.
  La apariencia general de esta distribución es similar a la de la distribución Normal estándar: ambas son simétricas y unimodales y el valor máximo de la ordenada se alcanza en la media $\mu = 0$. Sin embargo, esta distribución tiene colas más amplias que la Normal debido a que presenta mayor variabilidad.
  Una distribución t_{n-1} viene determinada por un único parámetro matemático llamado **grados de libertad** (v), que es función de n; por lo tanto existe una distribución t distinta para cada valor de n. A medida que aumentan los grados de libertad, la curva de densidad de la distribución t se aproxima a la curva de la distribución Normal estándar, como se observa en la Figura 7.10.

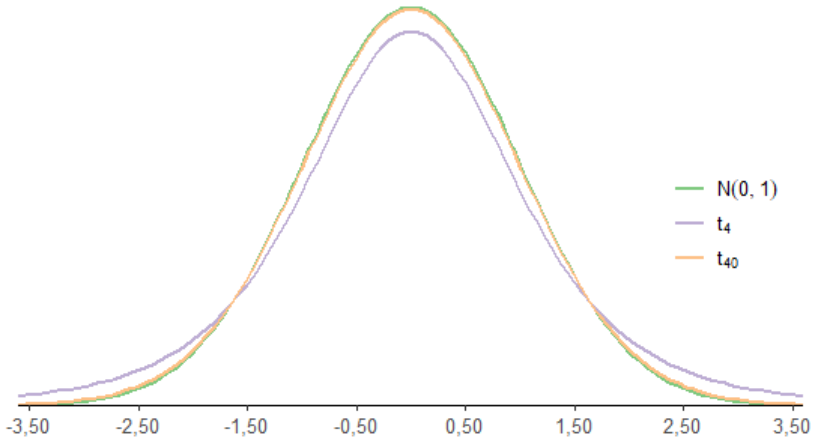


Figura 7.10 *Funciones de densidad de probabilidad de la distribución Normal estándar y de la t de Student con 4 y 40 grados de libertad*

- χ^2_{n-1} denota la distribución chi-cuadrado con $(n-1)$ grados de libertad.
  Esta distribución está definida para variables que solo toman valores positivos.
  Al igual que la distribución t_{n-1}, una distribución χ^2_{n-1} viene determinada por un único parámetro matemático llamado **grados de libertad** (v), que es función de n; por lo tanto existe una distribución χ^2_{n-1} distinta para cada valor de n.
  Si los grados de libertad son pequeños, la función de densidad de probabilidad es asimétrica a la derecha, pero a medida que aumentan los grados de libertad, las curvas de densidad se van haciendo más simétricas y se aproximan a la distribución Normal, como se observa en la Figura 7.11.

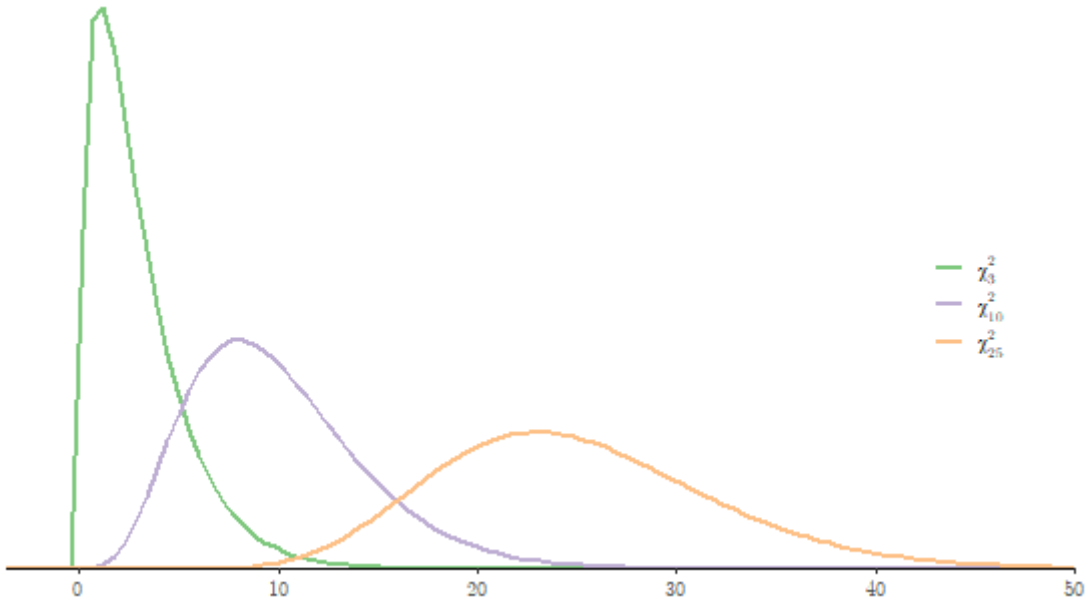


Figura 7.11 *Funciones de densidad de probabilidad de la distribución chi-cuadrado con 3, 10 y 25 grados de libertad*

7.8.5 Propiedades de los estimadores

Formalmente se definen dos propiedades para los estimadores, que se denominan **insesgamiento** y **consistencia**:

Un estimador $\widehat{\theta}$ es **insesgado** si su esperanza coincide con el valor del parámetro poblacional correspondiente θ.

$\widehat{\theta}$ es **insesgado** si $E(\widehat{\theta}) = \theta$.

Un estimador $\widehat{\theta}$ es **consistente** si se aproxima al valor del parámetro poblacional correspondiente θ, a medida que el tamaño de la muestra crece. Esto se cumple si:

- $\lim_{n\to\infty} E(\widehat{\theta}) = \theta$,
- $\lim_{n\to\infty} V(\widehat{\theta}) = 0$.

Por lo presentado en la Tabla 7.3, Sección 7.4.6, se puede afirmar que los tres estimadores considerados en este texto son **insesgados y consistentes**.

Existen otras propiedades de los estimadores; pero en este texto no se desarrollan.



8. Inferencia estadística. Estimación

8.1 Introducción

Cuando se realiza un análisis inferencial, se busca obtener conclusiones sobre una población a partir de la información que brinda una muestra. En muchas situaciones, interesan conclusiones sobre parámetros de la distribución de probabilidades de la variable de interés, pero también puede interesar concluir sobre la forma o modelo de dicha distribución, evaluar relaciones entre variables de interés o realizar predicciones.

En el proceso de inferencia sobre un parámetro, siempre interviene un estadístico muestral (estimador). Por esto, en el Capítulo 7 se trató el comportamiento de las distribuciones de probabilidades de algunos estadísticos cuando se toman muestras aleatorias simples.

Cuando se realiza concretamente la inferencia, se toma una única muestra a partir de la cual se obtienen las conclusiones, es decir, se obtiene un único valor del estadístico, de los infinitos posibles que corresponden a la distribución del mismo. Entonces, para poder sacar conclusiones sobre el parámetro poblacional, se debe considerar el margen de error que puede existir, es decir, la diferencia posible entre el valor del estadístico obtenido y el valor del parámetro, y es ahí donde el conocimiento de la distribución del estimador sirve de sustento. Conocer esta distribución, significa por ejemplo, saber con qué probabilidad el valor del estimador puede alejarse en más de una cierta magnitud del valor del parámetro y, a partir de esto, se puede acotar el margen de error de la estimación.

En la etapa de Análisis de los datos, las herramientas inferenciales completan un proceso que se inicia con la descripción de los datos obtenidos (Capítulo 2). El análisis descriptivo brinda conclusiones preliminares, permite detectar la presencia de observaciones atípicas y otras desviaciones, ayuda a probar si los requerimientos de la herramienta inferencial elegida se cumplen y también a considerar un modelo razonable para la población de la cual se extrajo la muestra. Esto es importante ya que todas las herramientas inferenciales se apoyan en supuestos sobre la población y pueden conducir a conclusiones erróneas si estos no se cumplen. En síntesis, las herramientas de análisis descriptivo y las inferenciales se complementan en la resolución de un problema.

En el presente capítulo, se describen las diferentes formas de realizar inferencias sobre los parámetros de la distribución de probabilidades y se analizan en detalle los **intervalos de confianza**, que constituyen una herramienta que sirve para estimar el valor de un parámetro a partir del valor del estimador en la muestra, indicando una cota de error y un riesgo asociado al hecho de trabajar con muestras aleatorias. Adicionalmente, se presenta una breve reseña sobre los **intervalos de tolerancia**, que dan información sobre posibles valores de la variable en la población.

Como se menciona en un párrafo anterior, la aplicación de las diferentes herramientas inferenciales supone el cumplimiento de ciertos requerimientos. En este capítulo se comenta cómo verificar dicho cumplimiento para poder obtener luego conclusiones válidas.

La técnica de muestreo y el tamaño de la muestra juegan un papel fundamental, no solo para las inferencias sobre los parámetros sino también para las inferencias sobre el modelo poblacional. Respecto del tipo de muestreo, en este libro solo se consideran muestras aleatorias simples (Capítulo 7).

Los objetivos de este capítulo son:

- Introducir las distintas formas de realizar inferencia estadística.
- Presentar los intervalos de confianza para los distintos parámetros y estudiar el efecto del tamaño de la muestra y del nivel de confianza en la precisión de los mismos.
- Completar el ciclo PPDAC utilizando los intervalos de confianza para obtener conclusiones en el contexto de los problemas planteados.
- Presentar los intervalos de tolerancia.

8.2 Inferencias sobre los parámetros de la distribución de probabilidades de la variable en estudio

Existen principalmente dos procedimientos utilizados para realizar inferencias estadísticas sobre parámetros: la estimación por intervalos de confianza y las pruebas de hipótesis.

- La estimación de parámetros se aplica cuando el objetivo es dar una idea aproximada del valor de estos:

  Se llama **estimación puntual** al valor que toma el estimador en la muestra obtenida. El cálculo de este valor es parte del análisis descriptivo, pero no se considera una forma apropiada de hacer inferencia ya que no brinda información sobre el error que se puede estar cometiendo ni sobre el riesgo de hacer inferencias erróneas.
  Esto sí se tiene en cuenta en la **estimación por intervalos de confianza**, donde se brinda una cota del error de estimación, la cual, para un riesgo y un tamaño de muestra dados, se obtiene a partir de la distribución de probabilidades del estimador.

- Las pruebas de hipótesis se aplican cuando se quiere tomar alguna decisión respecto de una hipótesis que se formula sobre el valor del parámetro de interés. Una de las formas de decidir sobre las hipótesis es a través de un intervalo de confianza.



Problema 1 (pág. 7): *Considere la población de barras producidas por la empresa metalúrgica, de las cuales interesa su longitud. La población estadística está formada por todas las longitudes.*
De acuerdo al objetivo del estudio se pueden definir diferentes parámetros de interés para dicha población. Suponga que en este caso interesa la proporción de barras que cumplen con las especificaciones, es decir, la proporción de barras con longitud entre 249,4 mm y 250,6 mm (π).
Entonces:

- *se construye un intervalo de confianza cuando se quiere saber aproximadamente qué valor tiene dicha proporción.*
- *se lleva a cabo una prueba de hipótesis si se quiere decidir si dicha proporción es igual, menor o mayor que cierto valor; por ejemplo, si es mayor que 0,10.*

En el Capítulo 7, se parte de una población donde se conoce la distribución de la variable de interés y se observa cómo varían de una muestra a otra los estadísticos media muestral, variancia muestral o frecuencia relativa muestral. A partir de la distribución muestral se pueden responder preguntas como, por ejemplo, "¿cuál es la probabilidad de que la media de la muestra se aleje en menos de una determinada magnitud de la media de la población?"

Con estos resultados, en este capítulo se presenta la forma de realizar inferencias respecto de los parámetros poblacionales: media, variancia y proporción, mediante la estimación por intervalos de confianza.

Dichas inferencias se realizan a partir de estimadores y por lo tanto, es importante que estos estimadores sean "buenos", es decir, que posean buenas propiedades (Sección 7.4.6).

Simbología

Cuando interesa estimar un parámetro, se define su correspondiente estimador. Una vez tomada la muestra, se calcula el valor de dicho estimador con los datos observados. Ese valor puede considerarse una **estimación puntual del parámetro**, sin ser suficiente para realizar inferencias sobre el mismo.

En general, la notación que se utiliza para los parámetros considerados en este libro se presenta en la Tabla 8.1.

Tabla 8.1 *Notación de parámetros, estimadores y estimaciones puntuales*

Parámetro	Estimador (antes de tomar la muestra)	Estimación puntual (con la muestra ya tomada)
Media poblacional, μ	$\bar{Y}$	$\bar{y}$
Variancia poblacional, σ^2	S^2	s^2
Proporción poblacional [1], π	f_r	f_o

[1] Se refiere a la proporción de unidades de la población que cumplen cierta condición.

Cabe señalar que, como se muestra en el Capitulo 7, para la variancia poblacional, σ^2, se dispone de un estimador insesgado, S^2, que además presenta otras propiedades deseables.
Sin embargo, en numerosas aplicaciones el interés no radica en la variancia en sí misma, sino en el desvío estándar poblacional, σ, ya que este se expresa en las mismas unidades de medida que la variable en estudio y resulta más sencillo de interpretar.
Por esta razón, la estimación de la variancia suele utilizarse principalmente como paso previo para realizar inferencias sobre el desvío estándar.

8.3 Intervalos de confianza

La **estimación por intervalo de confianza** se presenta de manera más detallada para el caso de la **media poblacional, μ, cuando se conoce el desvío estándar poblacional, σ.**

Para ese intervalo se desarrollan y analizan todos los conceptos involucrados (nivel de confianza, precisión del intervalo, diferencia entre el intervalo aleatorio y el intervalo de confianza, forma de interpretarlo, etc.). Estos conceptos son válidos también para el resto de los intervalos de confianza que se presentan en este capítulo.

8.3.1 Intervalo de confianza para la media poblacional

1. Estimación de la media poblacional, μ, cuando el desvío estándar poblacional, σ, es conocido

Desarrollo del intervalo de confianza

Como se menciona en la Sección 8.2, en la estimación por intervalo de confianza se brinda una cota del error de estimación, la cual, para un riesgo y un tamaño de muestra dados, se obtiene a partir de la distribución muestral del estimador.

Considere, a modo de ejemplo, los siguientes resultados obtenidos en el Capítulo 7:

Problema 1 (pág. 7): *Dada la variable longitud de las barras de acero (Y) y considerando muestras aleatorias simples de tamaño 10 se obtuvo la distribución de probabilidades de la media muestral:*

$\bar{Y} \sim N(\mu = 250, \frac{\sigma=0,2}{\sqrt{n=10}})$.

$\bar{Y} \sim N(\mu = 250, \sigma_{\bar{Y}} = 0,0632)$.

Conocida la distribución de la variable $\bar{Y}$, se planteó un intervalo centrado en la media poblacional [2], en el cual se encontraba, por ejemplo, el 95 % de las medias muestrales (pensando en los todos los posibles valores de la media muestral).

$$P(250-e<\bar{Y}<250+e) = P\left(\frac{(250-e)-250}{0,0632} < Z < \frac{(250+e)-250}{0,0632}\right) = P(z*_1 < Z < z*_2) = 0,95.$$

De allí se obtuvo que $z*_2 = z_{0,975} = \dfrac{(250+e)-250}{0,0632}$.

$z_{0,975} = 1,96$, *de donde* $e = 1,96 \cdot 0,0632 = 0,1239$.

Observación:
Debido a la simetría de la distribución Normal estándar respecto de 0, $z*_1 = -z*_2$ lo que conduce al mismo resultado para e.

El valor de e se puede interpretar de la siguiente manera: si se selecciona al azar una muestra de n = 10 barras, hay una chance de 0,95 de que la longitud promedio, para dicha muestra, se aleje del promedio poblacional como máximo en 0,1239 mm.
Se observa en la expresión anterior que ***el valor de*** *e* ***(semiampitud del intervalo obtenido) no depende del valor de la media poblacional*** μ *y es igual al desvío estándar de la variable media muestral ($\sigma_{\bar{Y}}$) multiplicado por una constante que es el percentil de orden 0,975 de la distribución Normal estándar.*

[2] Al estar centrado en la media poblacional, que en este caso vale 250, el intervalo puede expresarse de la forma $250 \pm e$.

Concretamente, el valor de e, que no depende del valor de μ, se puede considerar una "cota superior del error de estimación" (o error máximo de estimación), asociada a una cierta probabilidad. Entonces, cuando se desee estimar la media poblacional, se puede construir un intervalo sumando y restando esa cota a la media muestral observada; es decir, un intervalo centrado en la media muestral y de semiamplitud igual a e.

Generalizando este resultado se obtiene la expresión de un intervalo de confianza, como se muestra a continuación:

- Se cuenta con una variable aleatoria Y con $E(Y) = \mu$ y $V(Y) = \sigma^2$, donde σ es conocido.
- Para estimar a μ, se va a tomar una M.A.S. de n unidades.
- El estimador más adecuado de la media poblacional es la media muestral, $\bar{Y}$. Con respecto a la distribución de probabilidades de este estimador, a partir de lo visto en el Capítulo 7, el Teorema Central del Límite permite afirmar que para n suficientemente grande, $\bar{Y}$ tiene distribución aproximadamente Normal. Por otra partes, si $Y \sim N$, por la propiedad reproductiva $\bar{Y}$ también tendrá distribución Normal para cualquier valor de n. El siguiente desarrollo se realiza a partir de suponer que $\bar{Y}$ tiene esa distribución.
- El error de estimación se define como la diferencia entre el valor del estadístico y el valor del parámetro, $(\bar{Y} - \mu)$. Antes de tomar la muestra, el error de estimación es una variable y se pretende buscar un valor e que, con una probabilidad alta, no sea superado al tomar una muestra. A esta probabilidad se la simboliza $(1-\alpha)$. Por ejemplo, se puede tomar $(1-\alpha) = 0,95$.

Con todo esto, se plantea lo siguiente:

$$P(|\bar{Y} - \mu| < e) = 0,95 \Longrightarrow P(\mu - e < \bar{Y} < \mu + e) = 0,95.$$

Trabajando con la variable estandarizada $Z = \dfrac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$, queda:

$$P\left(\frac{-e}{\sigma/\sqrt{n}} < Z < \frac{e}{\sigma/\sqrt{n}}\right) = 0,95,$$

de lo cual se obtiene $\dfrac{e}{\sigma/\sqrt{n}} = z_{0,975}$, siendo $z_{0,975}$ el percentil de orden 0,975 de la distribución Normal estándar.

Luego se despeja $e = 1,96\,\dfrac{\sigma}{\sqrt{n}}$.

La cota superior e para el error de estimación de la media poblacional μ, cuando se conoce el desvío estándar poblacional σ, para una probabilidad de $0,95$ resulta:

$e = 1,96 \frac{\sigma}{\sqrt{n}}$.

Reemplazando e en la expresión original, queda:

$$P\left(\mu - 1,96 \frac{\sigma}{\sqrt{n}} < \bar{Y} < \mu + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95.$$

Despejando μ de ambas desigualdades, se obtiene:

$$P\left(\bar{Y} - 1,96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95.$$

El intervalo $(\bar{Y} - e\ ,\ \bar{Y} + e)$ es un **intervalo aleatorio** de probabilidad $0,95$ para la media poblacional μ, suponiendo conocida la desviación estándar poblacional σ. Representa la familia de intervalos que pueden obtenerse a partir de muestras aleatorias de tamaño n extraídas de la población, para la cual la probabilidad de que el intervalo que resulte de la muestra observada contenga al verdadero valor de μ es $0,95$.

En la expresión anterior, el promedio poblacional no se ha convertido en una variable, sigue siendo una constante en la población definida. Lo aleatorio es cada extremo del intervalo, que depende de la variable aleatoria *media muestral*. Es decir, se trata de la expresión de un intervalo donde sus extremos son aleatorios y la probabilidad hace referencia a la posibilidad de que esos dos valores aleatorios queden a ambos lados de μ, es decir, que el intervalo *cubra* al valor del parámetro que se desea estimar.

Cabe destacar que la amplitud del intervalo, en este caso, no es aleatoria, ya que no depende de lo observado en la muestra.

Esta probabilidad se define antes de tomar la muestra, asociada a la variable aleatoria $\bar{Y}$. Su valor (en este caso $0,95$), como toda probabilidad, puede interpretarse de las siguientes maneras:

- Antes de tomar una muestra de tamaño n, la chance de que el intervalo que se va a calcular a partir de la muestra observada cubra al valor del parámetro es 0,95.
- Si con cada una de las muestras posibles de tamaño n se construye un intervalo como el expresado, aproximadamente el $95\,\%$ de esos intervalos cubrirán al valor del parámetro.
- También puede afirmarse que la probabilidad de que un intervalo no cubra al valor del parámetro es 0,05 (el $5\,\%$ de todos los intervalos no cubrirán al valor del parámetro), lo cual puede interpretarse como un riesgo.

Todo lo expresado hasta ahora se refiere a lo que sucede antes de tomar la muestra. Luego de tomada la misma, se calcula el promedio, el cual es un valor numérico, y se obtiene el **intervalo de confianza**, el cual se define como:

$$IC_{95\,\%,\mu} = \left(\bar{y} - 1,96 \frac{\sigma}{\sqrt{n}}\ ,\ \bar{y} + 1,96 \frac{\sigma}{\sqrt{n}}\right).$$

En la expresión del intervalo de confianza, ambos extremos son valores numéricos, ya no hay nada aleatorio y por lo tanto **no se puede hablar de probabilidad**. Además, no puede saberse si el intervalo cubre o no al parámetro desconocido. Pero como antes de tomar la muestra había una probabilidad de 0,95 de que el intervalo que se iba a obtener cubra al valor del parámetro, se dice que dicho intervalo cubre al parámetro con una confianza del 95 %. Para comprender el concepto de la probabilidad $0,95$, se presenta la siguiente situación:



Problema 1 (pág. 7): *Considere las barras producidas en una empresa metalúrgica. Suponga que su longitud (en mm) sigue una ley Normal con media $\mu = 250$ y desviación estándar $\sigma = 0,2$.*
Se van a tomar muestras de tamaño 10 y, a partir de cada una, se va a construir un intervalo de confianza del 95 % para la media poblacional.

*El **intervalo aleatorio** para la media poblacional en este caso es:*

$$IA_{0,95;\mu} = \left(\bar{Y} - 1,96 \frac{0,2}{\sqrt{10}} \, , \, \bar{Y} + 1,96 \frac{0,2}{\sqrt{10}} \right) = \bar{Y} \pm 0,124.$$

Suponga que para una muestra de tamaño 10, el valor de la media muestral resulta $\bar{y} = 249,9$ mm.
*Luego el **intervalo de confianza** del 95 % para la media poblacional es:*

$$IC_{95\%,\mu} = 249,9 \pm 1,96 \frac{0,2}{\sqrt{10}} = 249,9 \pm 0,124 = (249,776; 250,024).$$

Este intervalo contiene a la media poblacional $\mu = 250$, por lo que esta muestra llevaría a decir que 250 es un valor posible de μ. Análogamente, se suponen otros valores para la media muestral como los que se presentan en la Tabla 8.2.

Tabla 8.2 *Intervalos de confianza para la longitud promedio poblacional de las barras (μ), para diferentes valores de la media muestral $\bar{Y}$*

Muestra N°	***Valor de la media muestral*** $\bar{y}$	***Intervalo de confianza del*** 95 % ***para la media poblacional*** $\bar{y} \pm 0,124$	***¿El intervalo cubre el valor de la longitud promedio poblacional?***
1	*250,08*	*(249,956; 250,204)*	*SI*
2	*249,85*	*(249,726; 249,974)*	*NO*
3	*249,94*	*(249,816; 250,064)*	*SI*
4	*250,03*	*(249,906; 250,154)*	*SI*
5	*250,16*	*(250,036; 250,284)*	*NO*

Todas estas situaciones se pueden visualizar en la Figura 8.1.

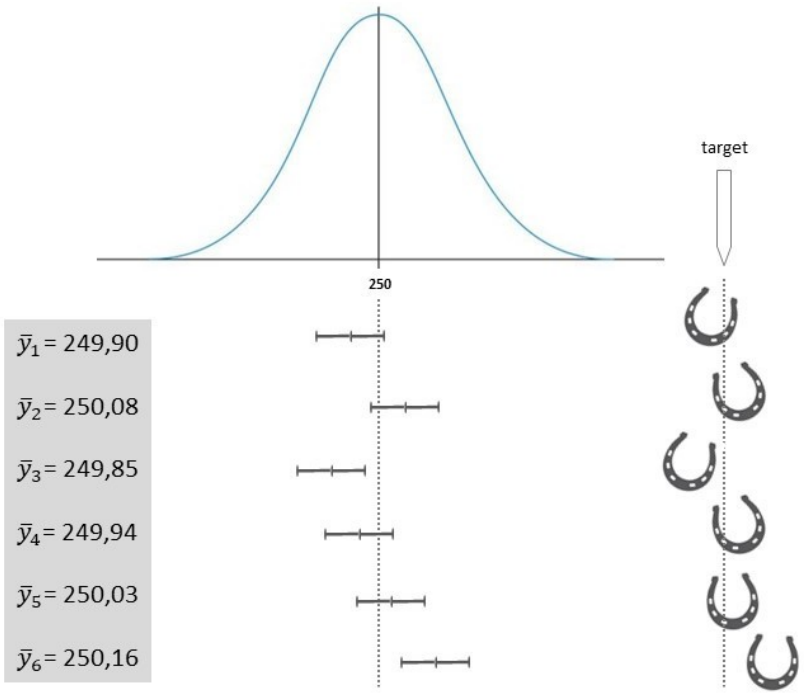


Figura 8.1 *Intervalos del 95 % de confianza para la longitud promedio de todas las barras de acero del proceso, construidos a partir de los valores de las medias de diferentes muestras de tamaño 10*

Como puede observarse, muestras diferentes dan valores diferentes de la media muestral, y, por lo tanto, intervalos diferentes. El proceso es análogo a arrojar herraduras para acertar en el blanco: algunos intervalos cubren al valor del parámetro y otros, no.
En el ejemplo, siempre que la media de la muestra esté situada a una distancia de a lo sumo 0,124 de la media poblacional, el intervalo cubrirá al valor de la media poblacional. Pensando en un gran número de muestras de tamaño $n = 10$ que se podrían tomar de la población, esto sucederá aproximadamente en el 95 % de todas las muestras posibles.

> *Para este caso se conoce el valor de la media poblacional μ y entonces se puede saber si un intervalo de confianza lo cubre o no.*
>
> *Cuando se realiza la inferencia en un problema real, se desconoce el valor de μ, de modo que nunca se tiene la certeza de que el intervalo de confianza obtenido lo cubra, solo se tiene una determinada confianza de que esto suceda.*

Hasta aquí se trabajó con una probabilidad de 0,95 para el intervalo aleatorio, lo cual significa un nivel de confianza del 95 % para el intervalo de confianza. El valor 1,96 es el percentil de orden 0,975 de la distribución Normal estándar, es decir, el valor que acumula una probabilidad igual a 0,975 en dicha distribución. Si se desea fijar otro valor para el nivel de confianza, cambiará el valor del percentil.

En general, definido un nivel de confianza $(1-\alpha)\,100\,\%$, la probabilidad, α, se divide en partes iguales a ambos lados de la distribución muestral, como se observa en la Figura 8.2. El percentil correspondiente es el percentil de orden $(1-\frac{\alpha}{2})$.

Por ejemplo, si se busca un nivel de confianza del 99 %, el percentil correspondiente es $P_{0,995} = z_{0,995} = 2,575$; mientras que, si interesa una confianza del 98 %, el percentil correspondiente es $P_{0,99} = z_{0,99} = 2,33$. En la Figura 8.2 se presentan los percentiles correspondientes a un nivel de confianza de $(1-\alpha)\,100\,\%$.

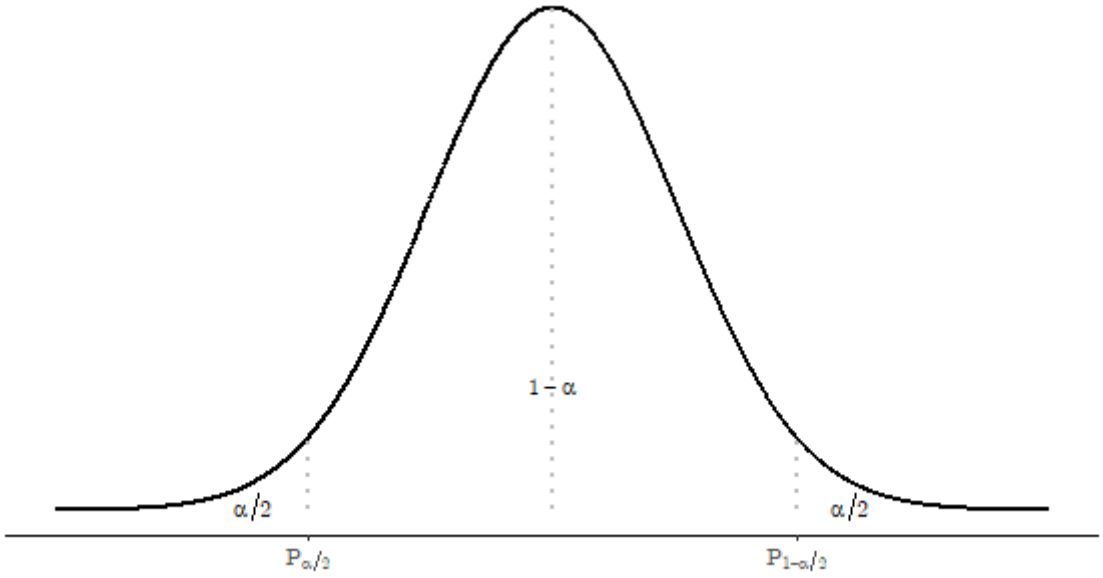


Figura 8.2 *Percentiles utilizados en la construcción de los intervalos con un nivel de confianza de* $(1-\alpha)\,100\,\%$

Las expresiones del **intervalo aleatorio** y del **intervalo de confianza** bilaterales para μ cuando σ es conocido, son las siguientes:

$$IA_{(1-\alpha),\mu} = \left(\bar{Y} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,,\,\bar{Y} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = \bar{Y} \pm z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},$$

$$IC_{(1-\alpha)\,100\,\%,\mu} = \left(\bar{y} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,,\,\bar{y} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = \bar{y} \pm z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

¿En función de qué elementos se determina el valor del nivel de confianza?

b. Intervalo aleatorio versus intervalo de confianza

- El **intervalo aleatorio** se plantea antes de tomar la muestra, cuando la media muestral se define como una variable aleatoria, y está centrado en $\bar{Y}$. El valor $(1-\alpha)$, fijado de antemano, es la probabilidad de que el intervalo que se vaya a obtener contenga al valor del parámetro (riesgo igual a α de que no lo contenga). Este intervalo es conceptual y representa a todos los intervalos que podrían obtenerse a partir de muestras del mismo tamaño.
- El **intervalo de confianza** se calcula luego de tomada la muestra y está centrado en el valor observado de la media muestral, $\bar{y}$. Es el intervalo numérico que se construye para un problema concreto y a partir del cual se obtienen conclusiones con un nivel de confianza del $(1-\alpha)\,100\,\%$.

Ambos intervalos tienen semiamplitud e, que se conoce como **cota superior del error de estimación**. En el intervalo aleatorio, e representa la distancia máxima que puede alejarse la media muestral de la media poblacional, con probabilidad $(1-\alpha)$. Luego, al sumar y restar e al valor de la media observada en la muestra, se obtiene el intervalo de confianza, que muestra los posibles valores del parámetro μ para los cuales sería razonable obtener el valor del promedio muestral observado.

c. Interpretación del intervalo de confianza

El intervalo de confianza obtenido da información sobre los posibles valores del parámetro desconocido (con un cierto nivel de confianza). No se puede afirmar con certeza cuánto vale ese parámetro (en este caso, μ) porque se cuenta con información de una muestra, pero se puede afirmar que se encuentra dentro del intervalo obtenido, con una cierta confianza.



Problema 1 (pág. 7): *Siguiendo con el ejemplo de las longitudes de las barras, suponga que se extrae una muestra y la media muestral resulta igual a* $\bar{y} = 250,2$ *mm.*
El intervalo de confianza del 95 % para la media poblacional es $IC_{95\%,\mu} = (250,076; 250,324)$.
En ese caso, se dice que se tiene una confianza de 95 % de que el intervalo $(250,076; 250,324)$ *cubra a la media poblacional desconocida.*
Si bien no se puede asegurar que este intervalo cubra al valor del parámetro, brinda información que puede servir para tomar decisiones.
Suponga, por ejemplo, que si la longitud promedio de las barras supera 250 mm, se requiere hacer ajustes en el proceso para que dicho promedio disminuya. Con un nivel de confianza del 95 %, los posibles valores para μ *son mayores que 250 mm. De ahí que pueda considerarse (corriendo un riesgo del 5 %) que es razonable pensar que* $\mu > 250$ *mm y, por lo tanto se recomendaría realizar ajustes en el proceso* [3].

[3] Recuerde que además de la información brindada por el intervalo de confianza, se debe considerar información del contexto como, por ejemplo, costos, cuestiones técnicas, etc. para tomar finalmente decisiones sobre un proceso.

Como se ve en el ejemplo, los intervalos de confianza pueden utilizarse con dos objetivos:

- Para conocer un rango de valores posibles para el parámetro de interés.
- Para responder preguntas sobre algún valor del parámetro, o bien para decidir si alguna hipótesis sobre el parámetro es razonable o no.

En este caso, una de las hipótesis es que la media poblacional es mayor a 250 mm ($\mu > 250$ mm). Al construir el intervalo, el mismo queda ubicado completamente a la derecha de dicho valor y eso permite concluir que el promedio poblacional es mayor que 250 mm (corriendo un riesgo del 5 %).

Es importante destacar que si antes de tomar la muestra la sospecha es que el promedio poblacional es menor (o mayor) a un determinado valor; al tomar la muestra, esta solo se podrá confirmar si el extremo derecho (izquierdo) del intervalo también es menor (mayor) que el valor planteado. Esto significa que solo toma relevancia uno de los extremos del intervalo:

- Si se sospecha que el parámetro es menor a cierto valor, solo toma relevancia el extremo superior del intervalo.
- Si se sospecha que el parámetro es mayor a cierto valor, solo toma relevancia el extremo inferior del intervalo.

Esto da lugar a la construcción de un intervalo de confianza unilateral.

d. Intervalos de confianza unilaterales

Cuando se plantea como objetivo conocer si un parámetro es mayor o menor a un cierto valor, se pueden construir **intervalos de confianza unilaterales** en los cuales el valor del riesgo, α, se localiza íntegramente a un solo lado en la distribución del estimador (solo a la izquierda o solo a la derecha).

En consecuencia, se obtiene la expresión de uno de los extremos del intervalo aleatorio (antes de tomar la muestra) y del intervalo de confianza (luego de tomar la muestra). El valor del otro extremo no entra en consideración y se le puede asignar el valor ∞ o $(-\infty)$ según corresponda.

Las expresiones de los intervalos aleatorios y de confianza unilaterales para μ cuando σ es conocido, son las siguientes:

- Intervalos unilaterales a la derecha:
  $$IA_{(1-\alpha),\mu} = \left(\bar{Y} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\ ,\ \infty\right),$$
  $$IC_{(1-\alpha)\,100\,\%,\mu} = \left(\bar{y} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\ ,\ \infty\right).$$
- Intervalos unilaterales a la izquierda:
  $$IA_{(1-\alpha),\mu} = \left(-\infty\ ,\ \bar{Y} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right),$$
  $$IC_{(1-\alpha)\,100\,\%,\mu} = \left(-\infty\ ,\ \bar{y} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right).$$

Cuando interesa saber si la media poblacional es mayor que cierto valor, toma relevancia el extremo inferior y el intervalo de confianza correspondiente se denomina **unilateral a la derecha**. Si interesa saber si la media poblacional es menor que cierto valor, toma relevancia el extremo superior y el intervalo correspondiente se denomina **unilateral a la izquierda**.

La denominación parece contradictoria, pero obedece a la hipótesis que se plantea sobre el parámetro de interés.

Problema 1 (pág. 7): *Siguiendo con el ejemplo de las longitudes de las barras, si antes de tomar la muestra se sospechaba que el promedio poblacional era mayor a 250 mm, se podría haber construido un intervalo de confianza unilateral a la derecha. El extremo relevante, en este caso, es el inferior.*
Suponga que se extrae una muestra y la media muestral resulta igual a $\bar{y} = 250,2$ *mm.*
Con un 95 % de confianza, se obtiene el siguiente intervalo:

$$IC_{95\,\%,\mu} = \left(\bar{y} - z_{0,95}\frac{0,2}{\sqrt{10}}\ ,\ \infty\right) =$$

$$= \left(250,2 - 1,645\,\frac{0,2}{\sqrt{10}}\ ;\ \infty\right) =$$

$$= (250,2 - 0,104\ ;\ \infty) = (250,096;\ \infty).$$

> *En este caso se puede informar, con una confianza del 95 % , que el intervalo* $(250,096;\infty)$ *cubre a la media poblacional desconocida. Como el extremo inferior, 250,096, es mayor que 250, se puede afirmar con un 95 % de confianza que* $\mu > 250$ *mm.*



Note que si no hay ninguna hipótesis que verificar respecto del parámetro de interés, el intervalo de confianza se utiliza para brindar una estimación del mismo. En ese caso solo tiene sentido construir un intervalo bilateral.

e. Relación entre el tamaño de la muestra, la cota del error de estimación y el nivel de confianza

De la expresión general para la cota superior del error de estimación, e, cuando se estima μ y σ es conocido, se observa que el valor de esa cota (e) (e indirectamente la precisión del intervalo de confianza), el nivel de confianza $(1-\alpha)\,100\,\%$ y el tamaño de la muestra (n) están vinculados entre sí. Un aumento o disminución en cualquiera de ellos afecta a los restantes.

- Si se mantiene fijo el nivel de confianza, a mayor tamaño de la muestra, e disminuye. El intervalo se hace más preciso, es decir, hay menos incertidumbre en la información que brinda.
- Si se mantiene fijo el tamaño de la muestra, a menor nivel de confianza, e también disminuye. El intervalo se hace más preciso; pero en este caso, a costa de tener mayor probabilidad (antes de tomar la muestra) de obtener un intervalo que no cubra al valor del parámetro.

Las relaciones mencionadas en los ítems anteriores se esquematizan en la Figura 8.3.

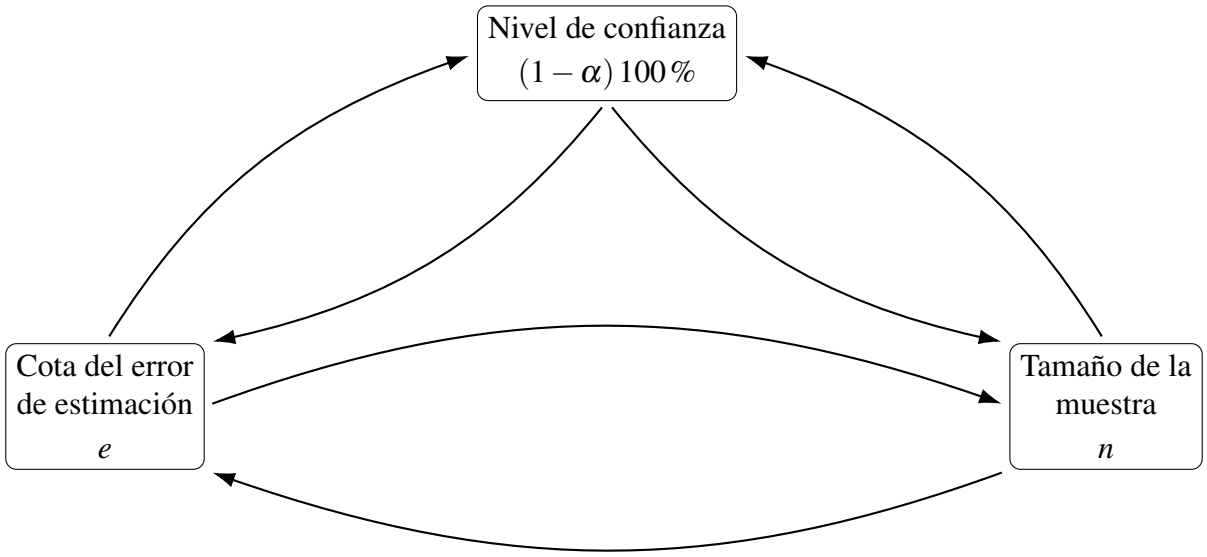


Figura 8.3 *Nivel de confianza, cota del error de estimación y tamaño de muestra, relacionados entre sí*

f. Uso del intervalo de confianza para la toma de decisiones sobre parámetros

En los ejemplos de este capítulo se vio cómo, a través de un intervalo de confianza, se puede concluir (en algunos casos), sobre alguna hipótesis relativa a un parámetro.

Por ejemplo, en el problema de las barras, se sospechaba que **el valor de la longitud promedio poblacional, μ, era mayor que 250 mm**, y como el intervalo obtenido resultó $IC_{95\%,\mu} = (250,096; \infty)$, se pudo concluir, con una confianza del 95 %, que la sospecha era acertada.

Por el contrario, si el intervalo obtenido hubiera resultado $IC_{95\%,\mu} = (248,2; \infty)$, no se hubiera podido concluir sobre la hipótesis planteada, ya que este abarca valores tanto mayores como menores a 250. Es decir que, según la información que brinde el intervalo de confianza, la hipótesis planteada podría descartarse o no.

- Como primer corolario, se puede decir que un intervalo sirve para concluir si un parámetro es mayor, menor o distinto a un valor determinado. Esa conclusión se obtiene si dicho intervalo no incluye a ese valor, incluyendo solamente valores mayores o menores.
- El hecho de que siempre haya incertidumbre no permite concluir nunca que el parámetro es exactamente igual a un cierto valor. Por eso, aunque el intervalo incluya al valor hipotético, en principio no se puede concluir nada.
  Sin embargo, así el intervalo incluya o no al valor planteado, se puede profundizar el análisis teniendo en cuenta los márgenes de error en el contexto del problema. Es aquí donde interviene la amplitud del intervalo y en consecuencia, el tamaño de la muestra.
  - Si el intervalo incluye al valor planteado, no es lo mismo que su amplitud sea muy pequeña o muy grande. Si bien no se puede afirmar que el parámetro es exactamente igual a ese valor planteado, si el intervalo es muy estrecho, en el contexto del problema quizás puede considerarse que es "prácticamente igual". En cambio, si el intervalo es muy amplio, claramente no se puede concluir (Recuerde que pensar que el intervalo es amplio o estrecho siempre depende del contexto de cada problema).
    Por ejemplo, en la situación de las longitudes de las barras, el contexto del problema indica que el promedio ideal es 250 mm ($\mu = 250$) y que, si el promedio resultara mayor que 250 mm ($\mu > 250$) se deben realizar ajustes en el proceso.
    Suponga que estos ajustes se justifican económicamente solo si $\mu > 250,15$ mm.
    Un intervalo de amplitud grande podría ser uno que incluyera a ambos valores, y eso sucede cuando la amplitud del intervalo es mayor a 0,15 unidades. Por ejemplo, si se hubiera obtenido el intervalo $IC_{95\%,\mu} = (249,94; 250,22)$, sería un error concluir que el promedio es 250 mm simplemente porque ese valor pertenece al mismo, ya que el intervalo llega a incluir valores mayores a 250,15.
    Ahora bien, si se hubiera obtenido el intervalo $IC_{95\%,\mu} = (249,98; 250,03)$, si bien no se puede asegurar que el promedio es exactamente 250 mm; en el contexto del problema se puede decir que "prácticamente" está muy cercano al valor ideal.
  - Si el intervalo no incluye al valor planteado, observando los valores que abarca, se puede tener una idea de cuáles valores podrían considerarse para el parámetro. En este caso se puede analizar, además de la amplitud, la "ubicación" del intervalo. Si todo el intervalo queda muy cerca del valor planteado (aunque no lo incluya), quizás en la práctica las diferencias pueden despreciarse. Por ejemplo el intervalo $IC_{95\%,\mu} = (250,002; 250,013)$ no incluye al valor 250 pero abarca valores que en el contexto del problema pueden considerarse muy cercanos. Por el contrario, si queda muy alejado, claramente el valor del parámetro es muy diferente al planteado, como por ejemplo en el caso que el intervalo resulte $IC_{95\%,\mu} = (250,18; 250,33)$, que indicaría que deben realizarse ajustes.

También podría pasar que el intervalo sea muy amplio y quede muy cerca a los valores de interés para el parámetro, por lo que abarcaría situaciones diferentes y se dificultaría concluir claramente acerca del valor planteado. Por ejemplo, si el intervalo resulta $IC_{95\%,\mu} = (250,03; 250,25)$ no se podría decidir si μ es 250 y el proceso funciona adecuadamente, o si μ es mayor que 250,15 y se justifica la realización de ajustes.

g. Determinación del tamaño de la muestra

Teniendo en cuenta lo explicado en el apartado anterior, con información de contexto del problema se puede pensar en una amplitud adecuada para el intervalo de confianza y, en base a eso, en cuál es el menor tamaño de muestra necesario para obtenerla.

Fijados el máximo valor admitido para la cota superior del error de estimación, e y el nivel de confianza $(1-\alpha)\,100\,\%$ deseado, es posible despejar el mínimo valor de n.

Si se va a construir un intervalo bilateral:

$$z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq e \Longrightarrow n \geq \frac{z^2_{1-\frac{\alpha}{2}}\,\sigma^2}{e^2},$$

donde $z_{1-\frac{\alpha}{2}}$ es el percentil de orden $(1-\frac{\alpha}{2})$ de la distribución Normal estándar y σ es el desvío estándar poblacional.
Si se va a construir un intervalo unilateral, se reemplaza el percentil $z_{1-\frac{\alpha}{2}}$ por el percentil $z_{1-\alpha}$ en la expresión anterior.



Problema 1 (pág. 7): *Siguiendo con el ejemplo de las longitudes de las barras, suponga que se desea determinar el mínimo tamaño de muestra requerido para estimar a la longitud promedio poblacional a partir de un intervalo de confianza bilateral.*
Se pretende que la amplitud del intervalo de confianza sea como máximo 0,15 mm, lo que equivale a que la cota superior del error de estimación sea como máximo 0,075 mm. Se desea también que el nivel de confianza sea del 95 %.
Suponga que se conoce que el desvío estándar poblacional, σ, vale 0,2 mm.

¿Cuál es el mínimo tamaño de muestra que se debe tomar?

$e \leq 0,075$
$(1-\alpha)\,100\,\% = 95\,\% \Rightarrow z_{1-\frac{0,05}{2}} = 1,96$
$\sigma = 0,2$

El mínimo tamaño de muestra requerido se obtiene de la siguiente manera:

$$n \geq \left(\frac{1,96 \cdot 0,2}{0,075}\right)^2 \approx 27,3 \quad \Rightarrow \quad n \geq 28.$$

Es decir, si se pretende que el máximo valor para la cota del error de estimación sea 0,075 mm y que la confianza sea del 95 %, se debe tomar una muestra de tamaño $n \geq 28$ barras.

Si se deseara realizar una estimación del parámetro de interés a través de un intervalo de confianza unilateral, en el cálculo del mínimo tamaño de muestra requerido se modifica el valor del percentil correspondiente.
$e \leq 0,075$
$(1-\alpha)\,100\,\% = 95\,\% \Rightarrow z_{1-0,05} = 1,645$
$\sigma = 0,2$

El mínimo tamaño de muestra requerido en este caso se obtiene de la siguiente manera:

$$n \geq \left(\frac{1,645 \cdot 0,2}{0,075}\right)^2 \approx 19,24 \quad \Rightarrow \quad n \geq 20.$$

Es decir, si se pretende que el máximo valor para la cota del error de estimación sea 0,075 mm y se va a constuir un intervalo unilateral con nivel de confianza del 95 %, se debe tomar una muestra de tamaño $n \geq 20$ barras.

¿En qué etapa del Ciclo PPDAC se define el nivel de confianza deseado?, ¿y el tamaño de la muestra?

Con lo visto hasta ahora, se pueden resolver las Actividades 1 a 3, Sección 8.6.

2. Estimación de la media poblacional, μ, cuando el desvío estándar poblacional, σ, es desconocido

En el caso que se quiera estimar μ y se desconozca el valor de σ, en lugar del estadístico pivote $Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ se utiliza el estadístico pivote $T = \frac{\bar{Y}-\mu}{S/\sqrt{n}}$, el cual tiene distribución t de Student con $(n-1)$ grados de libertad [4] si $Y \sim N(\mu, \sigma)$ (Ver Capítulo 7, Sección 7.8.4). Esta distribución tiene colas más pesadas que la distribución Normal estándar, lo cual hace que el intervalo de confianza resulte más amplio. Esto ocurre debido a que la incertidumbre es mayor ya que no se conoce el desvío estándar poblacional. La distribución t de Student se utiliza precisamente en estos casos.

Siguiendo un procedimiento similar al empleado en el caso de conocer σ, se obtiene la expresión de la cota superior del error de estimación:

$$e = t_{(n-1),1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}.$$

Si se compara esta expresión con la obtenida en el caso de conocer el valor del desvío estándar poblacional, se observa que en lugar del percentil de la distribución Normal estándar se utiliza el percentil de la distribución t_{n-1} y en lugar del valor del desvío estándar poblacional, σ, se utiliza el valor del desvío estándar muestral, S.

[4] El valor correspondiente a los grados de libertad constituye el parámetro matemático de la distribución t de Student.

Fijado al nivel de confianza, $(1-\alpha)$, el percentil $P_{1-\frac{\alpha}{2}} = t_{n-1,1-\frac{\alpha}{2}}$ es mayor que el percentil z del mismo orden, especialmente para tamaños de muestra pequeños. Es allí donde se observa la pérdida de precisión en la estimación de μ por no conocer el valor de σ.

Por otro lado, al aparecer S en lugar de σ, el valor de e en este caso depende de la muestra que se obtenga y resulta variable.

Las expresiones del **intervalo aleatorio** y del **intervalo de confianza** bilaterales para μ cuando σ es desconocido son las siguientes:

$$IA_{(1-\alpha),\mu} = \left(\bar{Y} - t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \; , \; \bar{Y} + t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right),$$

$$IC_{(1-\alpha)\,100\,\%,\mu} = \left(\bar{y} - t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \; , \; \bar{y} + t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right).$$

El valor $t_{n-1,1-\frac{\alpha}{2}}$ que aparece en la expresión de los intervalos anteriores, depende del nivel de confianza, $(1-\alpha)\,100\,\%$ y del tamaño de la muestra, n.

Por ejemplo, para un nivel de confianza del 95 %, el percentil correspondiente es $P_{0,975} = t_{(n-1);0,975}$.

Si el tamaño de muestra es $n = 10$, $t_{9;0,975} = 2,262$; si $n = 20$, $t_{19;0,975} = 2,093$ y si $n = 100$, $t_{99;0,975} = 1,984$.
Observe que, a medida que n aumenta, el valor del percentil de la distribución t se aproxima al valor del percentil del mismo orden de la distribución Normal estándar, que, en este caso es $z_{0,975} = 1,96$.

Problema 1 (pág. 7): *Reconsidere la situación de la empresa metalúrgica interesada en estudiar el comportamiento de la longitud de las barras. Suponga que ahora interesa estimar a la longitud promedio de la totalidad de las barras, μ, con un nivel de confianza del 99 %; pero que no se conoce el valor del desvío estándar poblacional, σ.*
A partir de una muestra aleatoria de $n = 120$ barras se obtuvo la longitud promedio de la muestra, $\bar{y}$ y el desvío estándar muestral, s.
$\bar{y} = 249,994$
$s = 0,301$
Si la longitud de las barras (Y) se puede considerar normalmente distribuida, $\bar{Y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.
Al no conocerse el desvío estándar poblacional, se define el estadístico $T = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$, *que se comporta según la distribución* t_{n-1}.
Para un nivel de confianza del 99 % y un tamaño de muestra de 120 barras, el percentil correspondiente es $t_{(119);0,995} = 2,618$.

$$IC_{99\,\%,\mu} = \left(\bar{y} - t_{0,995;119} \frac{s}{\sqrt{120}} \; , \; \bar{y} + t_{0,995;119} \frac{s}{\sqrt{120}}\right) =$$

$$= \left(249,994 - 2,618 \frac{0,301}{\sqrt{120}} \; ; \; 249,994 + 2,618 \frac{0,301}{\sqrt{120}}\right) =$$

$$= (249,994 - 0,072\ ;\ 249,994 + 0,072) =$$

$$= (249,922; 250,066).$$

Este intervalo cubre al valor de la longitud promedio poblacional con una confianza del 99 %.

Las expresiones de los intervalos aleatorios y de confianza unilaterales para μ cuando σ es desconocido, son las siguientes:

- Intervalos unilaterales a la derecha:
  $IA_{(1-\alpha),\mu} = \left(\bar{Y} - t_{n-1,1-\alpha}\frac{S}{\sqrt{n}}\ ,\ \infty\right)$,
  $IC_{(1-\alpha)100\%,\mu} = \left(\bar{y} - t_{n-1,1-\alpha}\frac{s}{\sqrt{n}}\ ,\ \infty\right)$.
- Intervalos unilaterales a la izquierda:
  $IA_{(1-\alpha),\mu} = \left(-\infty\ ,\ \bar{Y} + t_{n-1,1-\alpha}\frac{S}{\sqrt{n}}\right)$,
  $IC_{(1-\alpha)100\%,\mu} = \left(-\infty\ ,\ \bar{y} + t_{n-1,1-\alpha}\frac{s}{\sqrt{n}}\right)$.

¿Cómo determinar el mínimo tamaño de muestra requerido para estimar μ cuando σ es desconocido?

Así como se hizo en el caso de la estimación de μ con σ conocido, se podría despejar el mínimo tamaño de muestra requerido a partir de fijar un máximo valor para la cota superior del error de estimación, e y el nivel de confianza $(1-\alpha)\,100\,\%$ deseado.

En este caso, $t_{(n-1),1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \leq e \Longrightarrow n \geq \frac{t^2_{(n-1),1-\frac{\alpha}{2}}\,S^2}{e^2}$.

Observe que esa expresión no se puede utilizar ya que:

- El valor de s, desviación estándar muestral, se va a poder obtener una vez que la muestra ya esté tomada.
- El valor del percentil de la distribución t de Student depende de n.

Para la determinación del mínimo tamaño de muestra requerido, entonces, se propone lo siguiente:

- utilizar un valor del desvío estándar obtenido en un estudio anterior o en un estudio piloto.
- reemplazar el percentil de la distribución t_{n-1} por el percentil del mismo orden de la distribución Normal estándar, ya que este no depende de n.[5]

[5] Recuerde que el percentil correspondiente varía según interese construir un intervalo de confianza unilateral o bilateral.

Fijados el máximo valor admitido para la cota superior del error de estimación, e y el nivel de confianza $(1-\alpha)\,100\,\%$ deseado, es posible despejar el mínimo valor de n.

Si se va a construir un intervalo bilateral:

$$z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n}} \leq e \Longrightarrow n \geq \frac{z^2_{1-\frac{\alpha}{2}}\,\hat{\sigma}^2}{e^2},$$

donde $z_{1-\frac{\alpha}{2}}$ es el percentil de orden $(1-\frac{\alpha}{2})$ de la distribución Normal estándar y $\hat{\sigma}$ es una estimación del desvío estándar poblacional obtenida a partir de estudios previos o de una muestra piloto.
Si se va a construir un intervalo unilateral, se reemplaza el percentil $z_{1-\frac{\alpha}{2}}$ por el percentil $z_{1-\alpha}$ en la expresión anterior.

Una vez definido el valor de n y tomada la muestra, en la construcción del intervalo de confianza se utiliza el valor del desvío estándar muestral observado, s, y el valor del percentil t correspondiente.

Problema 1 (pág. 7): *Reconsidere la situación de la empresa metalúrgica interesada en estudiar el comportamiento de la longitud de las barras.*
Suponga ahora que interesa estimar a la longitud promedio de la totalidad de las barras, μ, mediante un intervalo bilateral del 99 % de confianza, pero no se conoce el valor del desvío estándar poblacional, σ. Suponga además que se pretende un máximo de 0,03 para la cota superior del error de estimación y que, de estudios anteriores, se puede considerar que el desvío estándar es alrededor de 0,2.

¿Cuál debe ser el mínimo tamaño de muestra en este caso?

$e \leq 0,03$
$(1-\alpha)\,100\,\% = 99\,\% \Rightarrow z_{1-\frac{0,01}{2}} = 2,57$ *(reemplaza al percentil t que no se puede obtener sin conocer n)*
$\hat{\sigma} = 0,2$ *(valor aproximado de estudios anteriores)*

En este caso, el cálculo quedaría de la siguiente manera:

$$n \geq \left(\frac{2,57 \cdot 0,2}{0,03}\right)^2 \approx 293,55 \quad \Rightarrow \quad n \geq 294.$$

Es decir que se requiere una muestra compuesta como mínimo por 294 barras para estimar a la longitud promedio a partir de un intervalo de confianza bilateral, con una cota de error de a lo sumo 0,03 mm y un nivel de confianza del 99 %.

Para utilizar el estadístico T se requiere que la distribución de la población de la cual se obtuvo la muestra sea Normal.

Ahora bien, ¿puede utilizarse en otros casos? Según resultados de estudios de simulación, si la distribución de la población es aproximadamente simétrica, con tamaños de muestra no muy pequeños (mayores a 10), puede usarse la distribución t-student.

En cambio, para distribuciones fuertemente asimétricas, como la Exponencial, usar la distribución t-student daría resultados erróneos para cualquier tamaño de muestra: los intervalos tendrían un nivel de confianza real menor al planteado originalmente.



Con lo visto hasta ahora, se pueden resolver las Actividades 4 a 7, Sección 8.6.

8.3.2 Intervalo de confianza para la variancia y el desvío estándar poblacionales

En el Capítulo 7 se estudió la distribución de probabilidades de la variancia muestral, S^2.

Si la población de la cual se extraen las muestras es Normal, el estadístico pivote $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2}$ tiene una distribución chi-cuadrado (χ^2) con $(n-1)$ grados de libertad [6].

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Con el mismo razonamiento hecho para obtener el intervalo de confianza para la media poblacional, a partir de plantear una probabilidad para el estadístico con distribución χ^2 con $(n-1)$ grados de libertad se puede despejar el parámetro σ^2 y encontrar una expresión para el intervalo aleatorio correspondiente:

$$P\left({\chi^2}_a < \frac{(n-1)S^2}{\sigma^2} < {\chi^2}_b\right) = 1-\alpha.$$

Trabajando algebraicamente se obtiene que:

$$P\left(\frac{(n-1)S^2}{{\chi^2}_b} < \sigma^2 < \frac{(n-1)S^2}{{\chi^2}_a}\right) = 1-\alpha,$$

donde ${\chi^2}_a$ es el percentil de orden $(\frac{\alpha}{2})$ y ${\chi^2}_b$ es el percentil de orden $(1-\frac{\alpha}{2})$ de una distribución χ^2_{n-1}.

De la expresión anterior se desprende la del intervalo aleatorio y luego, reemplazando la variable S^2 por el valor observado en la muestra, s^2, se obtiene la del intervalo de confianza.

Las expresiones del **intervalo aleatorio** y del **intervalo de confianza** bilaterales para σ^2 son las siguientes:

$$IA_{(1-\alpha),\sigma^2} = \left(\frac{(n-1)S^2}{{\chi^2}_{1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{{\chi^2}_{\frac{\alpha}{2}}}\right)$$

$$IC_{(1-\alpha)\,100\,\%,\sigma^2} = \left(\frac{(n-1)s^2}{{\chi^2}_{1-\frac{\alpha}{2}}}, \frac{(n-1)s^2}{{\chi^2}_{\frac{\alpha}{2}}}\right).$$

Dado que la función raíz cuadrada es estrictamente creciente en los números reales positivos, es posible obtener un intervalo de confianza para el desvío estándar poblacional a partir del intervalo de confianza para la variancia poblacional, conservando el mismo nivel de confianza. Para ello, se aplica la raíz cuadrada a ambos extremos del intervalo de confianza para la variancia.

[6] El valor correspondiente a los grados de libertad constituye el parámetro matemático de la distribución χ^2.

La expresión del **intervalo de confianza** bilateral para σ es la siguiente:

$$IC_{(1-\alpha)\,100\,\%,\sigma} = \left(\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}} \; , \; \sqrt{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}} \right).$$



Problema 1 (pág. 7): *Reconsidere la situación de la empresa metalúrgica interesada en estudiar el comportamiento de la longitud de las barras.*
Suponga que ahora interesa estimar a la desviación estándar poblacional de las longitudes, σ, con un nivel de confianza del 99 %.
A partir de una muestra aleatoria de $n = 120$ barras se obtuvo $s^2 = 0,0906\ mm^2$.
Si la longitud de las barras (Y) se puede considerar normalmente distribuida, $\frac{119\,S^2}{\sigma^2} \sim \chi^2_{119}$.
Dado que el nivel de confianza es 99 %, los percentiles de la distribución mencionada son los siguientes: $\chi^2_{119;0,005} = 83,018$ y $\chi^2_{119;0,995} = 162,481$.

En primer lugar se obtiene el intervalo de confianza para la variancia poblacional, que resulta:

$$IC_{99\,\%,\sigma^2} = \left(\frac{119 \cdot 0,0906}{162,481} \; ; \; \frac{119 \cdot 0,0906}{83,018} \right)$$
$$IC_{99\,\%,\sigma^2} = (0,0663; 0,1297).$$

Luego, se aplica la función raíz cuadrada a ambos extremos del intervalo de confianza para la variancia poblacional y se obtiene el correspondiente intervalo de confianza para el desvío estándar poblacional:
$IC_{99\,\%,\sigma} = (\sqrt{0,0663}; \sqrt{0,1297}) = (0,2575; 0,3601)$.
Este intervalo cubre al valor de la desviación estándar poblacional de la longitud de las barras con una confianza del 99 %.

Las expresiones de los intervalos aleatorios y de confianza unilaterales para σ^2 y de los intervalos de confianza unilaterales para σ son las siguientes:

- Intervalos unilaterales a la derecha:

$$IA_{(1-\alpha),\sigma^2} = \left(\frac{(n-1)S^2}{\chi^2_{1-\alpha}} \; , \; \infty \right),$$
$$IC_{(1-\alpha)\,100\,\%,\sigma^2} = \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha}} \; , \; \infty \right),$$
$$IC_{(1-\alpha)\,100\,\%,\sigma} = \left(\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha}}} \; , \; \infty \right).$$

- Intervalos unilaterales a la izquierda:

$$IA_{(1-\alpha),\sigma^2} = \left(0 \; , \; \frac{(n-1)S^2}{\chi^2_{\alpha}} \right),$$
$$IC_{(1-\alpha)\,100\,\%,\sigma^2} = \left(0 \; , \; \frac{(n-1)s^2}{\chi^2_{\alpha}} \right),$$
$$IC_{(1-\alpha)\,100\,\%,\sigma} = \left(0 \; , \; \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha}}} \right).$$

El intervalo de confianza para σ^2 basado en la distribución chi cuadrado es muy sensible a la falta de normalidad y su desempeño no mejora al aumentar el tamaño muestral. Si la normalidad de la población no es razonable, el nivel de confianza real del intervalo puede resultar mucho menor que el pretendido.
En esos casos, el intervalo de confianza puede construirse mediante el método de Bonett [13], el cual es robusto frente a la falta de normalidad.
En este texto no se desarrolla analíticamente dicho intervalo, pero puede obtenerse con R.



Con lo visto hasta ahora, se pueden resolver las Actividades 8 a 10, Sección 8.6.

8.3.3 Intervalo de confianza para la proporción de unidades de la población que cumplen cierta condición

Esta estimación se basa en la distribución de probabilidades de la frecuencia relativa muestral, vista en el Capítulo 7.

En la población se cuenta con una variable aleatoria Y que toma el valor 1 si la unidad elemental cumple una cierta condición y 0 si no la cumple. Es decir, $Y \sim Be(\pi)$, donde π, proporción de unidades que cumplen la condición de interés en la población, es el parámetro que se desea estimar, a partir de una muestra de n unidades.

Como se muestra en el Capítulo 7, se definen los estadísticos muestrales C: cantidad de unidades de la muestra que cumplen la condición de interés, que tiene distribución binomial ($C \sim Bi(n,\pi)$) y $fr = \frac{C}{n}$: proporción de éxitos en la muestra.

fr es el estimador más adecuado para π.

Tanto la distribución de C como la de fr dependen de n y π, el parámetro a estimar.

Para estimar a la media poblacional, μ y a la variancia poblacional, σ^2, se define un estadístico pivote (Z o T y X^2) y a partir de la distribución del mismo, se plantea una probabilidad y luego se despejan los posibles valores para el parámetro, lo que permite obtener una expresión del intervalo aleatorio correspondiente. Luego, reemplazando al estadístico (variable aleatoria) por el valor observado en la muestra (valor numérico) se obtiene el intervalo de confianza.

En el caso de la proporción sucede que, al plantear la probabilidad $P(a < C < b) = 1 - \alpha$, o bien $P(\frac{a}{n} < fr < \frac{b}{n}) = 1 - \alpha$, como la distribución exacta de C es Binomial, se hace imposible encontrar una expresión analítica para luego despejar los posibles valores de π y así encontrar el intervalo aleatorio.

Para la construcción de intervalo de confianza exacto para la proporción poblacional, se desarrolla entonces un procedimiento que permite obtener los extremos del mismo sin necesidad de determinar explícitamente la expresión del intervalo aleatorio. Estos valores se obtienen a partir del valor observado en la muestra (como en los casos anteriores) pero no reemplazando en la expresión de un intervalo aleatorio, sino determinando computacionalmente los posibles valores del parámetro π para los cuales el valor observado de C es "razonable" con el nivel de confianza considerado (misma lógica que en los casos anteriores).

El procedimiento para obtener el intervalo de confianza bilateral exacto para la proporción poblacional π es el siguiente:

- A partir del valor observado del estadístico C, es decir de c, se obtiene el valor de π para el que se verifica que $P(C < c) = \frac{\alpha}{2}$, donde $C \sim Bi(n,\pi)$. Es decir, se hace variar el valor del parámetro π de la distribución Binomial hasta que se verifique la igualdad anterior. **El valor encontrado será menor que $\frac{c}{n}$ y corresponde al extremo inferior del intervalo**.
- Luego, de manera similar, se obtiene otro valor de π para el que se verifica que $P(C > c) = \frac{\alpha}{2}$. **Este valor será mayor que $\frac{c}{n}$ y corresponde al extremo superior del intervalo)**.

En el caso del intervalo unilateral a la derecha, a partir de c, se obtiene el valor de π para el que se verifica que $P(C < c) = \alpha$, donde $C \sim Bi(n,\pi)$. **El valor encontrado será menor que $\frac{c}{n}$ y corresponde al extremo inferior del intervalo. El extremo superior vale 1.**

En el caso del intervalo unilateral a la izquierda, a partir de c, se obtiene el valor de π para el que se verifica que $P(C > c) = \alpha$, donde $C \sim Bi(n,\pi)$. **El valor encontrado será mayor que $\frac{c}{n}$ y corresponde al extremo superior del intervalo. El extremo inferior vale 0.**

Este procedimiento, que se realiza con ayuda de software, garantiza que el nivel de confianza del intervalo obtenido sea igual o mayor al que se pretende.

Algunas particularidades del intervalo de confianza para la proporción son:

- El valor del parámetro está limitado entre 0 y 1 y el valor del estimador también.
- En el caso de la construcción de intervalos unilaterales, el extremo inferior correspondiente a un intervalo a la izquierda es 0; mientras que el extremo superior correspondiente a un intervalo a la derecha es 1.
- Salvo que el valor observado de fo sea 0,5, el intervalo será asimétrico (no estará centrado en el valor observado), ya que la distribución Binomial no es simétrica para valores de π diferentes de 0,5. Cuanto más cercano a 0 o a 1 esté fo, más asimétrico será.
- Si el valor observado del estimador en la muestra es 0 o 1, toma sentido solo un valor del intervalo. Es decir, cabe construir un intervalo de confianza unilateral, quedando toda la probabilidad α en una de las colas de la distribución.
- El nivel de confianza real no se puede conocer y es mayor o igual al planteado inicialmente en el cálculo.

El intervalo obtenido como se menciona en los párrafos anteriores se denomina **intervalo exacto de Clopper–Pearson para la proporción de unidades de la población que cumplen con cierta condición, π**. Dada su construcción no se puede definir fácilmente una expresión para el intervalo aleatorio y para el intervalo de confianza.

Problema 1 (pág. 7): *Reconsidere la situación de la empresa metalúrgica productora de barras metálicas. Suponga que ahora interesa estimar a la proporción de barras que cumplen las especificaciones en relación a la longitud, es decir, la proporción de barras con longitudes entre 249,4 mm y 250,6 mm, π, con un nivel de confianza del 95 %.*

A partir de una muestra aleatoria de $n = 120$ barras se encontró que 114 cumplían con las especificaciones.

El estadístico en este caso es C: cantidad de barras de la muestra que tienen longitudes entre 249,4 mm y 250,6 mm. $C \sim Bi(120, \pi)$.

También puede definirse el estadístico f_r: proporción de barras de la muestra que tienen longitudes entre 249,4 mm y 250,6 mm.

En esta muestra, el valor observado de C resultó $c = 114$ y el valor observado de f_r resultó $f_o = \frac{114}{120} = 0,95$.

Con esa información, y definido el nivel de confianza, se obtuvo el siguiente intervalo exacto de Clopper-Pearson para el parámetro de interés:

$IC_{95\%,\pi} = (0,894; 0,981)$.

Este intervalo cubre al valor de la proporción poblacional de barras que cumplen con las especificaciones, con una confianza del 95 %.

Observe que, como se manifestó en párrafos anteriores, este intervalo no está centrado en el valor de la proporción muestral de barras que cumplen las especificaciones ($f_o = 0,95$), ya que dicho valor no es cercano a 0,50.

Dado que se trata de una proporción de barras que satisfacen ciertas especificaciones, podría construirse un intervalo de confianza unilateral en el que interese el valor mínimo que podría tomar este parámetro (intervalo unilateral a la derecha). El mismo es el siguiente:

$IC_{95\%,\pi} = (0,904; 1)$.

Observe que en este caso, el extremo que interesa es el inferior, y al extremo superior se le asigna el valor 1, que es el máximo valor que puede tomar una proporción.

Este intervalo informa, con un 95 % de confianza, que la proporción de barras que cumplen con las especificaciones en relación a su longitud es como mínimo 0,904.

> ! *La construcción del intervalo de confianza constituye la última tarea en la etapa de Análisis de los datos y el punto de partida para la etapa de Obtención de conclusiones, en la cual se deben elaborar conclusiones en contexto. Como se indica en la Tabla 1.3 (página 29), se plantean las siguientes preguntas: ¿Las barras de acero de cierto tipo cumplen con los requerimientos de la automotriz en relación con su longitud? ¿Cuál es la proporción de barras que satisfacen las especificaciones en relación con la longitud?*
> *A partir de la información brindada por el intervalo de confianza para el parámetro π y del contexto del problema, se responde a estas preguntas evaluando el cumplimiento de las especificaciones técnicas y estimando la proporción de barras conformes. En función de ello, se decide si se puede abastecer a la automotriz o si es necesario realizar previamente modificaciones en el proceso de producción.*

En algunas situaciones se puede aproximar la distribución de fr con la distribución Normal y entonces puede realizarse un proceso similar al que se hizo para encontrar el intervalo para μ, a partir de la distribución del estadístico pivote (Z).

Si $n \to \infty$, $fr \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}})$ y la expresión para el intervalo aleatorio es la siguiente:

$$IA_{(1-\alpha),\pi} = \left(f_r - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}} \; ; \; f_r + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}} \right).$$

Observe que la expresión de la cota del error de estimación para este intervalo, depende del valor del parámetro π que se desea estimar. Por ese motivo, se reemplaza el valor del parámetro por el valor de su estimador (frecuencia relativa muestral, f_r) y se utilizan las siguientes expresiones (basadas en aproximaciones) para el intervalo aleatorio y para el intervalo de confianza:

Las expresiones del **intervalo aleatorio** y del **intervalo de confianza** bilaterales para π basadas en la aproximación Normal, son las siguientes:

$$IA_{(1-\alpha),\pi} = \left(f_r - z_{1-\frac{\alpha}{2}} \sqrt{\frac{f_r(1-f_r)}{n}} \; , \; f_r + z_{1-\frac{\alpha}{2}} \sqrt{\frac{f_r(1-f_r)}{n}} \right)$$

$$IC_{(1-\alpha)\,100\,\%,\pi} = \left(f_o - z_{1-\frac{\alpha}{2}} \sqrt{\frac{f_o(1-f_o)}{n}} \; , \; f_o + z_{1-\frac{\alpha}{2}} \sqrt{\frac{f_o(1-f_o)}{n}} \right).$$



Problema 1 (pág. 7): *Con la misma información se puede construir el intervalo de confianza para la proporción de barras que cumplen las especificaciones en relación a la longitud, es decir, la proporción de barras con longitudes entre 249,4 mm y 250,6 mm, π, con un nivel de confianza del 95 %; pero ahora utilizando la aproximación Normal.*

A partir de una muestra aleatoria de $n = 120$ barras se encontró que 114 cumplían con las especificaciones.

En esta muestra, el valor observado de C resultó $c = 114$ y el valor observado de f_r resultó $f_o = \frac{114}{120} = 0,95$.

En este caso, el tamaño de muestra es grande y no habría problemas en construir este intervalo.

El valor del percentil correspondiente es:

$z_{1-\frac{0,05}{2}} = z_{0,975} = 1,96$.

El intervalo de confianza aproximado es el siguiente:

$$IC_{95\,\%,\pi} = \left(0,95 - 1,96 \sqrt{\frac{0,95 \cdot 0,05}{120}} \; ; \; 0,95 + 1,96 \sqrt{\frac{0,95 \cdot 0,05}{120}} \right) =$$

$$= (0,95 - 0,039 \; ; \; 0,95 + 0,039) = (0,911; 0,989)$$

Este intervalo, a diferencia del anterior, está centrado en el valor observado de la frecuencia relativa muestral, f_o.

Las expresiones del **intervalo aleatorio** y del **intervalo de confianza** unilaterales para π basadas en la aproximación Normal, son las siguientes:

- Intervalos unilaterales a la derecha:

$$IA_{(1-\alpha),\pi} = \left(f_r - z_{1-\alpha} \sqrt{\frac{f_r\,(1-f_r)}{n}}\ ,\ 1 \right),$$

$$IC_{(1-\alpha)\,100\,\%,\pi} = \left(f_o - z_{1-\alpha} \sqrt{\frac{f_o\,(1-f_o)}{n}}\ ,\ 1 \right).$$

- Intervalos unilaterales a la izquierda:

$$IA_{(1-\alpha),\pi} = \left(0\ ,\ f_r + z_{1-\alpha} \sqrt{\frac{f_r\,(1-f_r)}{n}} \right),$$

$$IC_{(1-\alpha)\,100\,\%,\pi} = \left(0\ ,\ f_o + z_{1-\alpha} \sqrt{\frac{f_o\,(1-f_o)}{n}} \right).$$

Problema 1 (pág. 7): *El intervalo de confianza unilateral a la derecha para la proporción de barras que cumplen las especificaciones en relación a la longitud es el siguiente:*

$$IC_{95\,\%,\pi} = (0{,}95 - 1{,}65 \sqrt{\frac{0{,}95\,0{,}05}{120}}\ ;\ 1) =$$

$$= (0{,}95 - 0{,}033\,;\,1) = (0{,}917;\,1)$$

El intervalo de confianza basado en la aproximación Normal puede llevar a resultados erróneos, dado que el nivel de confianza real del intervalo puede ser mucho menor al planteado, especialmente si el tamaño de muestra es pequeño.

¿Cómo determinar el mínimo tamaño de muestra requerido en este caso?

Si se va a construir un intervalo de confianza para π, basado en la aproximación Normal, se podría despejar el valor mínimo de n a partir de fijar un máximo valor para la cota superior del error de estimación, e y el nivel de confianza $(1-\alpha)\,100\,\%$ deseado.

En este caso, $z_{1-\frac{\alpha}{2}} \sqrt{\dfrac{f_o\,(1-f_o)}{n}} \leq e \Rightarrow n \geq \dfrac{z^2_{1-\frac{\alpha}{2}}\, f_o\,(1-f_o)}{e^2}$.

Observe que el valor de f_o, frecuencia relativa o proporción muestral de unidades que cumplen la condición de interés, se va a poder obtener una vez que la muestra ya esté tomada. Entonces, para la determinación del mínimo tamaño de muestra requerido, se puede utilizar un valor aproximado de la proporción a estimar, obtenido en un estudio anterior o en un estudio piloto, $\hat{\pi}$.

Si no se cuenta con esta información, otra opción, más conservadora, es asignarle a f_o el valor 0,50, que maximiza el numerador de la expresión anterior y, en consecuencia, produce el mayor tamaño mínimo de muestra requerido.

Fijados el máximo valor admitido para la cota superior del error de estimación, e y el nivel de confianza $(1-\alpha)\,100\,\%$ deseado, es posible despejar el mínimo valor de n.

Si se va a construir un intervalo bilateral:

$$z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}\,(1-\hat{\pi})}{n}} \leq e \Longrightarrow n \geq \frac{z^2_{1-\frac{\alpha}{2}}\,\hat{\pi}\,(1-\hat{\pi})}{e^2},$$

donde $z_{1-\frac{\alpha}{2}}$ es el percentil de orden $(1-\frac{\alpha}{2})$ de la distribución Normal estándar y $\hat{\pi}$ es una estimación de la proporción poblacional de interés obtenida a partir de estudios previos o de una muestra piloto. Recuerde que $\hat{\pi}$ puede reemplazarse por el valor 0,50 si se carece de información sobre la proporción. Si se va a construir un intervalo unilateral, se reemplaza el percentil $z_{1-\frac{\alpha}{2}}$ por el percentil $z_{1-\alpha}$ en la expresión anterior.

Una vez definido el valor de n y tomada la muestra, en la construcción del intervalo de confianza se utiliza el valor de la frecuencia relativa muestral observada, f_o.

En el Capítulo 4, página 191, se presentó la definición frecuencial de la probabilidad de un suceso (A), a partir de repetir n veces un experimento y de analizar la frecuencia relativa de dicho suceso. En ese contexto, la proporción de veces que se presenta el suceso A en las n repeticiones se interpretó como una aproximación a su probabilidad.
Desde la perspectiva adoptada en este capítulo, esa misma situación puede reinterpretarse en términos inferenciales. Las n repeticiones del experimento constituyen una muestra aleatoria, y la proporción de unidades de la muestra que verifican la condición de interés constituye una estimación puntual de la proporción poblacional π o, equivalentemente, de la probabilidad de ocurrencia del suceso, $P(A)$. El intervalo de confianza permite incorporar explícitamente la variabilidad propia del procedimiento de muestreo, proporcionando un rango de valores posibles para la proporción poblacional o para $P(A)$, asociado a un determinado nivel de confianza. Si la cantidad de repeticiones del experimento es grande (tamaño de muestra n grande), el intervalo obtenido brindará una estimación precisa del parámetro de interés.



Problema 7 (pág. 8): *En la situación planteada en este problema, interesa realizar un cierto número de mediciones de una resistencia conocida, con un multímetro, para evaluar al procedimiento de medición de la resistencia que se realiza con dicho equipo.*
En el Capítulo 4, página 191, se supuso que se realizaba un número relativamente grande ($n = 100$) de mediciones de dicha resistencia, en Ω y se definió como condición de interés que la medición realizada se encuentre entre dos valores, a y b (suceso A).
Si se corrobora que en 82 de las 100 mediciones realizadas, la resistencia se encuentra entre a y b, $f_o = \frac{82}{100} = 0,82$ constituye una estimación puntual de $P(A)$ (probabilidad de que una medición de la resistencia conocida se encuentre entre a y b) y consecuentemente de π (proporción de mediciones de la resistencia conocida que toman valor entre a y b).
Con los contenidos que se desarrollan en el presente capítulo, se puede estimar al parámetro de interés con

un cierto nivel de confianza (por ejemplo, del 95 %). En este caso, se concluye, con un 95 % de confianza que el intervalo $IC_{95\%,\pi} = (0,73; 0,89)$ *cubre a la proporción poblacional de mediciones de la resistencia conocida con valor entre a y b o, equivalentemente, a la probabilidad de que una medición de la resistencia conocida esté entre a y b.*



Con lo visto hasta ahora, se pueden resolver las Actividades 11 a 13, Sección 8.6.

8.4 Intervalo de tolerancia

En las secciones anteriores se presentaron intervalos de confianza que se construyen en ocasión de estimar algún parámetro de interés. Existen otras situaciones en las cuales interesa contar con información sobre posibles valores de una variable, como la que se presenta, a modo de ejemplo, a continuación.



Problema 1 (pág. 7): *Reconsidere la situación de la empresa metalúrgica productora de barras. Las mismas deben tener longitud en el intervalo* $(250 \pm 0,6)$ *mm.*
Suponga que se pretende que como mínimo un 99 % de las barras tengan longitud en el intervalo mencionado. ¿Se cumplirá con lo pretendido?

En esta sección se presenta el **intervalo de tolerancia**, el cual no brinda información sobre parámetros, sino que proporciona un intervalo de valores tal que, con un determinado nivel de confianza, al menos una proporción P de las unidades de la población toma valores de la variable dentro del mismo.



Problema 1 (pág. 7): *En el caso de la producción de barras, interesa el intervalo de valores de la variable que garantice que, como mínimo, el 99 % de las barras* $(P = 0,99)$ *tenga longitud dentro del mismo. Esta estimación también tendrá asociado un nivel de confianza* $(1-\alpha)\,100\,\%$*, por ejemplo, 95 %.*

En los intervalos de tolerancia se especifican dos cantidades: por un lado, el nivel de confianza de la estimación, $(1-\alpha)\,100\,\%$, y por el otro, la proporción mínima de la población que se desea abarcar, P.

La idea consiste en construir un intervalo de la forma $\bar{Y} \pm k\,S$, donde k es una constante que depende del nivel de confianza, de la proporción P y del tamaño muestral. El intervalo de tolerancia también puede ser unilateral a la derecha o a la izquierda.

Antes de tomar la muestra, el valor $(1-\alpha)$ representa la probabilidad de que el intervalo que se vaya a obtener contenga los percentiles poblacionales correspondientes a la proporción P especificada. Para un intervalo de tolerancia bilateral, se trata de los percentiles de orden $\left(\frac{1-P}{2}\right)$ y $\left(P + \frac{1-P}{2}\right)$. Para un intervalo unilateral a la derecha (en el que solo está definido el extremo inferior), se trata del percentil de orden $1-P$ y para un intervalo unilateral a la izquierda (en el que solo está definido el extremo superior), se trata del percentil de orden P.

La construcción de un intervalo de tolerancia requiere suponer una forma para la distribución de la variable de interés. En particular, cuando se asume que dicha distribución es Normal, el valor de k

se obtiene a partir de la distribución t no central, cuyo desarrollo no se aborda en este texto. Por este motivo, los intervalos de tolerancia se calcularán directamente mediante el software R (Ver Sección 8.8.3, página 442).

Una vez construido el intervalo de tolerancia, se realizan las interpretaciones correspondientes y se extraen las conclusiones pertinentes.



Problema 1 (pág. 7): *En el caso de las barras, $P = 0,99$, de donde los percentiles de interés son $p_{0,005}$ y $p_{0,995}$ y el nivel de confianza $(1-\alpha)\,100\,\% = 95\,\%$.*
El intervalo de tolerancia obtenido con R, suponiendo normalidad, es:

$IT_{P=0,99;95\,\%} = (249,122;\ 250,866)$.

Esto puede interpretarse de la siguiente manera: con un 95 % de confianza se puede afirmar que como mínimo el 99 % de las longitudes de las barras se encuentran entre 249,122 mm y 250,866 mm.
En la empresa se pretendía que como mínimo el 99 % de las longitudes se encuentren entre 249,4 mm y 250,6 mm.
El límite inferior del intervalo de tolerancia es menor a 249,4 mm y el límite superior es mayor a 250,6 mm, por lo tanto no se puede afirmar que se cumple lo que pretende la empresa en relación a las longitudes.
Si el intervalo de tolerancia hubiera sido (249,5; 250,3 mm), ¿qué se concluiría?
En este caso, con un 95 % de confianza se puede afirmar que como mínimo el 99 % de las longitudes se encuentran entre 249,5 y 250,3 mm, de modo que también puede afirmarse que como mínimo el 99 % de las mismas se encuentra entre 249,4 y 250,6 mm (que es lo que se pretende en la empresa).



Un aspecto fundamental para la construcción de intervalos de tolerancia es el conocimiento de la forma de la distribución de probabilidades de la variable de interés. Existen diversos enfoques para su obtención bajo distintos supuestos distribucionales, así como también alternativas no paramétricas que no requieren especificar un modelo probabilístico particular.

Cuando el interés se centra en estimar la proporción de unidades de la población que verifican una determinada condición sobre una variable cuantitativa, $\pi = P(a < Y < b)$, es posible recurrir a procedimientos basados en un modelo probabilístico para la variable Y, en lugar de estimar π directamente a partir de la clasificación de las observaciones según cumplan o no la condición, como se vio en la Sección 8.3.3. Estos métodos permiten construir intervalos de confianza para π aprovechando la información contenida en la distribución de la variable y en la estimación de sus parámetros poblacionales, como la media y la desviación estándar. En este enfoque, la información muestral no se reduce a una variable indicadora de cumplimiento, sino que se utiliza la magnitud completa de los valores observados, lo que puede conducir a procedimientos más eficientes bajo el modelo asumido. Un tratamiento de este tipo de métodos puede encontrarse en [14]. Sin embargo, la construcción de estos intervalos de confianza específicos para π no será considerada en este texto. Asimismo, los intervalos de tolerancia proporcionan información indirecta sobre $\pi = P(a < Y < b)$, ya que permiten realizar afirmaciones probabilísticas acerca de la fracción de la población contenida en determinados intervalos.



Con lo visto hasta ahora, se pueden resolver las Actividades 14 y 15, Sección 8.6.

8.5 Evaluación de los requerimientos de las herramientas presentadas

En un estudio estadístico, una vez finalizada la recolección de los datos, se inicia la etapa de análisis. Como primer paso, se realiza un tratamiento primario de la información con el objetivo de evaluar y garantizar su calidad, detectando posibles problemas tales como datos faltantes, inconsistencias, errores de registro o valores atípicos que requieran revisión. Posteriormente, se aplican herramientas descriptivas, que permiten organizar, resumir y visualizar la información recolectada (como se presentó en el Capítulo 2).

Cuando los datos corresponden a una muestra, el análisis descriptivo resulta insuficiente para obtener conclusiones definitivas acerca de la población en estudio, por lo que debe complementarse con herramientas inferenciales. Sin embargo, la validez de las conclusiones obtenidas mediante estos procedimientos depende del cumplimiento de ciertos requerimientos.

En este contexto, el análisis descriptivo adquiere también un rol diagnóstico, ya que permite evaluar el cumplimiento de dichos requerimientos, al menos de forma aproximada, antes de aplicar la técnica inferencial correspondiente. Verificar su cumplimiento es fundamental, pues la aplicación de herramientas inferenciales sin esta evaluación previa puede conducir a conclusiones erróneas o poco confiables. De este modo, el análisis descriptivo no solo contribuye a comprender los datos, sino que también orienta la selección adecuada del método inferencial y respalda la validez de los resultados obtenidos.

Entre los aspectos que pueden analizarse antes de realizar el análisis inferencial, pueden mencionarse dos de particular importancia: la forma en que se seleccionaron las unidades que componen la muestra y la distribución que tiene la variable de interés.

Si se observara que los datos no corresponden a una muestra aleatoria simple (M.A.S.), no sería válido aplicar las herramientas inferenciales presentadas en este capítulo. En cuanto a la distribución de la variable de interés, algunas técnicas pueden requerir que dicha variable siga una distribución determinada, por lo que su aplicación dependerá del cumplimiento de este requerimiento.

Con respecto a la muestra, es imposible verificar, solamente con la información de los datos, que los mismos corresponden a una M.A.S. dado que esto tiene que ver, principalmente, con la forma (aleatoria o no) en la que se eligieron las unidades de la muestra (Sección 7.2). Sin embargo, existe una situación en la que la falta de aleatoriedad sí puede detectarse a partir de los datos: cuando el proceso de selección y/o medición de la variable de interés introduce una tendencia o patrón a lo largo del tiempo (desde la primera unidad hasta la última). Este comportamiento puede observarse mediante un gráfico de series temporales. Para ello, es necesario contar con información sobre el orden en que fueron seleccionadas las unidades y/o medidos los valores de la variable.
En la Figura 2.13, Capítulo 2 (página 67), se presentan diferentes diagramas de series cronológicas que dan cuenta, en su mayoría, de patrones no aleatorios.



Problema 1 (pág. 7): *Suponga que la información de las longitudes de las barras se encuentra según el orden en que se realizaron las mediciones en las barras seleccionadas.*
En la Figura 8.4 se presenta un diagrama en el cual los valores de las longitudes se grafican en el orden

en que se realizaron las mediciones. En este caso, el diagrama pone de manifiesto la ausencia de patrones, tendencias o cualquier comportamiento no aleatorio que haga descartar la posibilidad de usar esos datos para hacer inferencia dado que no corresponden a una muestra aleatoria.
Esto, de todas maneras no permite asegurar que la muestra observada corresponda a una muestra aleatoria simple ya que para ello habría que conocer con detalle cómo fueron seleccionadas las barras y medidas las longitudes.

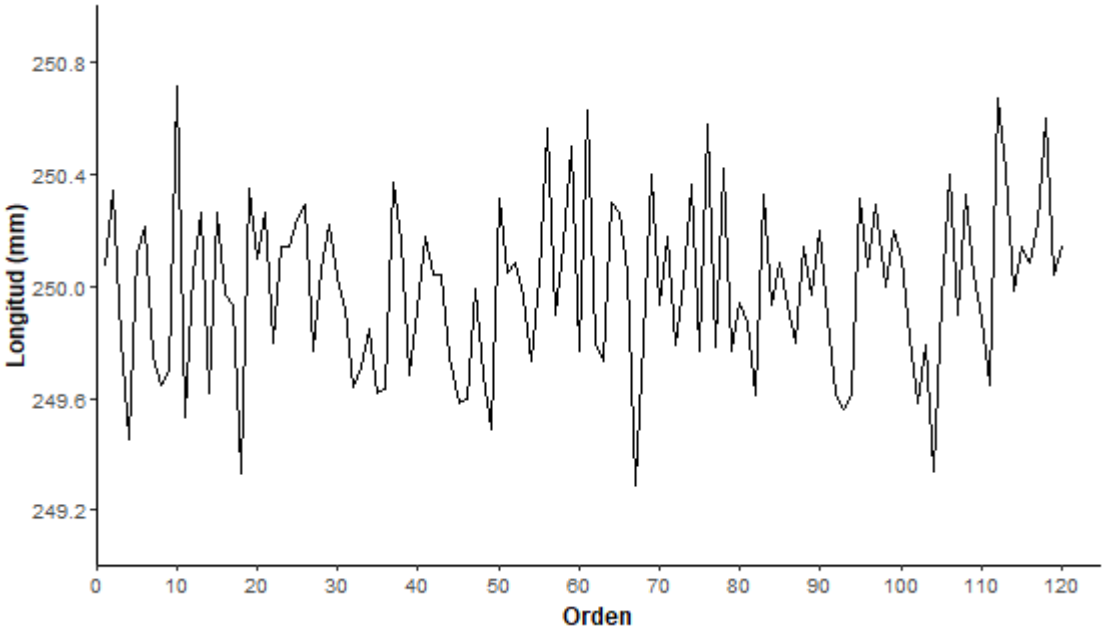


Figura 8.4 *Longitudes de las barras, en el orden en que fueron observadas*

Con respecto a la distribución de la variable, en el Capítulo 2 se vio que existen distintos gráficos que muestran la distribución de frecuencias. En los mismos se puede analizar la simetría de la distribución, la existencia de valores extremos, etc., y en algunos casos lo observado puede ayudar a decidir cuáles herramientas inferenciales utilizar.

Existe otra herramienta descriptiva, no presentada en el Capítulo 2, que permite analizar la distribución de los datos y comparar la distribución observada en una muestra con un modelo teórico, como los estudiados en el Capítulo 5. Esta herramienta se denomina **diagrama de probabilidad**.

Se trata de un gráfico en el que se representan los cuantiles [7] muestrales frente a los cuantiles teóricos de una distribución específica (Normal, Exponencial, entre otras). El diagrama se construye de tal manera que, si la distribución de los datos se ajusta adecuadamente al modelo considerado, los puntos tenderán a alinearse aproximadamente sobre una recta. En consecuencia, cuanto más próximos se encuentren los puntos a dicha recta, mejor será el ajuste entre la distribución observada y el modelo teórico propuesto. Por el contrario, desviaciones sistemáticas respecto de la recta sugieren discrepancias entre los datos y la distribución postulada.

Por ejemplo, al realizar un diagrama de probabilidad Normal (comparación con la distribución Normal), si los datos están distribuidos como una campana simétrica (similar a una campana de Gauss), quedarán puntos próximos a una recta. Si los datos están distribuidos de una forma marcadamente asimétrica, se verá un patrón de puntos curvo y con muchos puntos alejados de la recta.

[7] En este texto se utilizan indistintamente los términos *cuantil* y *percentil*.

La observación del diagrama, entonces, orienta sobre el modelo asociado a la población de la cual se extrajo la muestra.

Problema 1 (pág. 7): *Como complemento del análisis descriptivo realizado en el Capítulo 2 para estudiar el comportamiento de las longitudes de las barras, se construyó el diagrama de probabilidad Normal que se presenta en la Figura 8.5. Como se observa en dicha figura los puntos están bastante próximos a la recta, lo cual sugiere que la muestra podría haber sido tomada de una población normal (las longitudes de las barras se podrían comportar según el modelo normal).*

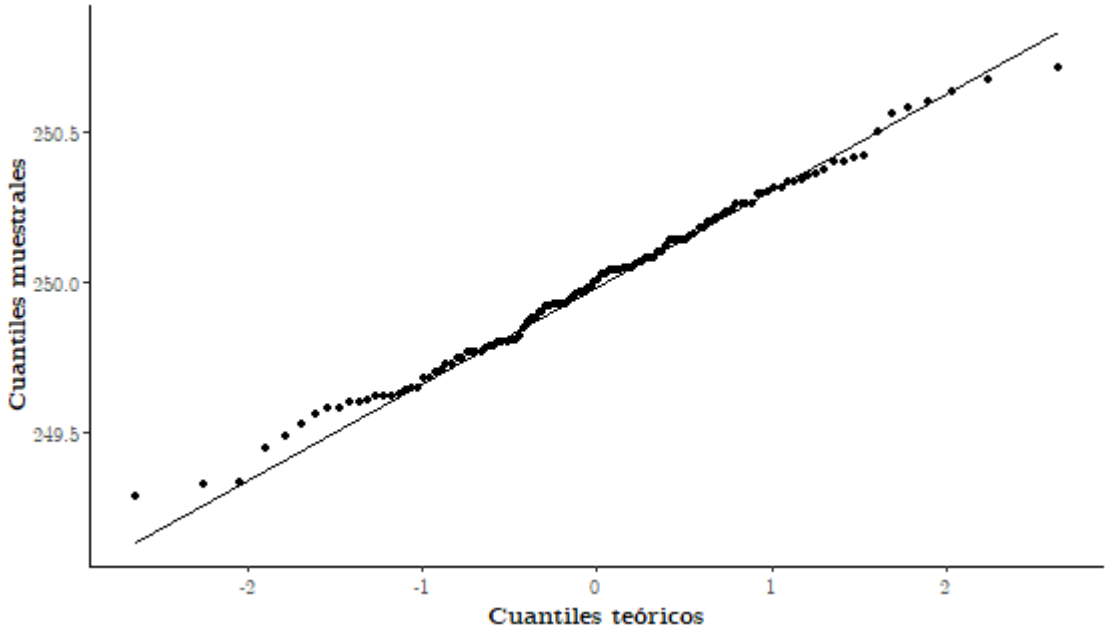


Figura 8.5 *Diagrama de probabilidad Normal construido con las longitudes de las n = 120 barras de acero seleccionadas*

Problema 8 (pág. 8): *Para analizar el comportamiento de los tiempos hasta la falla de los n = 45 dispositivos electrónicos analizados, se construyó el diagrama de probabilidad Normal que se presenta en la Figura 8.6.*
Como se observa en dicha figura, los puntos presentan un patrón curvo y se alejan de la recta, lo cual sugiere que la muestra no fue tomada de una población normal (o también que los tiempos hasta la falla de los dispositivos no se comportan según el modelo normal).

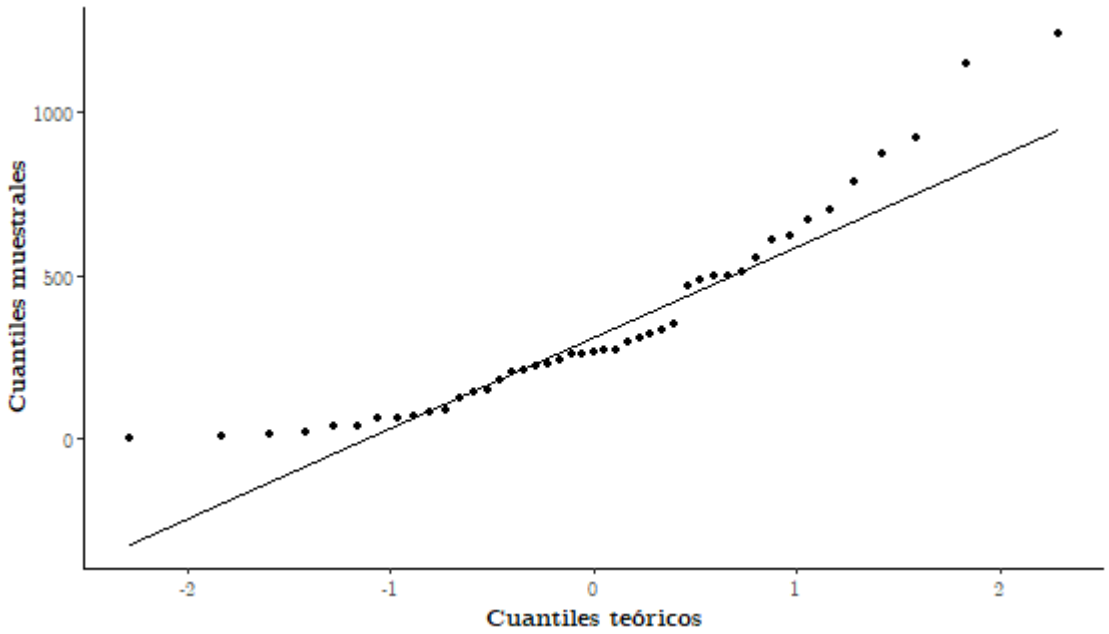


Figura 8.6 *Diagrama de probabilidad Normal construido con los tiempos hasta la falla de n = 45 dispositivos electrónicos analizados*

Es importante tener en cuenta que existen herramientas inferenciales (no paramétricas) que se utilizan para concluir sobre la distribución de probabilidades de la población. En este capítulo solo se utilizan las herramientas descriptivas para este propósito.

En resumen, la etapa Análisis de los datos comienza con el tratamiento primario de los mismos. Luego se realiza el análisis descriptivo que permite observar los primeros resultados y obtener conclusiones preliminares. A continuación, se verifican los requerimientos necesarios para aplicar herramientas inferenciales (como intervalos de confianza o de tolerancia), comprobando que se cumplan al menos de forma aproximada. Si estos requerimientos se satisfacen, se procede a calcular dichos intervalos. Los resultados obtenidos constituyen la base para la etapa siguiente, Obtención de conclusiones, en la que se interpretan en el contexto del problema bajo estudio.

Con lo visto hasta ahora, se pueden resolver las Actividades 16 a 29, Sección 8.7.

8.6 Síntesis

Existen distintas herramientas para realizar el análisis inferencial. Cuando el objetivo es estimar los parámetros de la distribución de probabilidades de una variable de interés, los intervalos de confianza constituyen una herramienta fundamental. Esta estimación se basa en el valor de un estadístico (estimador) calculado a partir de una muestra, e incorpora una cota de error (o margen de error) y un riesgo asociado al uso de muestras aleatorias.

Antes de seleccionar la muestra, cuando aún existe incertidumbre sobre qué unidades la integrarán, se define un intervalo aleatorio. Una vez tomada la muestra y observadas las unidades seleccionadas, se calcula el intervalo de confianza.

Previo a la toma de la muestra, se fija la probabilidad $(1-\alpha)$ de que el intervalo de confianza que se obtenga cubra al valor del parámetro poblacional. Tomada la muestra y obtenido el intervalo, este cubre o no al parámetro, es decir, ya no existe aleatoriedad y no puede hablarse de probabilidad en sentido estricto. Se dice que el intervalo cubre al parámetro con una confianza de $(1-\alpha)\,100\,\%$.

El valor de la cota de error depende de la probabilidad $(1-\alpha)$ definida sobre la distribución muestral del estimador correspondiente, sin necesidad de conocer el valor del parámetro a estimar. La amplitud del intervalo (asociada a la cota), el tamaño muestral y el nivel de confianza están relacionados entre sí y es posible determinar el tamaño de muestra requerido fijando previamente las otras dos magnitudes.

Otra herramienta de inferencia son los intervalos de tolerancia, que no brindan información sobre parámetros, sino sobre posibles valores de la variable en la población.

La aplicación de las diferentes herramientas inferenciales supone ciertos requerimientos cuyo cumplimiento debe verificarse previamente para poder obtener luego conclusiones válidas. Uno de estos requerimientos puede ser la normalidad de la variable de interés, cuyo cumplimiento se estudia a partir de las observaciones muestrales, de manera descriptiva o inferencial (la inferencial se ve en el Capítulo 9).

En la Tabla 8.3 se presentan las expresiones de los intervalos aleatorios y de confianza considerados en este texto.

Tabla 8.3 *Expresión del intervalo aleatorio y del intervalo de confianza para μ, σ^2 y π*

Parámetro	**Expresión del intervalo**	**Intervalo aleatorio de nivel** $(1-\alpha)$	**Intervalo del** $(1-\alpha)\,100\,\%$ **de confianza**
μ cuando se conoce σ	bilateral	$\bar{Y} \pm z_{1-\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}$	$\bar{y} \pm z_{1-\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}$
	unilateral a la derecha	$\left(\bar{Y} - z_{1-\alpha} \dfrac{\sigma}{\sqrt{n}} \; ; \; \infty\right)$	$\left(\bar{y} - z_{1-\alpha} \dfrac{\sigma}{\sqrt{n}} \; ; \; \infty\right)$
	unilateral a la izquierda	$\left(-\infty \; ; \; \bar{Y} + z_{1-\alpha} \dfrac{\sigma}{\sqrt{n}}\right)$	$\left(-\infty \; ; \; \bar{y} + z_{1-\alpha} \dfrac{\sigma}{\sqrt{n}}\right)$
μ cuando no se conoce σ	bilateral	$\bar{Y} \pm t_{n-1,1-\frac{\alpha}{2}} \dfrac{S}{\sqrt{n}}$	$\bar{y} \pm t_{n-1,1-\frac{\alpha}{2}} \dfrac{s}{\sqrt{n}}$
	unilateral a la derecha	$\left(\bar{Y} - t_{n-1,1-\alpha} \dfrac{S}{\sqrt{n}} \; ; \; \infty\right)$	$\left(\bar{y} - t_{n-1,1-\alpha} \dfrac{s}{\sqrt{n}} \; ; \; \infty\right)$
	unilateral a la izquierda	$\left(-\infty \; ; \; \bar{Y} + t_{n-1,1-\alpha} \dfrac{S}{\sqrt{n}}\right)$	$\left(-\infty \; ; \; \bar{y} + t_{n-1,1-\alpha} \dfrac{s}{\sqrt{n}}\right)$
σ^2	bilateral	$\left(\dfrac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}} \; ; \; \dfrac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}\right)$	$\left(\dfrac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}} \; ; \; \dfrac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}\right)$
	unilateral a la derecha	$\left(\dfrac{(n-1)S^2}{\chi^2_{1-\alpha}} \; ; \; \infty\right)$	$\left(\dfrac{(n-1)s^2}{\chi^2_{1-\alpha}} \; ; \; \infty\right)$
	unilateral a la izquierda	$\left(0 \; ; \; \dfrac{(n-1)S^2}{\chi^2_{\alpha}}\right)$	$\left(0 \; ; \; \dfrac{(n-1)s^2}{\chi^2_{\alpha}}\right)$
π	bilateral	$f_r \pm z_{1-\frac{\alpha}{2}} \sqrt{\dfrac{f_r(1-f_r)}{n}}$	$f_o \pm z_{1-\frac{\alpha}{2}} \sqrt{\dfrac{f_o(1-f_o)}{n}}$
	unilateral a la derecha	$\left(f_r - z_{1-\alpha} \sqrt{\dfrac{f_r(1-f_r)}{n}} \; ; \; 1\right)$	$\left(f_o - z_{1-\alpha} \sqrt{\dfrac{f_o(1-f_o)}{n}} \; ; \; 1\right)$
	unilateral a la izquierda	$\left(0 \; ; \; f_r + z_{1-\alpha} \sqrt{\dfrac{f_r(1-f_r)}{n}}\right)$	$\left(0 \; ; \; f_o + z_{1-\alpha} \sqrt{\dfrac{f_o(1-f_o)}{n}}\right)$

Observaciones:

- En cada caso se indica en primer lugar la expresión del intervalo bilateral; luego la del intervalo unilateral a la derecha y finalmente, la del intervalo unilateral a la izquierda.
- En el caso de la media:
  - Cuando el desvío estándar poblacional σ es conocido, se utiliza la distribución Normal estándar; cuando es desconocido, se emplea la distribución t de Student con $n-1$ grados de libertad.
  - El uso exacto de la distribución t de Student requiere normalidad de la población. En la práctica, puede utilizarse de manera aproximada cuando la distribución es aproximadamente simétrica y el tamaño muestral no es pequeño (típicamente $n > 10$).
  - En distribuciones fuertemente asimétricas, la aproximación puede no ser adecuada incluso para tamaños muestrales grandes, pudiendo obtenerse intervalos con niveles de confianza inferiores al nivel nominal.

- En el caso de la variancia:
  - Se brindan las expresiones correspondientes a los intervalos (bilateral y unilaterales) basados en el estadístico pivote χ^2, válidos bajo el supuesto de normalidad de la variable de interés.
  - Estos intervalos son sensibles a desviaciones respecto de la normalidad, y su desempeño no mejora necesariamente con el aumento del tamaño muestral. En tales situaciones, puede considerarse la construcción de intervalos alternativos, por ejemplo mediante el método de Bonett.
  - Los intervalos de confianza para el desvío estándar se obtienen aplicando la raíz cuadrada a los extremos de los intervalos correspondientes a la varianza.
  - En los intervalos unilaterales a la izquierda, el extremo inferior vale 0, ya que $\sigma^2 \geq 0$.
- En el caso de la proporción:
  - Se brindan las expresiones correspondientes a los intervalos (bilateral y unilaterales) basados en la aproximación normal.
  - Los intervalos exactos se obtienen mediante métodos computacionales.
  - En los intervalos unilaterales a la izquierda, el extremo inferior vale 0, y en el caso de intervalos unilaterales a la derecha, el extremo superior vale 1, ya que $0 \leq \pi \leq 1$.

8.7 Actividades propuestas

1. En una empresa metalúrgica dedicada a la fabricación de recipientes sometidos a presión, la resistencia al estallamiento de las soldaduras es una característica crítica para garantizar la seguridad y la calidad del producto final. Con el objetivo de evaluar el desempeño de un determinado tipo de soldadura, se realizó un estudio en el que se midió la resistencia al estallamiento, expresada en $lb/pulg^2$, de las soldaduras correspondientes a un conjunto de recipientes seleccionados al azar del proceso productivo. Los valores obtenidos se presentan a continuación:

   7200 - 6100 - 7300 - 7300 - 8000 - 7400 - 7300 - 7300 - 8000 - 6700.

   Suponga que las 10 observaciones obtenidas se pueden considerar una muestra aleatoria de una población con distribución Normal y que se conoce que el desvío estándar de la resistencia al estallamiento para ese tipo de soldaduras (σ) es de 500 $lb/pulg^2$.
   Con la información recabada, se obtuvo la siguiente salida de R.

```
summary(resistencia)
    resistencia
 Min.   : 6100
 1st Qu.: 7225
 Median : 7300
 Mean   : 7260
 3rd Qu.: 7375
 Max.   : 8000
```

   a) Defina población y variable de interés.
   b) Construya e interprete un intervalo de confianza del 95 % para la resistencia promedio al estallamiento en ese tipo de soldaduras.
   c) Antes de tomar la muestra, ¿qué significa el valor 0,95?
   d) Con la misma información se construyeron los intervalos del 90 %, 92 % y 99 % de confianza. Los intervalos fueron (6983; 7537), (6853,65; 7666,35) y (6999,11; 7520,89). Sin hacer cálculos, establezca la correspondencia entre los intervalos y los niveles de confianza mencionados. Justifique.

2. Reconsidere la misma situación mencionada en la actividad anterior, relativa a las soldaduras.
   a) Construya un intervalo del 95 % de confianza unilateral que le permita decidir si la resistencia promedio es superior a 6800 $lb/pulg^2$.
   b) Analice el intervalo obtenido e informe sus conclusiones.

3. En una empresa en la que se fabrican tapas metálicas para bebidas gaseosas, un aspecto importante es el peso (en mg) de la capa de PVC que lleva cada una, la cual determina el espesor de la película que hace que la botella quede bien cerrada.
   Se conoce por experiencia que este peso se comporta según el modelo Normal con un desvío estándar (σ) de 5 mg.
   En la empresa quieren conocer el valor del peso promedio de las capas y para ello toman una muestra aleatoria de 40 tapas.
   Con la información del peso de cada una de las tapas de la muestra, se utiliza el siguiente comando de R:

```
MeanCI(pesocapa, sd=5, method ="classic", conf.level =0.99)
```

Finalmente, se obtiene el intervalo de confianza $(213,416; 217,489)$

a) Defina población y variable de interés.
b) ¿Cuál es el parámetro de interés? Defínalo e indique cuánto vale la estimación puntual del mismo.
c) Observe el intervalo de confianza presentado en la salida. Interprételo en términos del problema.
d) Indique el valor de la cota superior del error de estimación.
e) ¿Qué hubiera ocurrido con el valor de dicha cota si se hubieran muestreado 60 tapas? Explique y relacione con la precisión de la estimación.
f) Suponga que en la empresa desean realizar un nuevo estudio que les permita estimar al promedio poblacional con un 99 % de confianza y un máximo error de estimación de 1 mg. ¿Cuántas tapas deben seleccionar como mínimo en este caso?

4. En una determinada zona (Zona 1) se conoce, por experiencia, que la concentración media de dióxido de carbono en el aire (en partes por millón en volumen) es de 420. En un estudio interesa conocer, para otra zona (Zona 2), la concentración media de dióxido de carbono en el aire. Para ello se seleccionan aleatoriamente 20 puntos en esa nueva zona y se mide en cada uno la concentración de dióxido de carbono, obteniendo un promedio de 378 ppmv y un desvío estándar de 80 ppmv. Suponga que la variable de interés se comporta según el modelo Normal.
   a) Defina población y variable. Mencione el parámetro de interés.
   b) Estime la concentración media de dióxido de carbono en la Zona 2 con un nivel de confianza del 98 %.
   c) Explicite los requerimientos que debieron verificarse para que esta estimación sea válida.
   d) Señale en el contexto del problema cuáles de los datos dados en el enunciado son parámetros y cuáles son estadísticos. Para el caso de los parámetros indique a qué población corresponden.
   e) ¿Considera que la concentración media de dióxido de carbono en la Zona 2 es menor que la concentración media en la Zona 1? Plantee el intervalo unilateral correspondiente y concluya.

5. En una industria automotriz, los ingenieros a cargo de uno de los procesos de ensamblado están investigando el tiempo (en segundos) que demanda dicho proceso. Por experiencia se puede considerar que esa variable se comporta según el modelo normal.
   Se toma una muestra aleatoria simple de 20 realizaciones del proceso y los tiempos de cada una de ellas son los siguientes:

   93- 90 - 97 - 90 - 93 - 91 - 96 - 94 - 91 - 91 - 88 - 93 - 95 - 91 - 89 - 92 - 87- 88 - 90 - 86.

   La siguiente es la salida de R, obtenida con los 20 tiempos:

```
mean(tiempo)
91.25

sd(tiempo)
2.936
```

```
MeanCI(tiempo, sd=NULL, method ="classic", conf.level =0.98)
 mean    lwr.ci    upr.ci
91.25    89.583    92.917
```

a) Defina población y variable.

b) Se desea estimar el tiempo promedio de ensamblado con un nivel de confianza del 98 %.

1) Indique cuál es el estadístico que brinda información sobre este parámetro.
2) Interprete el intervalo de confianza correspondiente en términos del problema.

c) Indique porqué las siguientes afirmaciones son falsas.

1) Si se deseara un nivel de confianza superior, el intervalo resultaría más preciso.
2) Manteniendo fijo el nivel de confianza del 98 %, si aumenta el tamaño de la muestra siempre se obtendrá un intervalo de confianza más preciso.

d) Suponga que se desea realizar un nuevo estudio para estimar el tiempo promedio de ensamblado con el 98 % de confianza y un máximo error de estimación de 0,8 s. Obtenga el mínimo tamaño de muestra requerido para ello (considere a estos 20 tiempos analizados como un estudio piloto).

6. Reconsidere la situación de la Actividad 5. Un encargado del proceso observa el intervalo obtenido y afirma que el 98 % de las veces que se realiza el ensamblado, el tiempo oscila entre 89,583 s y 92,917 s.
   Usted, ¿qué opina de esa afirmación? Justifique.

7. Un proceso produce cierto tipo de cojinetes cuyo diámetro interior debe ser, en promedio, igual a 3 centímetros. Por experiencia los fabricantes conocen que el diámetro se comporta según el modelo Normal pero no están seguros del valor del diámetro promedio y para averiguarlo deciden llevar a cabo un estudio estadístico.
   En el marco del mismo, seleccionan en forma aleatoria, 12 cojinetes y en cada uno miden su diámetro interior, obteniendo los siguientes valores:

   $3,01 - 3,05 - 2,99 - 3,03 - 3,00 - 3,02 - 3,05 - 2,99 - 3,04 - 2,97 - 3,02 - 3,01.$

   A continuación se presentan algunas salidas obtenidas con R.

```
summary(cojinetes)
Min. 1st Qu.   Median   Mean   3rd Qu.    Max.
2.970   2.998   3.015   3.015   3.033   3.050
sd(cojinetes)
0.025
MeanCI(cojinetes, method = "classic")
mean   lwr.ci   upr.ci
3.015   2.999   3.031
```

a) Defina población, variable y parámetro de interés.

b) Indique un posible objetivo del estudio estadístico. Expréselo en función del parámetro definido en el ítem anterior.

c) Interprete la estimación obtenida en el contexto del problema. ¿Qué puede decir del cumplimiento del requisito establecido para el diámetro interior promedio de los cojinetes? Justifique.

d) ¿La estimación obtenida es válida? Explique.

8. Reconsidere el proceso de producción de los cojinetes de la Actividad 7.
   a) Estime el desvío estándar de los diámetros interiores con un nivel de confianza del 95 %.
   b) Interprete la estimación obtenida en términos del problema.
   c) Indique qué ocurrirá con la amplitud del intervalo si se considera un nivel de confianza del 99 % y se mantiene el mismo tamaño de la muestra.

9. En un proceso de producción de válvulas especiales, se concluyó que se debía reducir la variancia de los diámetros de las mismas, que en ese momento era de 16 mm^2. Con ese objetivo se dispusieron algunos cambios en el proceso de producción, después de una investigación detallada sobre los factores que influían en esta variabilidad.
   Las modificaciones se considerarán efectivas si se logra una reducción de 40 % (o mayor) en el desvío estándar.
   Una vez implementadas las modificaciones, y con el proceso ya estabilizado, se tomó una muestra de 60 válvulas para evaluar si estas modificaciones resultaron efectivas.
   Con la información recabada, se concluyó que no se puede descartar la normalidad para los diámetros de las válvulas y se obtuvo un promedio muestral de 59,94 mm y un desvío estándar de 2,5 mm.
   a) Defina población y variable de interés.
   b) Plantee el objetivo del estudio estadístico en términos del parámetro de interés.
   c) ¿Considera que las modificaciones llevadas a cabo en el proceso lograron reducir la variabilidad de los diámetros de las válvulas? Justifique su respuesta
   d) ¿Considera que las modificaciones fueron efectivas?

10. En una investigación hidrográfica se utiliza un telémetro de láser manual de cierta marca. Con ese equipo se realizan 15 mediciones de una distancia conocida e igual a 500 metros. Los valores asociados a dichas mediciones, en m, son los siguientes:

    499,9 - 499,98 - 500,10 - 499,97 - 500,09 - 500,01 - 499,95 - 500,05 - 499,94 - 500,01 - 500,03 - 500,06 - 500,02 - 499,93 - 500,03.

    Con estos valores se calcularon algunos estadísticos:

    Promedio muestral: 500,0047 m
    Desvío estándar muestral: 0,059 m

    a) Defina población y variable de interés.
    b) Indique qué modelo se puede considerar para la variable de interés. Justifique.
    c) Estime el valor promedio de las mediciones de una distancia de 500 m realizadas con el equipo mencionado, con un nivel de confianza del 97 %. Interprete en términos del problema.
    d) ¿Podría afirmar que el equipo es exacto? Si necesita definir algún criterio para responder, hágalo y luego responda.
    e) Interesa analizar si el equipo es, además, preciso. Suponga que el equipo se considerará preciso si el desvío estándar de los resultados de las mediciones resulta menor que 0,07 m.
    Con `R` se obtuvo lo siguiente:

```
VarCI(errores, method = "classic")
var     lwr.ci   upr.ci
0.0035  0.0019   0.0086
```

```
VarCI(errores, method = "classic", sides = "left")
var     lwr.ci   upr.ci
0.0035  0.0021   Inf

VarCI(errores, method = "classic", sides = "right")
var     lwr.ci   upr.ci
0.0035     0     0.0074
```

1) Indique cuál es el intervalo apropiado para tomar la decisión sobre la precisión del equipo de medición. Justifique.
2) Interprete el intervalo elegido, en contexto e informe sobre la precisión de ese equipo de medición.

11. En una empresa productora de placas de madera consideran buenas a aquellas unidades que presentan como máximo 1 imperfección.
Con el objetivo de conocer la proporción de placas buenas, seleccionaron una muestra aleatoria simple de 200 placas y encontraron que 182 placas eran buenas.
   a) Defina población, variable y parámetro de interés.
   b) Estime la proporción de placas buenas con un nivel de confianza del 95 %, usando la aproximación normal. Interprete.
   c) ¿Qué ocurrirá con la precisión del intervalo de confianza si se decide aumentar el nivel de confianza? Explique.
   d) Si se desea reducir el error de estimación a la mitad, manteniendo el nivel de confianza original, ¿cuántas placas adicionales deberán analizarse? (Considere a las 200 placas como una muestra piloto).

12. Una maderera minorista inspecciona los embarques que le llegan de sus proveedores. En cada uno, las hojas se pueden considerar de primera o de segunda calidad.
Para un embarque particular, que trae decenas de miles de hojas de una madera de calidad selecta, el supervisor selecciona una muestra aleatoria de 144 hojas y en dicha muestra, encuentra que 18 hojas no pueden venderse como de primera calidad. La siguiente es una salida de R obtenida con esta información.

```
BinomCI(18, 144, method = "clopper-pearson",conf.level = 0.99)
 est     lwr.ci    upr.ci
0.125     0.064    0.212
BinomCI(18, 144, method = "clopper-pearson",
conf.level = 0.99, sides = "left")
 est     lwr.ci upr.ci
0.125     0.069    1
BinomCI(18, 144, method = "clopper-pearson",
conf.level = 0.99, sides ="right")
 est     lwr.ci    upr.ci
0.125       0       0.203
```

   a) Defina población, variable y parámetro de interés.

b) Mencione cuál es el estimador de dicho parámetro e informe qué valor tomó en la muestra.
c) Interesa estimar a ese parámetro a través de un intervalo de confianza exacto. Seleccione el intervalo correspondiente e interprételo en contexto.
d) Si el embarque tiene más del 15 % de hojas de segunda calidad, no es rentable para la empresa minorista.Seleccione el intervalo unilateral que permita concluir si el embarque es rentable o no. Analícelo e informe su decisión.

13. Un fabricante de autopartes utiliza un conjunto de pruebas para evaluar a sus productos. Todas las autopartes deben pasar las pruebas antes de ser enviadas al cliente. En una muestra aleatoria de 500 autopartes, 15 fallan en alguna de las pruebas, lo que permite obtener los siguientes intervalos de confianza para la proporción de autopartes que no pasan la prueba:

```
BinomCI(15, 500, method = "clopper-pearson")
 est     lwr.ci    upr.ci
0.03     0.017      0.049
BinomCI(15, 500, method = "clopper-pearson", sides = "left")
 est     lwr.ci    upr.ci
0.03     0.019         1
BinomCI(15, 500, method = "clopper-pearson", sides = "right")
 est     lwr.ci    upr.ci
0.03       0        0.046
```

a) Realice el planteo del problema
b) ¿Puede afirmar que la proporción de autopartes que fallan es superior a 0,025? Justifique claramente su respuesta.
c) ¿Qué ocurrirá con la precisión de la estimación obtenida en el punto anterior.
   1) si se aumenta el nivel de confianza manteniendo fijo el tamaño muestral?
   2) si se aumenta el tamaño de muestra, manteniendo fijo el nivel de confianza?

14. En una empresa están interesados en comprar componentes electrónicos que duren más de 1500 horas (aunque admiten que como máximo un 5 % de los mismos no alcance este valor). Uno de los posibles proveedores le informa que en el último tiempo realizaron un estudio estadístico sobre la duración de los componentes que fabrican y obtuvieron lo siguiente:

$IC_{95\%,\mu} = (1550, \infty)$.

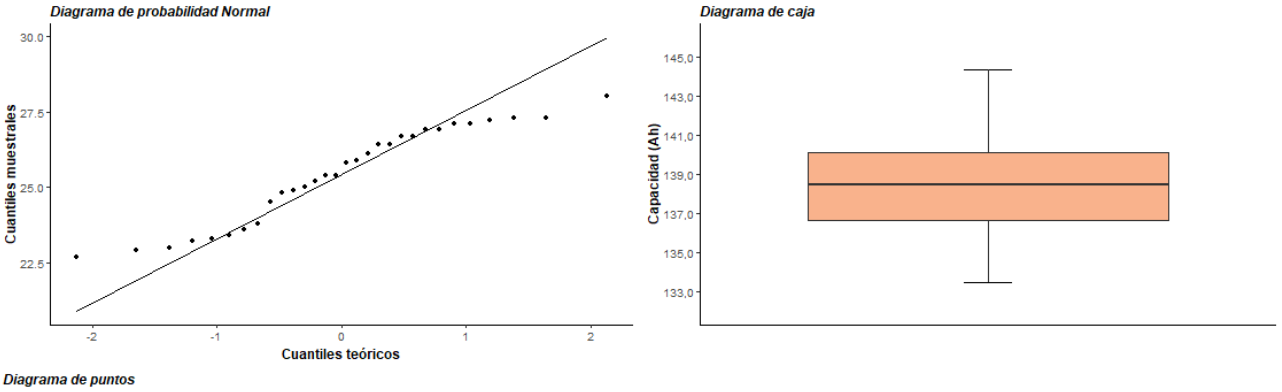
a) A partir de esta información, ¿considera que puede comprar los componentes a ese proveedor? Justifique su respuesta claramente.
b) Si su respuesta es "NO" indique sobre qué otro parámetro le interesaría contar con información para poder decidir.
c) ¿Qué información le brindaría un intervalo de tolerancia en este caso? Comente e indique en este caso qué valores de P y de $(1-\alpha)$ utilizaría.

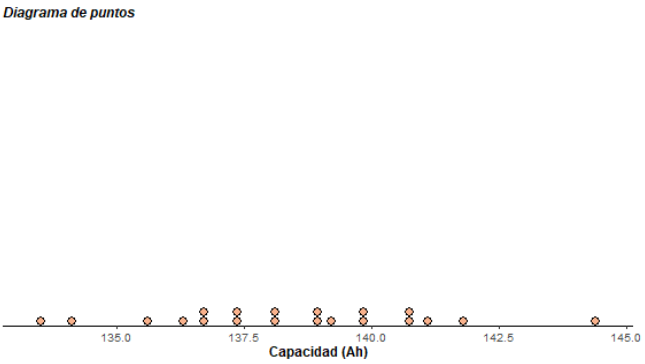
15. En una empresa productora de varillas quieren conocer los valores que toman los diámetros (en milímetros) de las mismas. Con la información de una muestra de tamaño 25 se obtuvo la siguiente salida de R.

```
normtol.int(diametro, alpha = 0.02, P = 0.99, side = 2)
alpha    P    x.bar 2-sided.lower 2-sided.upper
0.02   0.99   25.43       22.157      28.703
```

a) Defina población y variable de interés.
b) ¿Qué significa el valor 0,02 antes de tomar la muestra?
c) ¿Qué significa el valor 0,99 en el contexto del problema?
d) Interprete el intervalo de tolerancia brindado por R en contexto.
e) ¿Debe verificarse algún requerimiento para que este intervalo sea válido? Explique.

16. Un fabricante asegura que la capacidad media (en amperios-hora) de cierto tipo de baterías que produce la compañía es de al menos 140. Un grupo para la defensa del consumidor desea probar la credibilidad de la afirmación del fabricante y mide la capacidad de 20 baterías de ese tipo seleccionadas al azar, obteniendo lo siguiente.






```
mean(capacidad)
138.47
sd(capacidad)
2.659
MeanCI(capacidad, method = "classic")
mean     lwr.ci   upr.ci
138.47   137.226  139.714
MeanCI(capacidad, method = "classic", sides = "left")
mean     lwr.ci   upr.ci
138.47   137.442  Inf
MeanCI(capacidad, method = "classic")
mean     lwr.ci   upr.ci
138.47   -Inf     139.498
VarCI(capacidad, method = "classic")
var      lwr.ci   upr.ci
7.069    4.089    15.081
```

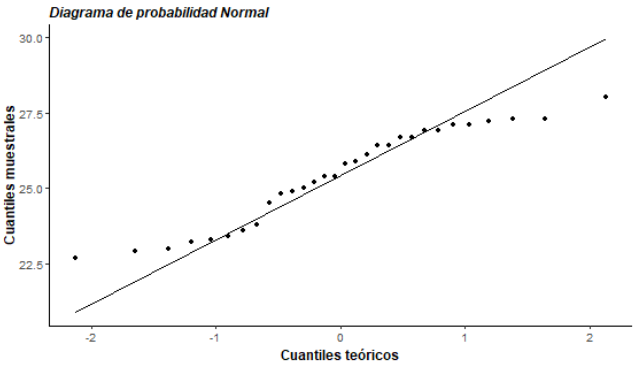
a) Para realizar inferencias acerca de la capacidad media, ¿debe verificarse algún requerimiento? Si su respuesta es afirmativa indique cuál/es y cómo procedería para evaluar su validez.

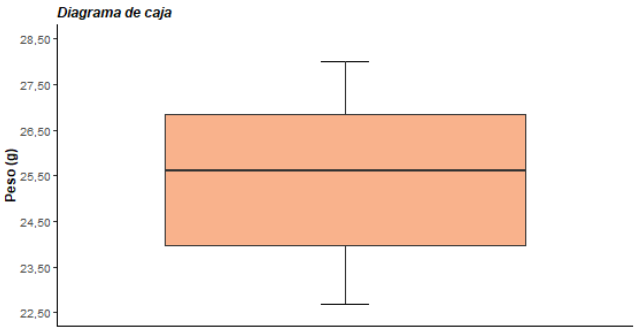
b) Evalúe la afirmación del fabricante.

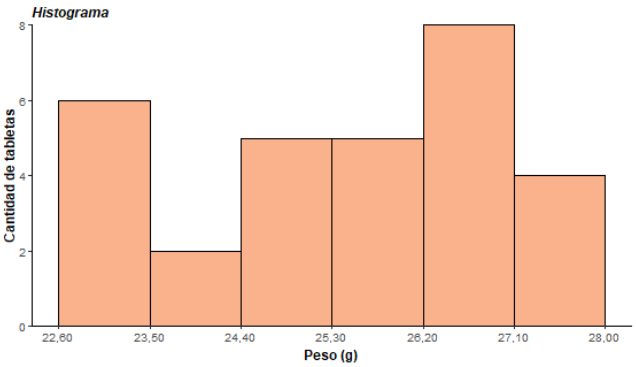
c) Estime el desvío estándar de la capacidad de las baterías.

17. Un laboratorio produce un cierto tipo de tabletas, cuya característica de interés es el peso (en gramos). Por normas de sus clientes, el desvío estándar del peso de las tabletas no debe superar 1,8 g.

En el laboratorio tomaron una muestra aleatoria simple de 30 tabletas y registraron el peso de cada una de ellas. Algunos resultados fueron los siguientes:








```
summary(peso)
Min. 1st Qu.   Median   Mean   3rd Qu.    Max.
22.70   23.98   25.60   25.43   26.85    28.00

VarCI(peso, method = "classic")
var   lwr.ci   upr.ci
2.542  1.612   4.594

VarCI(peso, method = "classic", sides = "right")
var   lwr.ci   upr.ci
2.542  0.00    4.163

VarCI(peso, method = "classic", sides = "left")
var   lwr.ci   upr.ci
2.542  1.732    Inf
```

```
VarCI(peso, method = "bonett")
var   lwr.ci   upr.ci
2.542  1.876   3.943

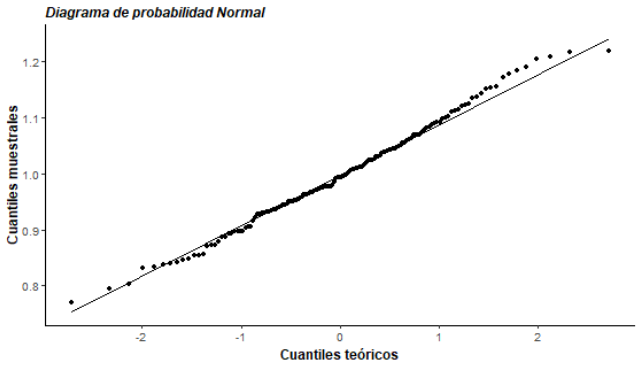
VarCI(peso, method = "bonett", sides = "right")
var   lwr.ci   upr.ci
2.542  0.00    3.660

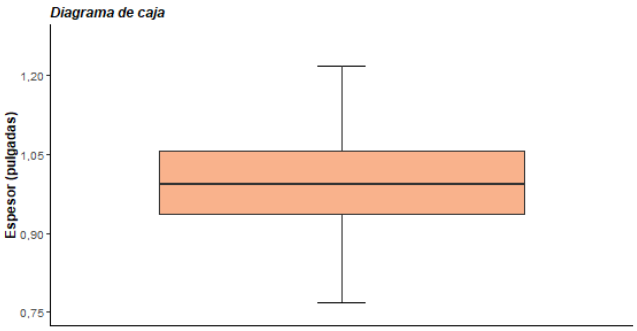
VarCI(peso, method = "bonett", sides = "left")
var   lwr.ci   upr.ci
2.542  1.976    Inf
```

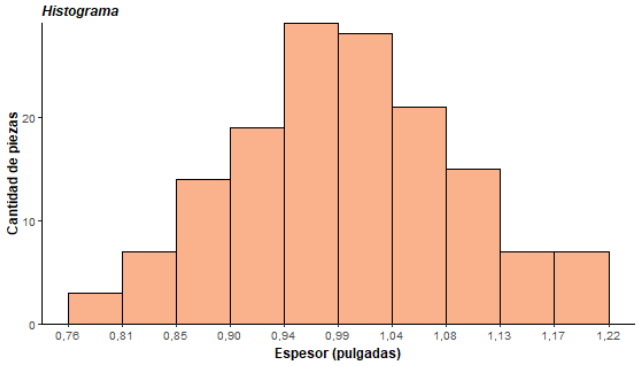
a) Realice el planteo del problema.
b) Comente algunos resultados que surgen del análisis descriptivo.
c) En este caso, ¿considera que el intervalo de confianza debe basarse en la distribución χ^2 o construirse con el método de Bonett? Justifique su respuesta.
d) Encuentre el intervalo unilateral apropiado para decidir si se cumple con las normas de los clientes. Justifique su elección, interprete dicho intervalo e informe su decisión.

18. Reconsidere la Actividad 29 (Capítulo 2, página 120), referida a una empresa de servicios informáticos que está considerando la contratación de un nuevo operador.
    a) Realice el planteo completo del problema y comente algunas cuestiones tenidas en cuenta en la etapa de planificación.
    b) Analice exhaustivamente la información obtenida e interprete los resultados en contexto.
    c) ¿Qué decisión deben tomar en la empresa en relación a la contratación de un nuevo operador?
    d) Indique si sus conclusiones son preliminares o definitivas.

19. Una empresa fabrica piezas moldeadas por inyección para un cliente que las compra en grandes cantidades. Una de las dimensiones críticas de estas piezas es su espesor (medido en pulgadas), el cual debe estar en el intervalo (0,97; 1,03)". La empresa pretende que como mínimo el 99,73 % de las piezas que produce tengan espesor dentro del intervalo mencionado.
    a) Si se puede suponer que el espesor de las piezas moldeadas (X) se comporta según el modelo Normal, ¿cuánto deberían valer el espesor promedio y el desvío estándar para que se cumpla con lo que pretende la empresa? Justifique.
    b) Interesa analizar si, con el proceso tal y como está funcionando actualmente, se puede cumplir con lo pretendido. Realice el planteo completo del problema. Incluya población, variable, parámetro de interés y objetivo del estudio estadístico.
    c) En el marco del estudio estadístico llevado a cabo, se tomó una muestra aleatoria simple de n = 150 piezas moldeadas y se obtuvo la salida que se presenta a continuación:








```
summary(espesor)
Min. 1st Qu.   Median   Mean   3rd Qu.    Max.
0.769  0.936   0.994    0.998   1.058     1.219

MeanCI(espesor, method = "classic", conf.level = 0.99)
var   lwr.ci   upr.ci
0.998  0.977   1.018

VarCI(espesor, method = "classic", conf.level = 0.99)
var   lwr.ci   upr.ci
0.009  0.0068  0.0124
```

1) ¿Cómo haría para detectar algún problema que indique que la muestra no es una muestra aleatoria simple?
2) ¿Podría indicar si la muestra proviene de una población con distribución normal?
3) Comente otras cuestiones que surgen del análisis descriptivo.
4) Analice exhaustivamente los datos e informe sus conclusiones.

20. Para estimar el valor de una magnitud δ (en gramos), se realizaron 15 mediciones repetidas de la misma con un instrumento que es exacto y se construyó el intervalo de confianza de 95 % para el promedio de las mediciones, que resultó (8,96; 9,01) g.
    a) Un estudiante interpreta este resultado como indicativo de que el promedio de las 15 mediciones tomará valores entre 8,96 g y 9,01 g con probabilidad 0,95. Otro estudiante interpreta que el 95 % de las mediciones están entre 8,96 g y 9,01 g. Analice si alguna de estas interpretaciones es correcta. En caso afirmativo, indique cuál; de lo contrario, dé su propia interpretación.
    b) Otro estudiante realiza una nueva medición de la magnitud δ que resulta igual a 9,02 g. En base a este resultado cuestiona los valores del intervalo de confianza ¿Está acertado en

este cuestionamiento? Justifique su respuesta.

c) ¿Cuántas mediciones deberán agregarse si, manteniendo la confianza, se desea disminuir el error de estimación para δ en 0,01 g?

d) Suponga que el instrumento utilizado se considera preciso si el parámetro correspondiente resulta menor que 0,08 g. Con la información obtenida, ¿puede Ud. informar si el instrumento es preciso o no? Justifique su respuesta.

Las siguientes actividades se refieren a los problemas presentados en la Sección 1.4 (página 7). Para algunos de ellos se fueron realizando tareas de las distintas etapas del ciclo PPDAC a lo largo de todo el texto. Se sugiere revisar todo lo realizado, completar las tareas que falten y luego realizar un breve informe con las principales conclusiones en contexto. Los datos se presentan en el archivo https://sites.google.com/view/libro-estadistica.

21. Considere el Problema 1 que se desarrolla en la empresa metalúrgica que fabrica barras metálicas. En la Tabla 1.3 se presentan algunas tareas que se desarrollan en su proceso de resolución, de acuerdo al ciclo PPDAC.

    a) Vuelva a escribir el proceso de resolución, completando con información relativa a las etapas de planificación y análisis de los datos, de acuerdo a lo efectivamente realizado.

    b) Plantee una situación de contexto de tal manera que, observando el intervalo de confianza obtenido, la conclusión sea que no hay que realizar ajustes en el proceso.

22. Reconsidere el Problema 1. Suponga que la longitud de las barras se comporta según el modelo Normal.

    a) Indique qué valor deben tener la longitud promedio (μ) y el desvío estándar (σ) si se pretende que como mínimo el 99,73 % de las barras cumplan con las especificaciones impuestas por la industria automotriz.

    b) Plantee nuevamente el problema, ahora en función de estos parámetros.

    c) Siga adelante con las otras tareas del ciclo PPDAC hasta obtener conclusiones en contexto.

23. Considere el Problema 2 sobre las placas de madera de un lote.

    a) Realice el planteo del problema y comente algunas cuestiones que se hayan definido en la etapa de planificación.

    b) ¿Por qué en este caso no se aplicaron herramientas inferenciales? ¿En qué caso se hubieran aplicado? Justifique.

24. Considere el Problema 3 relativo a un servidor neumático para la siembra mecánica de semillas.

    a) Realice el planteo completo del problema.

    b) Siga adelante con las restantes tareas del ciclo PPDAC hasta obtener conclusiones en contexto.

25. Considere el Problema 4 que tiene lugar en la empresa dedicada a la comercialización de productos en línea.

    a) En la empresa deben decidir si contratar o no a una nueva prestadora para realizar las entregas. Realice un planteo completo del problema y escriba el objetivo en función de algún/os parámetros de interés.

    b) Siga adelante con las otras tareas del ciclo PPDAC hasta obtener conclusiones en contexto.

26. Considere el Problema 5 que se refiere a una empresa que fabrica cables eléctricos de diferente tipo. La empresa está interesada en estudiar el comportamiento del número de imperfecciones en rollos de cable unipolar de 2,5 mm. Como tienen la posibilidad de abastecer a un cliente muy importante, pretenden que la proporción de rollos en los que se observan al menos 3 imperfecciones sea menor que 0,08 y llevan adelante un estudio para saber si sus pretensiones son razonables.
    a) Plantee el problema y mencione dos decisiones que se hayan tomado en la etapa de planificación.
    b) Analice los datos e informe sus conclusiones en relación al objetivo planteado.
    c) Suponga que se plantea otro objetivo: que por lo menos la mitad de los rollos tengan a lo sumo una imperfección. Plantee el nuevo parámetro de interés, analice nuevamente los datos e informe sus concusiones.
    d) ¿De qué depende la validez de los intervalos de confianza construidos en la resolución de esta actividad? Comente.

27. Considere el Problema 6, relativo a un comitente que quiere conocer ciertas características de un tramo de ruta de 5 km ya construido. En particular, para la resistencia interesa que como máximo el 5 % de los puntos de ese tramo tengan resistencia inferior a 30 MPa, y para el espesor interesa que su promedio sea como mínimo de 22 cm y que no haya puntos con espesor inferior a 19 cm.
    a) Plantee las variables y los parámetros de interés y escriba el objetivo del estudio en términos de dichos parámetros.
    b) Estime a los parámetros de interés e interprete las estimaciones en contexto.
    c) ¿Considera necesario construir algún intervalo de tolerancia en este caso? Explique.
    d) Informe si la ruta cumple con las exigencias de calidad en cuanto a espesor y resistencia.

28. Considere el Problema 7. Suponga que Ud. es la persona que tuvo a cargo la evaluación del multímetro y que para poder concluir si el mismo era exacto y preciso realizó 20 mediciones de un voltaje conocido e igual a A (en V). Los datos obtenidos se presentan a continuación.

    13,013 - 9,772 - 10,507 - 11,316 - 12,686 - 10,432 - 11,164 - 10,785 - 12,321 - 12,090 - 11,256 - 11,415 - 10,778 - 11,834 - 10,004 - 11,219 - 11,922 - 11,733 - 11,940 - 11,068

    a) Realice el planteo del problema y comente algunas cuestiones que se tuvieron en cuenta en la etapa de planificación.
    b) Indique un posible valor de "A" si la conclusión obtenida, con cierto nivel de confianza, es que no se puede descartar que el equipo sea exacto. Justifique.
    c) Defina un posible criterio de "instrumento preciso" si la conclusión obtenida, con cierto nivel de confianza, es que el equipo es preciso. Justifique.
    d) A partir de sus respuestas a los ítems *b* y *c*, complete el planteo del problema.

29. Considere el Problema 8, relativo al tiempo hasta la primera falla de dispositivos electrónicos. Suponga que usted es el usuario que debe decidir si comprar o no un cierto tipo de dispositivo.
    a) Plantee el problema y escriba el objetivo en función de algún parámetro de interés.
    b) Siga adelante con las restantes tareas del ciclo PPDAC hasta obtener conclusiones en contexto.

8.8 Estadística con R

8.8.1 Diagrama de probabilidad Normal

Para la construcción del diagrama de probabilidad Normal es posible recurrir a las funciones del paquete `ggplot2`. Suponga que se cuenta con un conjunto de datos (*dataset*) denominado `data` que incluye una variable denominada `y`, entonces el mencionado diagrama se puede obtener mediante el siguiente código:

```
ggplot(data, aes(sample = y)) +
  stat_qq() +  # Puntos del QQ-plot
  stat_qq_line() +  # Línea de referencia normal
  labs(x = "Cuantiles teóricos",
       y = "Cuantiles muestrales",
       title="Diagrama de probabilidad normal")+
  theme(axis.title.y = element_text(angle = 90),
        axis.title.x = element_text(vjust = 0.5, hjust = 0.5))
```



Problema 1 (pág. 7): *A continuación se muestra el código para construir el diagrama de probabilidad Normal con los datos de las longitudes de las barras que se presenta en la Figura 8.5, Sección 8.5, 421. En este caso* `datos` *es el nombre del conjunto de datos que se importó desde el archivo Excel, el cual incluye la variable denominada* `longitud`.

```
# Se utiliza la librería ggplot2 para construir el diagrama de probabilidad.

library(ggplot2)
ggplot(datos, aes(sample = longitud)) +
  stat_qq() + # Puntos del QQ-plot
  stat_qq_line() + # Línea de referencia normal
  labs(x = "Cuantiles teóricos",
       y = "Cuantiles muestrales",
       title="Diagrama de probabilidad normal")+
  theme(axis.title.y = element_text(angle = 90),
        axis.title.x = element_text(vjust = 0.5, hjust = 0.5))
```

Con la mismo paquete `ggplot2`, se pueden construir diagramas de probabilidad para otras distribuciones. Por ejemplo, si se tiene una muestra y se quiere analizar si proviene de una distribución Exponencial, se procede de la siguiente manera:

```
#En primer lugar se estima el valor del parámetro de la distribución
#Exponencial, a partir de los datos muestrales

rate_hat <- 1 / mean(x)
```

```
#Luego, se construye el diagrama de probabilidad utilizando el paquete
#ggplot2, el cual compara los cuantiles muestrales con los cuantiles
#teóricos de una distribución Exponencial ajustada mediante el parámetro
#estimado.
library(ggplot2)

ggplot(data.frame(x = x), aes(sample = x)) +
  stat_qq(distribution = qexp, dparams = list(rate = rate_hat)) +
  stat_qq_line(distribution = qexp, dparams = list(rate = rate_hat)) +
  labs(x = "Cuantiles teóricos",
       y = "Cuantiles muestrales") +
  theme_bw() +
  theme(
    panel.grid = element_blank(),
    axis.line = element_line(color = "black"),
    axis.ticks = element_line(color = "black")
  )
```

8.8.2 Intervalos de confianza

Para la construcción de los intervalos de confianza se requiere instalar la librería `DescTools`.

Algunas cuestiones generales:

- El nivel de confianza que se considera por omisión es 0,95. Cualquier otro nivel de confianza debe indicarse en la opción `conf.level`. Por ejemplo `conf.level = 0,90` indica que interesa construir un intervalo con el 90 % de confianza.
- Por omisión se considera que el intervalo de confianza es bilateral. Si interesa construir intervalos unilaterales a la derecha o a la izquierda, esto debe indicarse con la opción `sides`: `sides = c("left")` indica que interesa construir un intervalo de confianza donde solo se indique el extremo inferior o izquierdo, mientras que `sides = c("right")` indica que interesa construir un intervalo de confianza donde solo se indique el extremo superior o derecho.

A continuación se detallan las particularidades para cada parámetro de interés:

1. Intervalo de confianza para la media poblacional

Suponga que se cuenta con un conjunto de datos (*dataset*) denominado `data` que incluye una variable denominada `variable`, para la cual se quiere realizar una estimación por intervalo de confianza para su media.

El primer paso es guardar los valores observados de `variable` en un vector numérico, llamado `y`. Esto se hace mediante el código:

```
y <- data$variable
```

Posteriormente, la función `MeanCI` de la librería `DescTools` permite obtener el intervalo de confianza para la media poblacional de la siguiente forma:

```
MeanCI(y, sd = a, method = "classic")
```

En este caso, `y` es el vector numérico donde están los valores obtenidos en la muestra. En `sd` se indica el valor del desvío estándar poblacional conocido. Si no se indica ningún valor, R asume que el desvío estándar poblacional se desconoce y lo estima con el desvío estándar muestral.

En el argumento `method`, la opción `classic` permite construir los intervalos presentados en este texto, ya sea basados en la distribución normal o en la distribución t-Student.



Problema 1 (pág. 7): *En el ejemplo de las longitudes de las barras, a partir de la muestra de* $n = 120$ *barras seleccionada aleatoriamente de la producción, se obtiene que la longitud promedio muestral es* $\bar{y} = 249,994$ *mm.*

- *Suponga que se conoce que el desvío estándar poblacional, en este caso, vale* $\sigma = 0,2$ *mm.*
  Con los datos de las 120 longitudes, contenidos en un vector denominado `longitud`*, se construye el intervalo de confianza mediante la función:*
  `MeanCI(longitud, sd = 0.2, method = "classic", conf.level = 0.99)`
  Los resultados obtenidos son:

mean	*lower.ci*	*upper.ci*
249,9943	*249,9473*	*250,0414*

- *Suponga que no se conoce el valor del desvío estándar poblacional*
  Con los mismos datos mencionados, se construye el intervalo de confianza mediante la función:
  `MeanCI(longitud, method = "classic", conf.level = 0.99)`
  Los resultados obtenidos son:

mean	*lower.ci*	*upper.ci*
249,9943	*249,9224*	*250,0662*

Como se observa, con la información de la muestra de $n = 120$ *barras, se construyeron dos intervalos bilaterales del* 99 % *de confianza para la longitud promedio poblacional. El primero se construyó suponiendo que se conocía que el desvío estándar poblacional era igual a* $0,2$ *mm; mientras que el segundo se construyó sin conocer el valor de dicho parámetro.*

2. Intervalo de confianza para la variancia poblacional

La función `VarCI` de la librería `DescTools` permite obtener el intervalo de confianza para la variancia poblacional de la siguiente forma:

```
VarCI(y, method = c("classic"))
```

Nuevamente, `y` es el vector numérico donde están los valores obtenidos en la muestra. En el argumento `method`, la opción `classic` permite construir el intervalo de confianza basado en la distribución χ^2.

Este intervalo es válido cuando la variable de interés se distribuye normalmente. Si en `method` se indica `bonett`, se construye el intervalo basado en la propuesta de Bonett, válido para poblaciones asociadas a variables continuas, no necesariamente distribuidas normalmente.



Problema 1 (pág. 7): *En el ejemplo de las longitudes de las barras, a partir de la muestra de $n = 120$ barras seleccionada aleatoriamente de la producción, se obtiene que el desvío estándar muestral es $s = 0,3$ mm.*
Con los datos de las 120 longitudes, contenidos en un vector denominado `longitud`*, se construye el intervalo de confianza mediante la función:*

```
VarCI(longitud, method = "classic")
```

Los resultados obtenidos son:

var	*lower.ci*	*upper.ci*
0,09051216	*0,07129092*	*0,11875409*

También se construye otro intervalo mediante la función:

```
VarCI(longitud, method = "classic", sides = "right")
```

Los resultados obtenidos son:

var	*lower.ci*	*upper.ci*
0,09051216	*0,00000000*	*0,11360412*

Como se observa, con la información de la muestra de $n = 120$ barras, se construyeron dos intervalos del 95 % de confianza para la variancia poblacional. El primero bilateral y el segundo unilateral. En ambos casos, R utilizó la distribución χ^2 ya que la variable de interés se puede considerar normalmente distribuida.

3. Intervalo de confianza para la proporción poblacional

La función `BinomCI` de la librería `DescTools` permite obtener el intervalo de confianza para la proporción poblacional de la siguiente forma:

```
BinomCI(x = 10 , n = 60, method = "clopper-pearson")
```

En `x` se indica la cantidad de "éxitos" observada en la muestra y en `n` se indica el tamaño de la muestra. En el ejemplo, anterior, se trata de una muestra de tamaño 60, en la que el "éxito" ocurrió 10 veces. En el argumento `method`, la opción `clopper-pearson` permite construir el intervalo de confianza exacto. La opción `wald` da lugar al intervalo de confianza basado en la aproximación normal.



Problema 1 (pág. 7): *En el ejemplo de las longitudes de las barras, a partir de la muestra de $n = 120$ barras seleccionada aleatoriamente de la producción, se obtiene que la cantidad de barras con*

longitud fuera de especificaciones (longitud < 249,4 mm o longitud > 250,6 mm) es 6.

```
sum(longitud< 249.4|longitud > 250.6)
[1] 6
```

Con esa información se puede estimar la proporción de barras fuera de especificaciones a través de un intervalo de confianza.

```
BinomCI(x = 6 , n = 120, conf.level = 0.9, method = "clopper-pearson")
```

Los resultados obtenidos son:

est	*lower.ci*	*upper.ci*
0,05	*0,02199613*	*0,09629632*

También se construye otro intervalo mediante la función:

```
BinomCI(x = 6 , n = 120, conf.level =0.9, method = "wald")
```

Los resultados obtenidos son:

est	*lower.ci*	*upper.ci*
0,05	*0,01727471*	*0,08272529*

Como se observa, con la información de la muestra de $n = 120$ barras, se construyeron dos intervalos del 90 % de confianza para la proporción de barras de la población que no cumplen con las especificaciones definidas por la empresa automotriz. El primero exacto y el segundo aproximado.

8.8.3 Intervalos de tolerancia

Para obtener los intervalos de tolerancia, se debe instalar la librería `tolerance`. La función `normtol.int` es la que permite su obtención para el caso que la muestra provenga de una población normalmente distribuida, de la siguiente manera:

```
normtol.int(y, alpha = 0.05, P = 0.99, side = 2)
```

Aquí también `y` es el vector numérico donde están los valores obtenidos en la muestra. En el argumento `alpha` se indica indirectamente el nivel de confianza. Por ejemplo, si se desea un nivel de confianza del 99 %, se indica `alpha=0.01`.

`P` es la mínima proporción de unidades de la población que se pretenden cubrir con el intervalo a obtener, y `side` indica si interesa un intervalo de tolerancia bilateral (`side=2`) o unilateral (`side=1`). En este último caso, la función da como resultado los dos intervalos de tolerancia unilaterales posibles.

En el ejemplo, se calcula el intervalo de tolerancia bilateral que, con una confianza del 95 %, abarca al menos el 99 % de las unidades de la población.

La librería `tolerance` también permite presentar los intervalos de tolerancia en un gráfico de series temporales y en un histograma, de la siguiente forma:

```
out <- normtol.int(y, 0.05, 0.99, side = 2)
plottol(out, y, plot.type = "both")
```

Una vez creado el intervalo de tolerancia, se pueden graficar los datos de la muestra junto con los extremos del intervalo.

Problema 1 (pág. 7): *Con los datos de las 120 longitudes, contenidos en un vector denominado* `longitud`*, se construye el intervalo de tolerancia mediante la función:*

```
normtol.int(longitud, 0.05, 0.99, side=2)
```

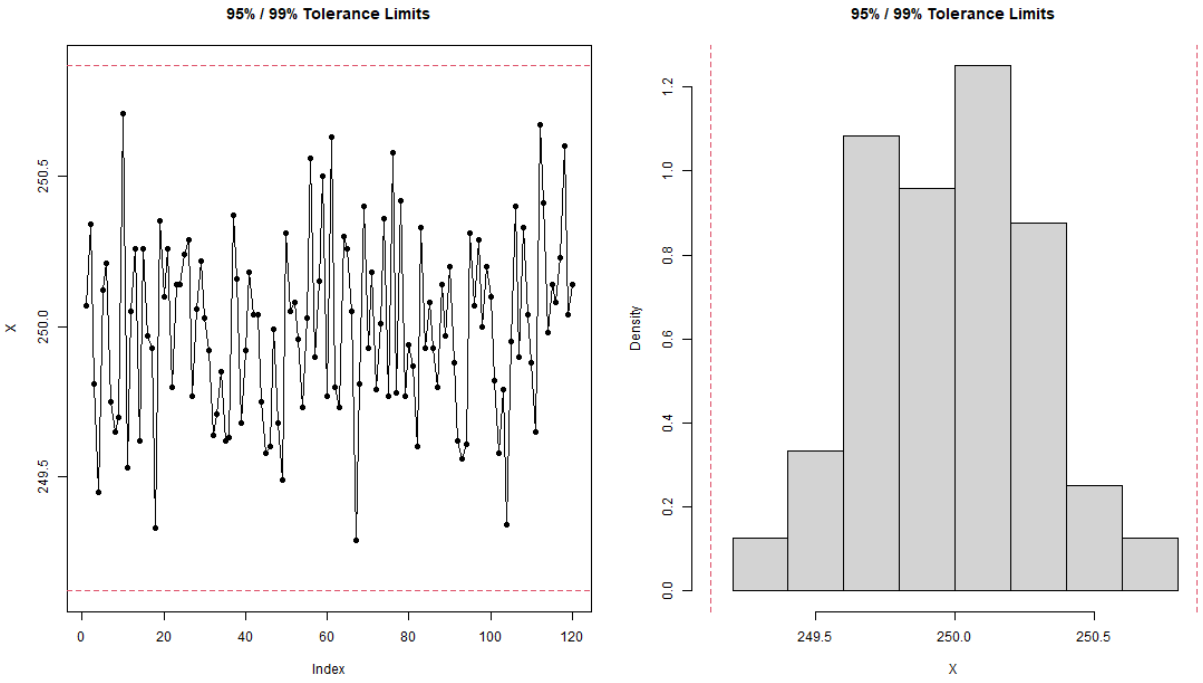
En este caso, se construye un intervalo de tolerancia bilateral con una cobertura del 99 % (como mínimo) y con una confianza del 95 %. Los resultados obtenidos son:

alpha	*P*	*x.bar*	*2-sided.lower*	*2-sided.upper*
0.05	*0.99*	*249.9942*	*249.1218*	*250.8667*

Construido el intervalo de tolerancia, se lo representa gráficamente mediante:

```
out <- normtol.int(longitud, 0.05, 0.99, side = 2)
plottol(out, longitud, plot.type = "both")
```

El resultado obtenido es:



Observe que en el primer gráfico se presentan las longitudes en el orden en que están en el archivo y con líneas punteadas se señalan los extremos del intervalo de tolerancia. Este gráfico tiene sentido si el orden en el que están en el archivo es el orden en que se obtuvieron las unidades o en que se realizaron las mediciones. En el segundo gráfico se observa el histograma y los extremos del intervalo de tolerancia señalados a ambos lados del mismo.



9. Inferencia estadística. Pruebas de Hipótesis

9.1 Introducción

En los Capítulos 7 y 8 se abordaron conceptos relativos al análisis inferencial. En el Capítulo 7, se presentó el estudio de las distribuciones muestrales de algunos estadísticos y a partir de estos conceptos, en el Capítulo 8 se introdujo el análisis inferencial mediante el uso de intervalos de confianza.

Tal como se menciona en la Sección 8.2, existen principalmente dos enfoques para realizar inferencias sobre parámetros poblacionales: por un lado, los intervalos de confianza, desarrollados en ese capítulo; y por otro, las pruebas de hipótesis, que se utilizan cuando se desea contrastar una afirmación formulada sobre el valor de un parámetro de interés.

Las pruebas de hipótesis permiten, además, probar hipótesis relacionadas con el modelo o la distribución que sigue una variable, la existencia de relación entre dos variables, etc. Estas últimas se denominan pruebas no paramétricas.

Aunque el procedimiento para realizar pruebas de hipótesis difiere del utilizado en la construcción de intervalos de confianza, ambos métodos se basan en los mismos principios fundamentales. Tanto en pruebas paramétricas como no paramétricas, es esencial definir un estadístico y conocer su distribución muestral para establecer márgenes de error o criterios de decisión basados en los datos observados.

Cuando la prueba se refiera a alguno de los parámetros tratados en el Capítulo 8, se retomarán los conceptos introducidos en el Capítulo 7 y se podrá concluir utilizando el procedimiento propio de las pruebas de hipótesis o a través de un intervalo de confianza.

En este capítulo se presentan los principales conceptos relacionados con las pruebas de hipótesis. Se comienza con pruebas paramétricas aplicadas a parámetros como la media, la desviación estándar y la proporción, para luego introducir pruebas no paramétricas orientadas a evaluar el modelo de distribución de una variable. Se describen los elementos clave de los distintos procedimientos, así como los riesgos asociados a las decisiones que se derivan de las pruebas.

Los objetivos de este capítulo son:

- Presentar las pruebas de hipótesis y las diferentes formas de llevarlas a cabo.
- Aplicar las pruebas de hipótesis junto a los intervalos de confianza y herramientas descriptivas en la resolución de problemas.

9.2 Pruebas de hipótesis

Además de los intervalos de confianza, las pruebas (o tests) de hipótesis constituyen otro procedimiento de inferencia estadística ampliamente utilizado en el análisis inferencial. En ellas se valora la evidencia muestral para probar alguna hipótesis que se enuncia acerca de una población.

En un problema de naturaleza estadística, al tener definida la población y las variable/s, pueden surgir hipótesis sobre algun parámetro de interés o sobre alguna característica referida a dichas variables.



Problema 1 (pág. 7): *Considere la población de barras producidas por la empresa metalúrgica, de las cuales interesa su longitud, en mm. Pueden surgir, por ejemplo, las siguientes hipótesis:*

- *La longitud promedio de las barras es 250 mm ($\mu = 250$).*
- *La proporción de barras que no cumplen con las especificaciones definidas por la automotriz es como mínimo 0,01 ($\pi \geq 0,01$).*
- *La variable de interés (longitud, Y) sigue una ley normal ($Y \sim N$).*

Una hipótesis estadística es una proposición sobre una o más poblaciones (variables).

Si se pudiera contar con información de toda la población se decidiría (sin riesgos inferenciales) si la hipótesis es correcta o no. Pero, en muchos casos, solo se cuenta con una muestra aleatoria de la población.

Así como cuando se trabaja con intervalos de confianza hay un margen de error y un riesgo de obtener conclusiones erróneas, en las pruebas de hipótesis también puede concluirse erróneamente debido a la aleatoriedad subyacente al trabajar con muestras. Por ejemplo, la información que brindan los datos de la muestra puede llevar a concluir que una hipótesis planteada es falsa cuando en realidad es cierta.

Según a qué característica de la población o poblaciones se refieran las hipótesis, las pruebas pueden clasificarse en:

- **Pruebas paramétricas**: enuncian algún o algunos valores para el o los parámetros de interés. Por ejemplo la referida a la longitud promedio de las barras o a la proporción de barras que no cumplen con las especificaciones. También puede plantearse que los promedios de dos poblaciones son iguales ($\mu_A = \mu_B$).
- **Pruebas no paramétricas**: las hipótesis hacen referencia a otras características de las variables en estudio. Entre ellas se encuentran las pruebas de bondad de ajuste, que concluyen sobre el modelo que sigue la distribución de una variable. Por ejemplo: la variable en estudio se comporta según el modelo Normal ($Y \sim N$) o la variable en estudio se comporta según el modelo Exponencial ($Y \sim Exp$). También se puede hipotetizar, entre muchas otras cuestiones, sobre la independencia de dos variables categóricas. En este capítulo se presentan brevemente las pruebas de bondad de ajuste.

En toda prueba de hipótesis se postulan siempre dos: **la hipótesis nula (H_o) y la hipótesis alternativa (H_a)**.

La hipótesis nula refiere siempre a un escenario único, donde si se trata del valor de un parámetro, aparece el signo "=". Se suele decir que corresponde al "no cambio". En contraposición, la hipótesis alternativa refiere generalmente a infinitos escenarios, donde aparecen otros signos ("$\neq$", "$<$" o "$>$") (si la prueba es paramétrica) y se dice que corresponde al "cambio".



Problema 1 (pág. 7): *Considere nuevamente el problema de la empresa productora de barras. En ese caso, uno de los parámetros de interés es la longitud promedio (μ) de las mismas. Suponga que hubo un cambio en los turnos del personal y se quiere analizar, a partir de una muestra, si después de esta modificación la longitud promedio de las barras que se producen cambió. Ante esto surgen dos posibilidades:*

- *La longitud media de las barras sigue siendo la misma.*
- *La longitud media de las barras cambió.*

Estas dos proposiciones o afirmaciones son las hipótesis en consideración en este caso. La hipótesis que refleja la sospecha de los ingenieros (la longitud media de las barras cambió), es la hipótesis alternativa. Lo contrario se indica como hipótesis nula, donde se postula que μ es igual a un determinado valor (μ_o). En este caso, $\mu_o = 250$.

En símbolos: $H_o : \mu = 250 \quad H_a : \mu \neq 250$.

En toda prueba de hipótesis se analiza si los datos de la muestra presentan evidencia en contra de la hipótesis nula, para así dar por probada la hipótesis alternativa.

Para resolver la prueba de hipótesis, hay caminos alternativos:

- Se puede utilizar un intervalo de confianza (como se hizo en algunos casos en el Capítulo 8).
- Se puede definir un procedimiento basado en una región de rechazo de H_o (forma clásica).
- Se puede utilizar el valor p.

Los tres son equivalentes y conducen a la misma conclusión, pero los intervalos de confianza solo pueden utilizarse cuando se trata de pruebas paramétricas.

Dado que se trabaja con información muestral para decidir sobre la población, existe la posibilidad de que la decisión no sea la acertada. Lo que se analiza es si los datos muestrales dan evidencia en contra de H_o o no, es decir, **la decisión que debe tomarse es si se rechaza o no a la hipótesis nula**. Se debe tener en cuenta que cuando se rechace H_o, será porque se encontró la evidencia suficiente para hacerlo y cuando no se rechace, será porque la evidencia en la muestra no fue suficiente.

Si la hipótesis nula es cierta, la evidencia muestral puede conducir a rechazarla (lo cual sería erróneo) o a no rechazarla (decisión acertada). Si la hipótesis nula es falsa, la evidencia muestral puede conducir a rechazarla (lo cual sería acertado en este caso) o a no rechazarla (decisión errónea). En la Tabla 1.3 se presentan las situaciones posibles y se clasifica a las decisiones como correctas o no, según corresponda.

Tabla 9.1 *Situaciones posibles al poner a prueba una hipótesis nula*

Decisión que se toma a partir de información muestral	**Situación real en la población**	
	H_o ***es cierta***	H_o ***es falsa, la cierta es*** H_a
Se rechaza H_o	Se comete error de tipo I, e_I	La decisión es correcta
No se rechaza H_o	La decisión es correcta	Se comete error de tipo II, e_{II}

La probabilidad de que se cometa error de tipo I es la probabilidad de rechazar una hipótesis nula cuando esta es cierta $\rightarrow P(e_I) = P(\text{rechazar } H_o | H_o \text{ cierta})$.

El máximo valor para esta probabilidad es α, que se denomina **"nivel de significación de la prueba"**.

La probabilidad de que se cometa error de tipo II es la probabilidad de no rechazar una hipótesis nula cuando esta es falsa $\rightarrow P(e_{II}) = P(\text{no rechazar } H_o | H_o \text{ falsa})$.

El valor de esta probabilidad es β. La probabilidad de rechazar H_o acertadamente, es decir, de rechazar H_o cuando esta es falsa, se denomina **"potencia de la prueba"** y es igual a $(1-\beta)$.

Cuando los datos de la muestra conducen al rechazo de la hipótesis nula, se dice que se tiene un **resultado estadísticamente significativo, al nivel** α. En ese caso existe la posibilidad de cometer error de tipo I, y α es el máximo riesgo de que ese error ocurra.

Cuando los datos de la muestra no conducen al rechazo de H_o, esto solo significa que no hay suficiente evidencia en contra de la misma. Se dice que no se rechaza H_o (o se retiene H_o), pero no se puede afirmar que H_o sea cierta porque se pudo haber cometido un error de tipo II, que se comete al no rechazar H_o cuando esta es falsa.

Observe que los errores de tipo I y II no pueden ocurrir simultáneamente. El primero puede ocurrir en el caso de rechazar H_o y el segundo puede ocurrir solo si H_o no se rechaza.

9.3 Pruebas de hipótesis sobre parámetros

Cuando en el planteo del problema aparece como objetivo determinar el cumplimiento de alguna condición respecto de un parámetro, el problema puede resolverse a través de una prueba de hipótesis paramétrica.

La prueba de hipótesis paramétrica se presenta de manera más detallada para el caso de la media poblacional, μ, cuando se conoce la desviación estándar poblacional, σ.
Para esta prueba se desarrollan y analizan todos los conceptos involucrados: las formas de proceder para rechazar o no a la hipótesis nula, los riesgos asociados a las decisiones que se toman, la relación entre el tamaño de muestra y los riesgos, entre otros aspectos relevantes.
Los ejemplos utilizados para ilustrar estos conceptos, consideran una muestra de tamaño $n = 10$, extraída de una población de barras, cuyas longitudes se suponen distribuidas normalmente con desviación estándar $\sigma = 0,2$.
Cabe señalar que los conceptos abordados son aplicables no solo a esta situación particular, sino también a cualquier prueba de hipótesis paramétrica, y varios de ellos resultan válidos en el contexto general de las pruebas de hipótesis.

9.3.1 Pruebas de hipótesis para la media poblacional

1. Prueba de hipótesis para la media poblacional, μ, cuando el desvío estándar poblacional, σ, es conocido

a. Formas de plantear las hipótesis

Las hipótesis (nula y alternativa) paramétricas pueden plantearse en cualquiera de las siguientes formas:

- Prueba bilateral. Se plantea cuando se sospecha que ha habido un cambio en el valor del promedio poblacional y este cambio puede verse reflejado en un aumento o disminución del mismo.
  $H_o: \ \mu = \mu_o$
  $H_a: \ \mu \neq \mu_o$
- Prueba unilateral a la derecha. Se plantea cuando se sospecha que ha habido un aumento en el valor del promedio poblacional.
  $H_o: \ \mu \leq \mu_o$
  $H_a: \ \mu > \mu_o$
- Prueba unilateral a la izquierda. Se plantea cuando se sospecha que ha habido una disminución en el valor del promedio poblacional.
  $H_o: \ \mu \geq \mu_o$
  $H_a: \ \mu < \mu_o$

En el caso de pruebas unilaterales, la hipótesis nula también puede expresarse $H_o: \ \mu = \mu_o$, ya que en toda prueba de hipótesis se observa si en la muestra hay evidencia para rechazar H_o, en favor de H_a. Entonces, por ejemplo, en una prueba unilateral a la derecha, rechazar que $\mu = \mu_o$, y probar que $\mu > \mu_o$, es lo mismo que rechazar $\mu \leq \mu_o$.

Además, como se verá luego, el procedimiento probabilístico que se utiliza requiere plantear un escenario concreto (donde el parámetro toma un único valor, es decir, una hipótesis con el signo "=") y no infinitos escenarios (en los cuales el parámetro puede tomar infinitos valores, como en las hipótesis con el signo "$<$" o "$>$").

Es importante destacar que **en la hipótesis alternativa nunca podría postularse el signo "="**, ya que por el margen de error existente en la inferencia, no hay forma de probar que el parámetro es exactamente igual a un valor.

La forma de llevar a cabo una prueba de hipótesis estadística consiste en analizar si existe evidencia muestral en contra de la hipótesis nula y a favor de la hipótesis alternativa. Se "contradice" la hipótesis nula cuando los datos de la muestra son poco probables bajo la suposición de que la hipótesis nula es cierta.

b. El procedimiento basado en una región de rechazo (forma clásica)

Como se mencionó en párrafos anteriores, hay tres procedimientos para resolver una prueba de hipótesis. Uno de ellos, basado en la **utilización de intervalos de confianza**, se presentó en el Capítulo 8 pero sin escribir ni formalizar las hipótesis, como en el ejemplo que se muestra a continuación.

Problema 1 (pág. 7): *Siguiendo con el ejemplo de las longitudes de las barras, suponga que se extrae una muestra de tamaño $n = 10$ y que la media muestral, $\bar{y}$, resulta 250,2 mm.*
El intervalo de confianza del 95 % para la media poblacional es $IC_{95\%,\mu} = (250,076; 250,324)$.
Suponga, por ejemplo, que si la longitud promedio de las barras supera 250 mm, se considerará hacer ajustes en el proceso para que dicho promedio disminuya. A partir del intervalo de confianza del 95 % se observa que los posibles valores para μ son mayores que 250 mm. De ahí que pueda considerarse (corriendo un riesgo del 5 %) que es razonable pensar que $\mu > 250$ mm y, por lo tanto, se considerará hacer ajustes en el proceso [1].
Las hipótesis nula y alternativa correspondientes a esta situación son:
$H_o: \mu = 250 \quad H_a: \mu > 250$
Con las hipótesis así formuladas, esta conclusión se traduce en un rechazo de H_o en favor de H_a.

[1] Recuerde que la información estadística debe siempre analizarse junto con información del contexto del problema para finalmente tomar la decisión de hacer ajustes o no.

Tanto el procedimiento basado en obtener una región de rechazo de H_o como el basado en el valor p, se sustentan en los mismos conceptos que el intervalo de confianza, es decir, en todo lo visto y definido en el Capítulo 7. La diferencia es que se determina, antes de tomar la muestra y suponiendo cierta la hipótesis nula, un criterio para rechazar o no dicha suposición según los datos que se obtengan en la muestra.

En todos los procedimientos se requiere en primer lugar plantear las hipótesis nula y alternativa para el parámetro de interés y definir el nivel de significación (α) que es el máximo valor que se admite para la probabilidad de error de tipo I. Luego se debe definir el estadístico de prueba y su distribución de probabilidades correspondiente.

Lo que sigue es exclusivo del **procedimiento basado en una región de rechazo**, en el cual la idea es determinar un criterio para decidir si se rechaza o no a la hipótesis nula, a partir del valor del estadístico observado en la muestra.

En este procedimiento, se continúa suponiendo que el parámetro de interés toma el valor postulado en H_o. A la distribución del estadístico, construida a partir de esta suposición, se la llama **"distribución bajo H_o"**.

Problema 1 (pág. 7): *Volviendo al problema de las longitudes de las barras, suponga que hubo un cambio en los turnos del personal y se quiere analizar, a partir de una muestra, si después de esta modificación la longitud promedio de las barras que se producen cambió.*

- *Se postulan las hipótesis: $H_o: \mu = 250 \quad H_a: \mu \neq 250$.*
- *Se define el nivel de significación de la prueba, $\alpha = 0,05$.*
- *El estadístico es la media muestral, $\bar{Y}$, que se comporta según el modelo Normal: $\bar{Y} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, donde la desviación estándar se conoce y es igual a 0,2, y el tamaño de la muestra es 10.*
  Suponiendo cierto el valor postulado para μ en H_o, la distribución queda:
  $\bar{Y} \sim N(250; \frac{0,2}{\sqrt{10}})$
  $\bar{Y} \sim N(250; 0,0632)$.

Resta definir la región de rechazo de H_o y tomar la decisión.

A partir de la distribución del estadístico "media muestral" bajo H_o, se obtiene una región de posibles valores de la media muestral que son poco probables si es cierta la hipótesis nula y que podrían indicar que dicha hipótesis debería rechazarse.

Claramente, los valores que verifican esto son los que están más alejados de μ_o. La región, llamada **"región de rechazo"** estará formada, entonces, por dos intervalos, como se puede observar en la Figura 9.1.

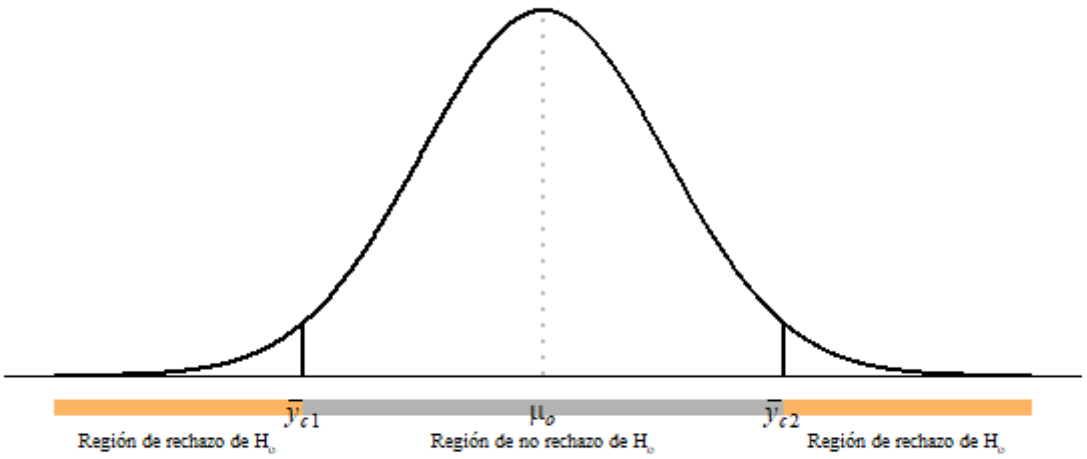


Figura 9.1 *Regiones de rechazo y no rechazo de H_o. Prueba bilateral para μ cuando σ es conocido. Distribución Normal*

El criterio o regla de decisión consiste en rechazar H_o si la media muestral toma un valor en la región de rechazo y no rechazarla en caso contrario.

Los valores que delimitan esta zona se denominan **valores críticos** y se suelen simbolizar con el subíndice c (o c_1 y c_2 si se trata de una prueba bilateral); mientras que para el valor observado del estadístico en la muestra se suele escribir el subíndice *obs*.

Formalmente, la regla de decisión para una prueba bilateral se escribe de la siguiente manera:

Rechazar H_o en favor de H_a si $\bar{y}_{obs} < \bar{y}_{c_1}$ **o** $\bar{y}_{obs} > \bar{y}_{c_2}$.

Observe que definida una región de rechazo, existe una probabilidad α de que la media de la muestra tome un valor perteneciente a la misma, aun cuando H_o sea cierta. Por lo tanto, al realizar la prueba se está asumiendo la posibilidad de "rechazar mal" H_o (rechazarla cuando es cierta), es decir, de cometer **error de tipo I** (e_I).

Los valores $\bar{y}_{c_1}$ e $\bar{y}_{c_2}$ dependen del nivel de significación elegido. Concretamente, el área sombreada en la Figura 9.1, asociada a la región de rechazo, debe ser igual a α. Los valores críticos mencionados son, respectivamente el percentil de orden $(\frac{\alpha}{2})$ y el percentil de orden $(1-\frac{\alpha}{2})$ de la distribución del estadístico bajo H_o, como se observa en la Figura 9.2.

Observe que, al fijar el valor de α, se determina la probabilidad de que el valor del estadístico que se observe en la muestra pertenezca a la región de rechazo, aun cuando H_o sea cierta.

Formalmente, $P(e_I) = P(\text{Rechazar } H_o | H_o \text{ es cierta}) = P((\bar{y}_{c_1} < \bar{Y} \text{ o } \bar{Y} > \bar{y}_{c_2}) | \mu = \mu_o)$.

Es decir, es la probabilidad de que la media muestral tome un valor en la región de rechazo, suponiendo que la media poblacional, μ, toma el valor μ_o. Esto se visualiza en la Figura 9.2.

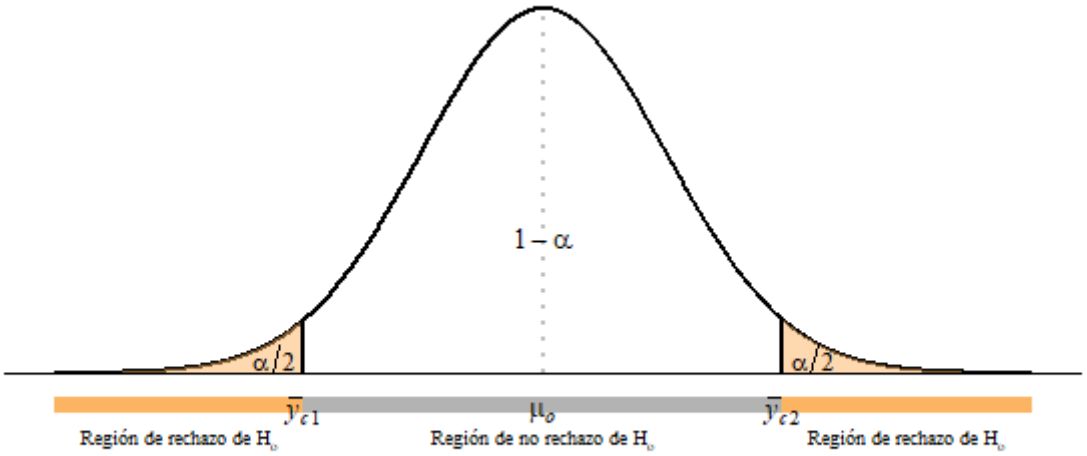


Figura 9.2 *Regiones de rechazo y no rechazo de H_o. Prueba bilateral para μ cuando σ es conocido. Distribución Normal. El área de la región en naranja representa la probabilidad de cometer error de tipo I (α)*

Problema 1 (pág. 7): *Suponga las siguientes hipótesis $H_o: \mu = 250$ $H_a: \mu \neq 250$.*
Si H_o es cierta, $\bar{Y} \sim N(250; 0,06325)$.
Si el nivel de significación se fija en $\alpha = 0,05$, los valores críticos son, respectivamente, los percentiles de orden 0,025 y 0,975 de la distribución del estadístico bajo H_o. La región de rechazo de la hipótesis nula es la formada por los valores menores que 249,876 mm o mayores que 250,124 mm (Figura 9.3). Se verifica que:
$\alpha = P(e_I) = P(\text{Rechazar } H_o | H_o \text{ es cierta}) = P((\bar{Y} < 249,876 \text{ o } \bar{Y} > 250,124) | \mu = 250) = 0,05$.

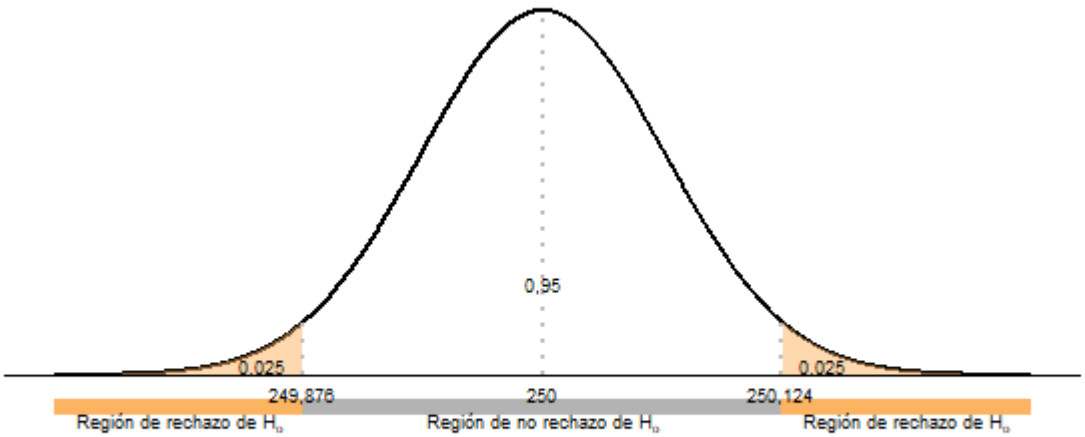


Figura 9.3 *Regiones de rechazo y no rechazo de H_o. Prueba $H_o: \mu = 250$ vs $H_a: \mu \neq 250$ cuando $\sigma = 0,2$. Distribución Normal, $\alpha = 0,05$*

¿Cómo se modificaría la región de rechazo si el nivel de significación α fuera 0,01?, ¿y si fuera 0,1?

1. *En el primer caso, los valores críticos son los percentiles de orden 0,005 y 0,995, los cuales valen, respectivamente,* $\bar{y}_{c_1} = 249,837$ *y* $\bar{y}_{c_2} = 250,163$. *Se verifica que:*
   $\alpha = P(e_I) = P(Rechazar\ H_o | H_o\ es\ cierta) = P((\bar{Y} < 249,837\ o\ \bar{Y} > 250,163) | \mu = 250) = 0,01.$
2. *En el segundo caso, los valores críticos son los percentiles de orden 0,05 y 0,95, los cuales valen, respectivamente,* $\bar{y}_{c_1} = 249,896$ *y* $\bar{y}_{c_2} = 250,104$. *Se verifica que:*
   $\alpha = P(e_I) = P(Rechazar\ H_o | H_o\ es\ cierta) = P((\bar{Y} < 249,896\ o\ \bar{Y} > 250,104) | \mu = 250) = 0,1.$
   Ambas regiones se presentan en la Figura 9.4.

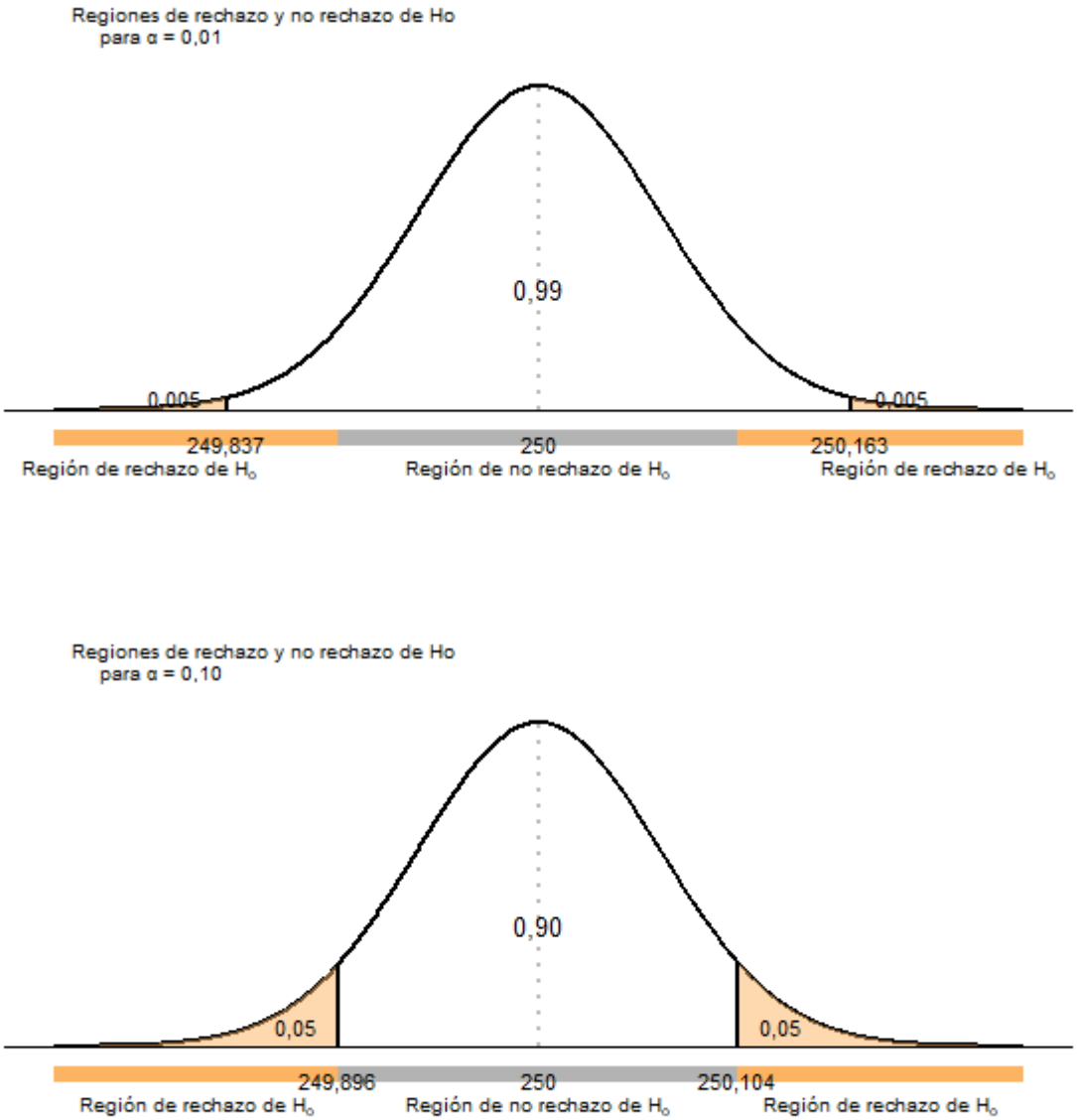


Figura 9.4 *Regiones de rechazo y no rechazo de* H_o. *Prueba* $H_o: \mu = 250$ *vs* $H_a: \mu \neq 250$ *cuando* $\sigma = 0,2$. *Distribución Normal, para distintos valores de* α

En general, en el proceso de resolución de un problema que involucre pruebas de hipótesis, se determina de antemano la máxima probabilidad tolerable para el error de tipo I (α), en función del contexto y de los costos asociados. Luego quedan determinadas las regiones de rechazo y no rechazo de H_o correspondientes.

Problema 1 (pág. 7): *El procedimiento basado en la región de rechazo seguido en el caso de las hipótesis bilaterales respecto de la longitud promedio poblacional, se resume a continuación:*

- *Planteo de las hipótesis nula y alternativa:* $H_o: \mu = 250 \quad H_a: \mu \neq 250$.
- *Definición del estadístico de la prueba y su distribución bajo* H_o*:* $\bar{Y} \sim N(250; \frac{0,2}{\sqrt{10}})$.
- *Determinación del valor de* α *y definición de las regiones de rechazo y no rechazo de* H_o *(y de la regla de decisión correspondiente):*
  $\alpha = 0,05$
  Región de rechazo: $(-\infty; 249,876) U (250,124; \infty)$
  Regla de decisión: Rechazar H_o *en favor de* H_a *si* $\bar{y}_{obs} < 249,876$ *o* $\bar{y}_{obs} > 250,124$.
- *Selección de la muestra, análisis de los datos, cálculo del valor observado del estadístico y conclusión según la regla de decisión establecida: La decisión dependerá en este caso del valor de* $\bar{y}_{obs}$.

En relación a este ejemplo, según el valor que arroje la media muestral observada, se pueden tomar distintas decisiones:

- *Suponga que en la muestra el valor observado de la media muestral resulta* $\bar{y}_{obs} = 250,15$. *Este valor es mayor que 250,124, es decir, pertenece a la región de rechazo y, por lo tanto, la decisión es rechazar* H_o *en favor de* H_a.
  De cometerse un error, este sería de tipo I.
- *Suponga ahora que en la muestra el valor observado de la media muestra resulta* $\bar{y}_{obs} = 249,993$. *Este valor se encuentra en la región de no rechazo (*$249,876 < \bar{y}_{obs} < 250,124$*) y por lo tanto la decisión es no rechazar* H_o.
  De cometerse un error, este sería de tipo II.

¿Y si en la muestra el valor observado de la media muestral fuera $\bar{y}_{obs} = 249,8$*? ¿qué decisión se debería tomar?, ¿qué error se podría estar cometiendo? ¿Y si fuera* $\bar{y}_{obs} = 250,05$*?*

Observe que en caso de rechazar H_o, si bien pudo haberse cometido error de tipo I, el riesgo de que esto ocurriera, antes de tomar la muestra, estaba controlado, siendo su máximo valor α.

No rechazar H_o no significa aceptarla. Tenga en cuenta lo explicado en el Capítulo 8 (página 402) sobre la obtención de conclusiones a partir de un intervalo de confianza: si el intervalo contiene al valor hipótetico (valor planteado en H_o), no significa que el mismo sea el valor real del parámetro. Al trabajar con muestras no se puede afirmar que un parámetro es exactamente igual a un valor.

c. Pruebas de hipótesis unilaterales

Suponiendo que la sospecha que surge del contexto del problema al realizar el planteo es que el promedio aumentó, puede plantearse una prueba unilateral a la derecha (o a la izquierda si la sospecha fuera que el promedio disminuyó).

Para la prueba unilateral a la derecha, las hipótesis son las siguientes:
$H_o: \mu \leq \mu_o \quad H_a: \mu > \mu_o$ o, más simplemente, $H_o: \mu = \mu_o \quad H_a: \mu > \mu_o$

Al aplicar el procedimiento basado en la región de rechazo, esta no se define de la misma manera que para la prueba bilateral. En este caso los valores del estadístico que son poco probables bajo la hipótesis nula —y que, por lo tanto, conducen a su rechazo y permiten afirmar la hipótesis alternativa— se encuentran ubicados en el extremo derecho de la distribución, como se muestra en la imagen superior de la Figura 9.5. Se determina un único valor crítico ($\bar{y}_c$) y la regla de decisión, para un cierto nivel de significación α, consiste en rechazar H_o en favor de H_a si $\bar{y}_{obs} > \bar{y}_c$, donde $\bar{y}_c$ es el percentil de orden $(1-\alpha)$ de la distribución del estadístico bajo H_o.

Note que si se utilizara el valor crítico ($\bar{y}_{c_2}$) obtenido para la prueba bilateral, el nivel de significación de la prueba unilateral sería $\frac{\alpha}{2}$.

Para la prueba unilateral a la izquierda, las hipótesis son las siguientes:
$H_o: \mu \geq \mu_o \quad H_a: \mu < \mu_o$ o, más simplemente, $H_o: \mu = \mu_o \quad H_a: \mu < \mu_o$

En este caso los valores del estadístico que son poco probables bajo la hipótesis nula —y que, por lo tanto, conducen a su rechazo y permiten afirmar la hipótesis alternativa— se encuentran ubicados en el extremo izquierdo de la distribución, como se muestra en la imagen inferior de la Figura 9.5. Se determina un único valor crítico ($\bar{y}_c$) y la regla de decisión, para un cierto nivel de significación α, consiste en rechazar H_o en favor de H_a si $\bar{y}_{obs} < \bar{y}_c$, donde $\bar{y}_c$ es el percentil de orden α de la distribución del estadístico bajo H_o.

Note que si se utilizara el valor crítico ($\bar{y}_{c_1}$) obtenido para la prueba bilateral, el nivel de significación de la prueba unilateral sería $\frac{\alpha}{2}$.

Los pasos restantes para tomar una decisión sobre las hipótesis planteadas son iguales en los tres tipos de pruebas.

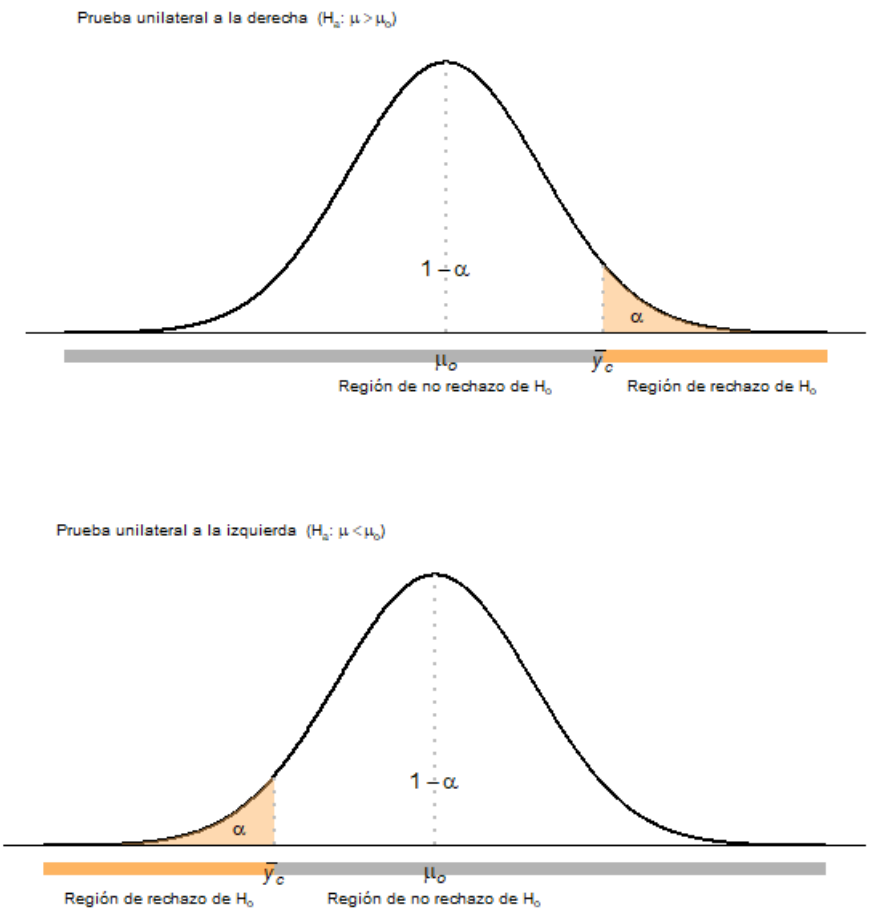


Figura 9.5 *Regiones de rechazo y no rechazo de H_o. Pruebas unilaterales para μ cuando σ es conocido. Distribución Normal. El área de la región en naranja representa el máximo valor de la probabilidad de cometer error de tipo I (α)*

Problema 1 (pág. 7): *En el problema de las barras, si se sospecha que la longitud promedio de las mismas aumentó, se plantean las siguientes hipótesis:*

$H_o: \mu = 250 \quad H_a: \mu > 250$

La distribución del estadístico bajo H_o es $\bar{Y} \sim N(250; \frac{0,2}{\sqrt{10}})$.

Para un nivel de significación $\alpha = 0,05$*, la regla de decisión consiste en rechazar H_o en favor de H_a si* $\bar{y}_{obs} > 250,104$ *(percentil de orden 0,95 de la distribución bajo H_o) (Figura 9.6, imagen superior).*

Si se hubiera sospechado una disminución en la longitud promedio de las barras, se plantearían las siguientes hipótesis:

$H_o: \mu = 250 \quad H_a: \mu < 250$

Para un nivel de significación $\alpha = 0,05$*, la regla de decisión consistiría en este caso, en rechazar H_o favor de H_a si* $\bar{y}_{obs} < 249,896$ *(percentil de orden 0,05 de la distribución bajo H_o) (Figura 9.6, imagen inferior).*

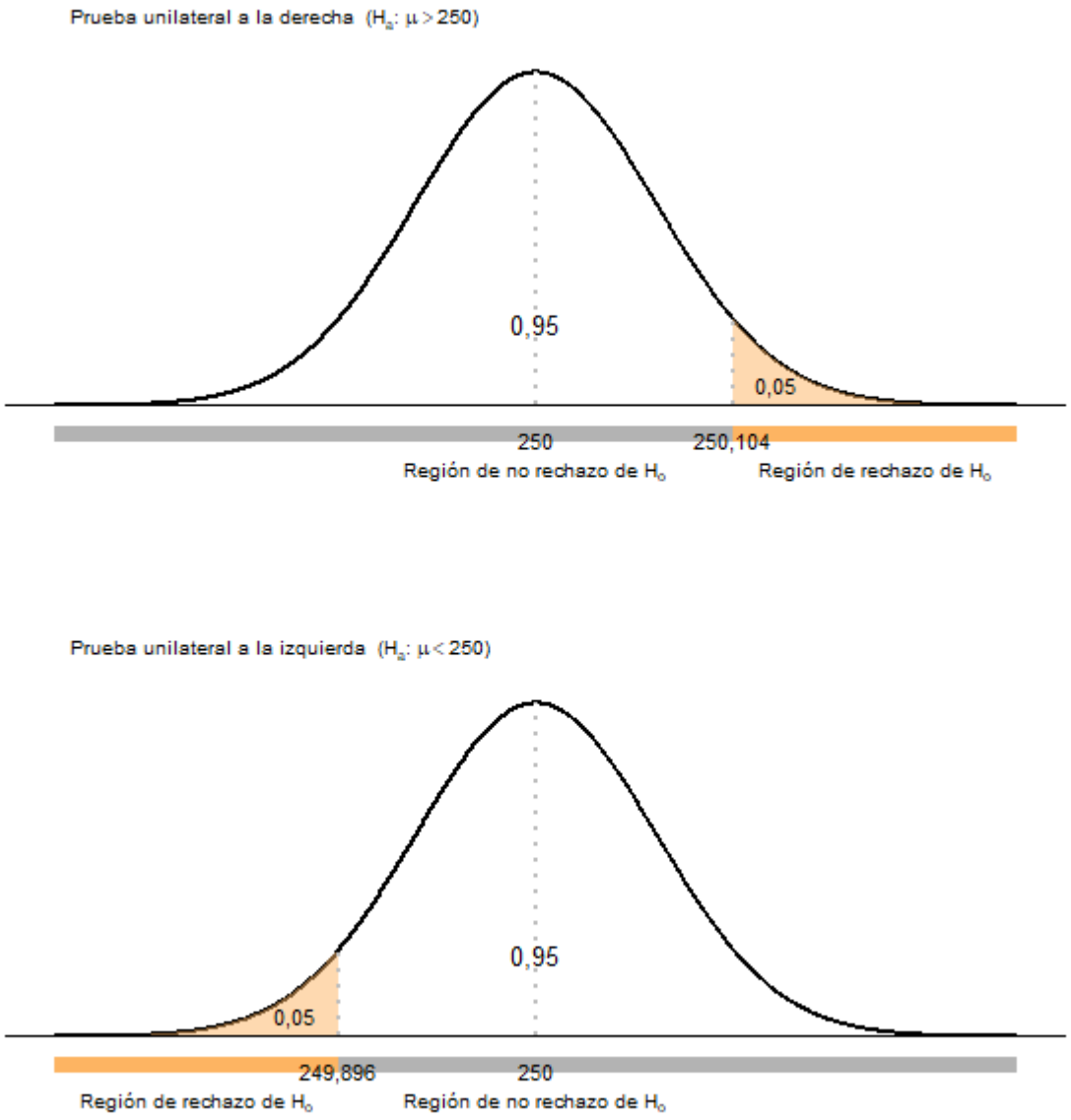


Figura 9.6 *Regiones de rechazo y no rechazo de H_o para las pruebas unilaterales planteadas sobre la longitud promedio de las barras, cuando* $\sigma = 0,2$*. Distribución Normal,* $\alpha = 0,05$

d. Estadísticos pivote

En muchos casos, la regla de decisión no se formula directamente sobre el estimador del parámetro de interés, sino sobre una función de dicho estimador.
En el caso estudiado, en el cual se conoce el valor de la desviación estándar poblacional, σ, se puede establecer la regla de decisión sobre el estadístico pivote $Z = \frac{\bar{Y}-\mu_o}{\frac{\sigma}{\sqrt{n}}}$, definido en el Capítulo 7.
Bajo H_o, $\bar{Y} \sim N(\mu_o; \frac{\sigma}{\sqrt{n}})$, y, por lo tanto, $Z \sim N(0, 1)$.

El valor observado del estadístico pivote se define de la siguiente manera: $z_{obs} = \frac{\bar{y}_{obs}-\mu_o}{\frac{\sigma}{\sqrt{n}}}$.

La regla de decisión, definida en función de dicho estadístico, consiste en rechazar H_o en favor de H_a si:

- $z_{obs} < z_{c_1}$ o $z_{obs} > z_{c_2}$ (prueba bilateral)
  donde z_{c_1} y z_{c_2} son los percentiles de orden $\frac{\alpha}{2}$ y $(1-\frac{\alpha}{2})$ de la distribución Normal estándar, respectivamente.
- $z_{obs} > z_c$ (prueba unilateral a la derecha)
  donde z_c es el percentil de orden $(1-\alpha)$ de la distribución Normal estándar.
- $z_{obs} < z_c$ (prueba unilateral a la izquierda)
  donde z_c es el percentil de orden α de la distribución Normal estándar.

Problema 1 (pág. 7): *En el problema de las barras, si se sospecha que la longitud promedio de las mismas cambió, se plantean las siguientes hipótesis:*
$H_o: \mu = 250 \quad H_a: \mu \neq 250$

Distribución del estadístico $\bar{Y}$ *bajo* H_o: $\bar{Y} \sim N(250; \frac{0,2}{\sqrt{10}})$.
Estadístico pivote: $Z = \frac{\bar{Y}-250}{\frac{0,2}{\sqrt{10}}}$.
Distribución del estadístico pivote bajo H_o: $Z \sim N(0, 1)$.
Para un nivel de significación $\alpha = 0,05$, *la regla de decisión consiste en rechazar* H_o *en favor de* H_a *si* $z_{obs} < z_{0,025} = -1,96$ *o* $z_{obs} > z_{0,975} = 1,96$ *(percentiles de orden 0,025 y 0,975 de la distribución Normal estándar, respectivamente).*

Si se hubiera sospechado un aumento en la longitud promedio de las barras, se plantearían las siguientes hipótesis: $H_o: \mu = 250 \quad H_a: \mu > 250$
Para un nivel de significación $\alpha = 0,05$, *la regla de decisión consistiría en este caso, en rechazar* H_o *favor de* H_a *si* $z_{obs} > z_{0,95} = 1,65$ *(percentil de orden 0,95 de la distribución Normal estándar).*

Si se hubiera sospechado una disminución en la longitud promedio de las barras, se plantearían las siguientes hipótesis: $H_o: \mu = 250 \quad H_a: \mu < 250$
Para un nivel de significación $\alpha = 0,05$, *la regla de decisión consiste en rechazar* H_o *en favor de* H_a *si* $z_{obs} < z_{0,05} = -1,65$ *(percentil de orden 0,05 de la distribución Normal estándar).*

e. El error de tipo II. La probabilidad β y la potencia de la prueba

Se comete **error de tipo II** si no se rechaza la hipótesis nula cuando esta es falsa. El riesgo de cometer este tipo de error, es decir, la probabilidad de error de tipo II se simboliza con β.

$(1-\beta)$ se denomina **potencia de la prueba**. Es la probabilidad de que se rechace H_o en favor de Ha, cuando H_a es verdadera.

Problema 1 (pág. 7): *En el problema de las barras, se plantean las siguientes hipótesis, sobre la longitud promedio de las barras, μ:*
$H_o: \mu = 250 \quad H_a: \mu \neq 250.$
En este caso,

- *el error de tipo I consiste en afirmar que la media del proceso ha cambiado, cuando en realidad sigue siendo 250 mm;*
- *el error de tipo II consiste en no rechazar que la media del proceso es 250 mm, cuando en realidad cambió.*

Cuando se planteó la probabilidad de error de tipo I (página 452), la condición de que H_o sea cierta, se expresa en un único escenario en el cual $\mu = \mu_o$. En cambio, al plantear la probabilidad de error de tipo II, la condición es que $\mu \neq \mu_o$, lo que implica infinitos escenarios posibles. No existe, entonces, un único valor para la probabilidad β, sino infinitos: para cada valor alternativo de μ (μ_a) se puede obtener un valor distinto de β (β_{μ_a}).

En la Figura 9.7 se presenta la probabilidad de error tipo II, β, para un posible valor de μ_a. Observe que dicha probabilidad se calcula para la distribución del estadístico suponiendo $\mu = \mu_a$ (distribución bajo H_a); mientras que la probabilidad de error de tipo I se obtiene para la distribución del estadístico suponiendo $\mu = \mu_o$ (distribución bajo H_o).

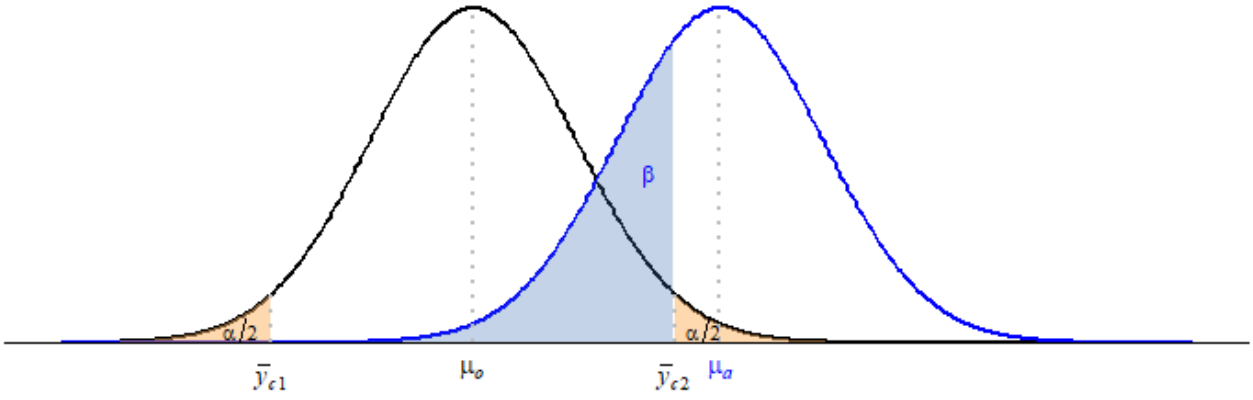


Figura 9.7 *Probabilidad de error de tipo II, β, para $\mu = \mu_a$ (área de la región en azul). Prueba de hipótesis bilateral para μ cuando σ es conocido. Distribución Normal. El área de la región en naranja indica la probabilidad de cometer error de tipo I (α)*

Problema 1 (pág. 7): *En el problema de las barras, suponga que un corrimiento de la longitud promedio de las mismas a 250,15 mm ($\mu_a = 250,15$) implica que el proceso puede tener inconvenientes debido al alto porcentaje de barras fuera de especificaciones que se producirían. Entonces tiene interés calcular la probabilidad de no rechazar H_o bajo ese supuesto, la cual resulta (manteniendo la regla de decisión asociada a $\alpha = 0,05$, página 454):*

$\beta_{\mu_a=250,15} = P(\text{no rechazar } H_o | \ H_o \text{ es falsa}) = P(249,876 < \bar{Y} < 250,124 | \mu_a = 250,15) = 0,34$

Esto quiere decir que, si la longitud promedio de las barras aumentara a 250,15 mm, existe una chance de 0,34 de no rechazar H_o en favor de H_a (en otras palabras, de no "detectar" dicho aumento con la información de la muestra).

La probabilidad de error de tipo II obtenida se representa gráficamente en la Figura 9.8.

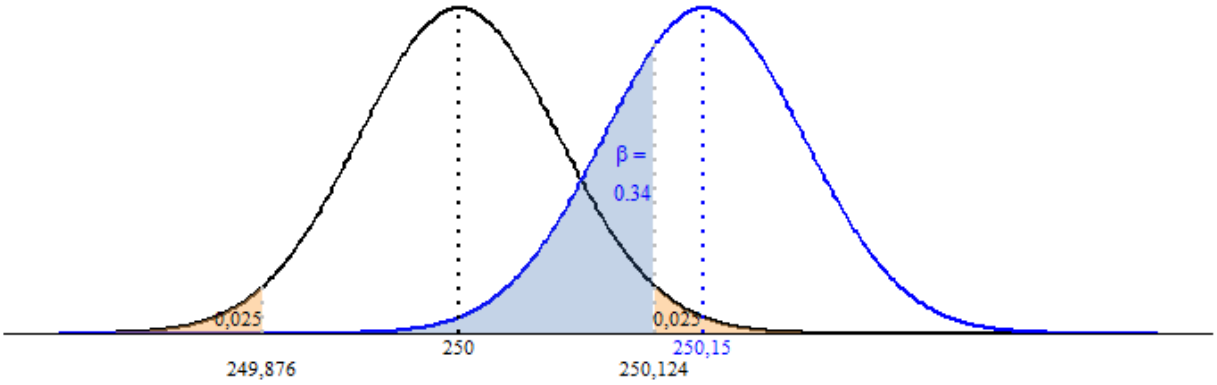


Figura 9.8 *Probabilidad de error de tipo II, β, para $\mu_a = 250,15$. Prueba $H_o: \mu = 250$ vs $H_a: \mu \neq 250$ cuando $\sigma = 0,2$. Distribución Normal, $\alpha = 0,05$*

En la Tabla 9.2, se muestran los valores de la probabilidad de error de tipo II (β), calculados para otros posibles valores de μ_a. En la segunda columna de la tabla, se realizan los mismos cálculos suponiendo que el tamaño de muestra fue de 50 barras. En todos los cálculos se consideró $\alpha = 0,05$.

Tabla 9.2 *Valores de β para diferentes valores de μ_a ($n = 10$ y $n = 50$, $\alpha = 0,05$)*

Valor de μ_a	**n = 10**	**n = 50**
$249,9$ *o* $250,1$	$0,6476$	$0,2334$
$249,85$ *o* $250,15$	$0,3405$	$0,0668$
$249,8$ *o* $250,2$	$0,1147$	$0,0111$
$249,7$ *o* $250,3$	$0,0027$	$0,0001$

Del análisis de los valores de β obtenidos para el caso de la longitud de las barras, se observa que, para un valor dado de μ_a y manteniendo el nivel de significación α, β disminuye si el tamaño de la muestra aumenta. Además, para un mismo valor de α y para un mismo tamaño de muestra, β decrece a medida que el valor alternativo del parámetro se aleja del valor postulado en la hipótesis nula.

¿Qué ocurre con la probabilidad β, si el nivel de significación se incrementa y el tamaño de muestra se mantiene fijo?

f. Relación entre tamaño de muestra, probabilidad de error de tipo I y probabilidad de error de tipo II. Determinación del tamaño de la muestra

En la Figura 9.9 se muestra la relación entre los conceptos asociados a los errores propios del muestreo en las pruebas de hipótesis.

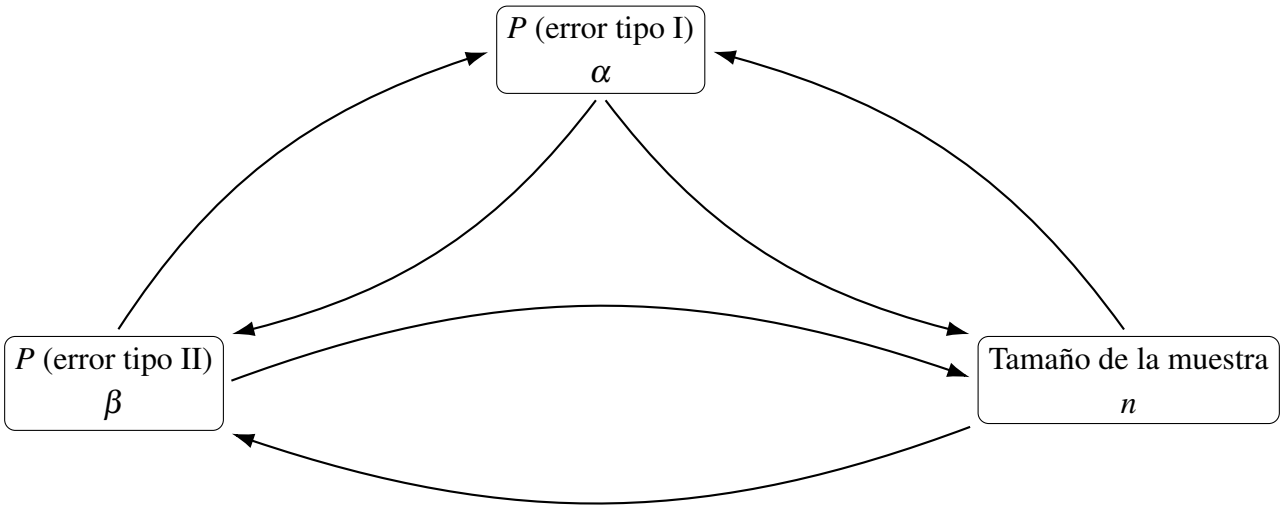


Figura 9.9 *Probabilidades de error de tipo I y tipo II y tamaño de muestra, relacionados entre sí*

¿Cómo se utiliza esta relación entre α, β y n en el momento de decidir el tamaño de muestra apropiado para una prueba de hipótesis?

En un problema concreto, donde se va a realizar una prueba de hipótesis para la media poblacional ($H_o: \mu = \mu_o$) y hay un valor alternativo del parámetro (μ_a) para el cual se quiere tener una baja probabilidad de error de tipo II (β), se pueden determinar los valores máximos pretendidos para α y β, y a partir de ellos obtener el mínimo tamaño de muestra requerido.

Fijados los máximos valores admitidos para α y β y definido un valor alternativo para el parámetro, μ_a, es posible despejar el mínimo valor de n. Si la prueba de hipótesis es bilateral:

$$n = \frac{(\left|z_{\frac{\alpha}{2}}\right| + |z_\beta|)^2 \sigma^2}{(\mu_a - \mu_o)^2},$$

donde $z_{\frac{\alpha}{2}}$ es el percentil de orden $\frac{\alpha}{2}$ de la distribución Normal estándar y z_β es el percentil de orden β de la misma distribución.
Para el caso de una prueba unilateral, en la expresión anterior se reemplaza $z_{\frac{\alpha}{2}}$ por z_α.



Problema 1 (pág. 7): *Como se mencionó en párrafos anteriores, un aumento de la longitud promedio de las barras a 250,15 mm ($\mu_a = 250,15$) implica que el proceso puede tener inconvenientes debido al alto porcentaje de barras fuera de especificaciones que se producirían. Suponga que se admite como máximo un riesgo de 0,10 de no detectar dicho aumento ($\beta = 0,10$).*

- *Se plantean las siguientes hipótesis: $H_o: \mu = 250$ $H_a: \mu \neq 250$.*
  Si se admite como máximo un riesgo de 0,05 de concluir erróneamente que la longitud promedio cambió ($\alpha = 0,05$), el mínimo tamaño de muestra requerido se obtiene de la siguiente manera:
  $n = \frac{(\left|z_{\frac{\alpha}{2}}\right| + |z_\beta|)^2 \sigma^2}{(\mu_a - \mu_o)^2} = \frac{(1,96+1,28)^2 \cdot 0,2^2}{(250,15-250)^2} = \frac{3,24^2 \cdot 0,2^2}{0,15^2} = 18,67$

  Es decir que se debe tomar como mínimo una muestra de n = 19 barras para que los riesgos sean los pretendidos.
- *Se plantean las siguientes hipótesis: $H_o: \mu = 250$ $H_a: \mu > 250$*
  Si se admite como máximo un riesgo de 0,05 de concluir erróneamente que la longitud promedio aumentó ($\alpha = 0,05$), el mínimo tamaño de muestra requerido se obtiene de la siguiente manera:
  $n = \frac{(|z_\alpha| + |z_\beta|)^2 \sigma^2}{(\mu_a - \mu_o)^2} = \frac{(1,65+1,28)^2 \cdot 0,2^2}{(250,15-250)^2} = \frac{2,93^2 \cdot 0,2^2}{0,15^2} = 13,74$
  En este caso, el mínimo tamaño de muestra requerido para que los riesgos sean los pretendidos es de 14 barras.

g. El valor p

Otra forma de decidir sobre rechazar o no la hipótesis nula, en lugar de comparar el valor de $\bar{y}_{obs}$ con el valor de $\bar{y}_c$ (o con los dos valores críticos para una prueba bilateral), consiste en comparar las probabilidades asociadas a los mismos. La probabilidad asociada a $\bar{y}_c$ es α y la probabilidad asociada a $\bar{y}_{obs}$ se llama valor p. Para que tenga sentido la comparación, el valor p debe calcularse de igual forma que la probabilidad α, es decir, debe ser la probabilidad de obtener en la muestra un valor del estadístico como el observado o más extremo, en la dirección de la hipótesis alternativa (siempre suponiendo cierta H_o).
Por ejemplo, para una hipótesis unilateral a la derecha sobre la media poblacional, el valor p se obtiene de la siguiente manera: valor $p = P(\bar{Y} > \bar{y}_{obs} | H_o \text{ cierta})$.

Cuando la prueba de hipótesis es bilateral, la región de rechazo está compuesta por dos subregiones, una en cada extremo de la distribución bajo H_o. Entonces, para poder comparar el valor p con α, el

mismo debe obtenerse como la suma de dos probabilidades iguales, correspondientes a dos regiones definidas a ambos lados de μ_o. Para definir estas regiones debe observarse si $\bar{y}_{obs}$ resultó menor o mayor que μ_o. Luego, se obtiene la probabilidad hacia el extremo correspondiente y se la multiplica por 2:

- si $\bar{y}_{obs} < \mu_o$, valor $p = 2.P(\bar{Y} < \bar{y}_{obs}|H_o \text{ cierta})$,
- si $\bar{y}_{obs} > \mu_o$, valor $p = 2.P(\bar{Y} > \bar{y}_{obs}|H_o \text{ cierta})$.

En la Figura 9.10 se observa el cálculo del valor p para pruebas unilaterales y bilaterales. En este último caso, se supone que el valor observado del estadístico resultó mayor que μ_o.

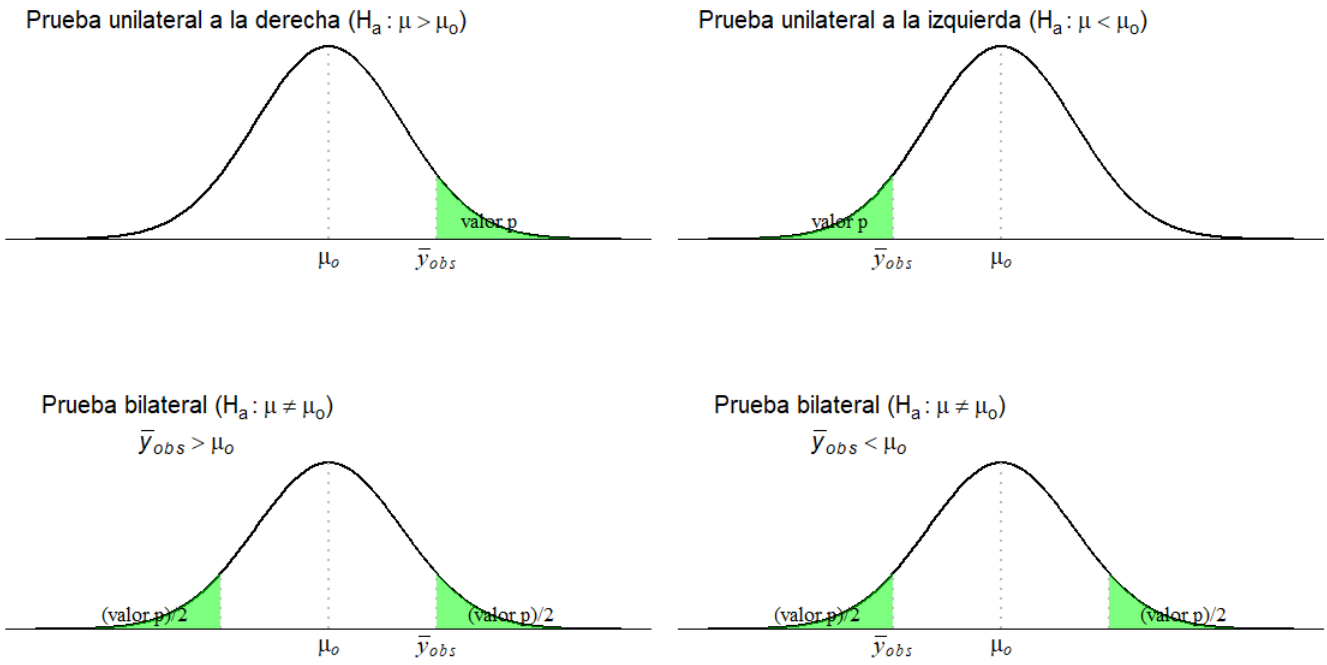


Figura 9.10 *Valor p (área de la región en verde). Pruebas uni y bilaterales para μ cuando σ es conocido. Distribución Normal*

Problema 1 (pág. 7): *En el problema de las longitudes de las barras, suponga que:*

- *se plantearon hipótesis bilaterales sobre la longitud promedio ($H_o: \mu = 250 \quad H_a: \mu \neq 250$),*
- *se fijó el nivel de significación de la prueba en 0,05,*
- *el valor observado para la media muestral fue $\bar{y}_{obs} = 250,12$ mm ($> \mu_o$).*

El valor p se obtiene de la siguiente manera: valor $p = 2.P(\bar{Y} > \bar{y}_{obs}|H_o) = 2.P(\bar{Y} > 250,12|H_o \text{ cierta}) = 0,018$.

En este caso, como valor $p < \alpha$, se rechaza H_o en favor de H_a.

Observe que para obtener el valor p para este ejemplo, se comenzó calculando una probabilidad "a la derecha", ya que el valor observado (250,12 mm) resultó mayor que el postulado en la hipótesis nula (250 mm). Luego se multiplicó a la probabilidad obtenida por 2, ya que la prueba era bilateral.

En la Figura 9.11 se presenta el valor p para esta prueba.

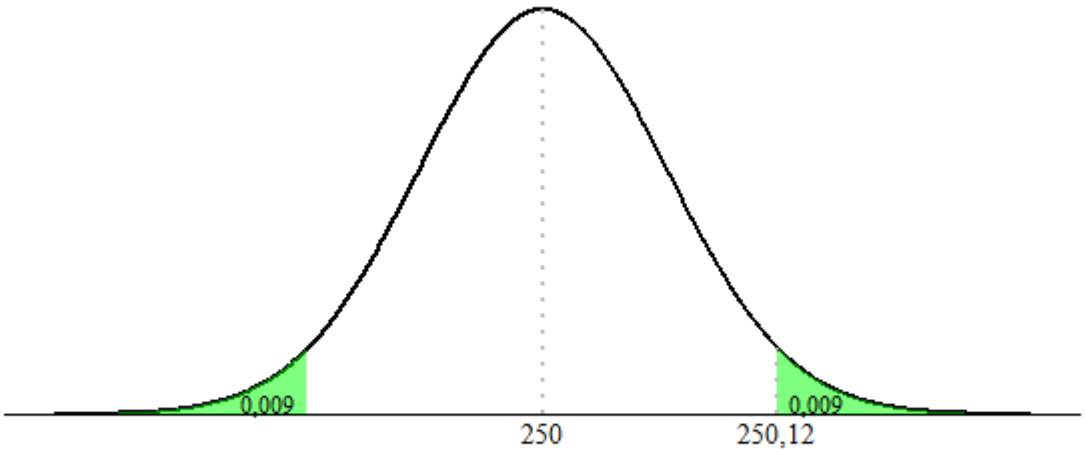


Figura 9.11 *Valor p. Prueba* $H_o: \mu = 250$ *vs* $H_a: \mu \neq 250$ *cuando* $\sigma = 0,2$. *Distribución Normal*

En la Figura 9.12 se observa el valor p asociado a la prueba de las siguientes hipótesis: $H_o: \mu = 250$ $H_a: \mu > 250$.
Dicho valor p se obtuvo suponiendo que la media muestral observada fue 250,02 mm.
En este caso, el valor p es 0,38, resulta mayor que cualquier α *razonable y, por lo tanto, no se debe rechazar* H_o.

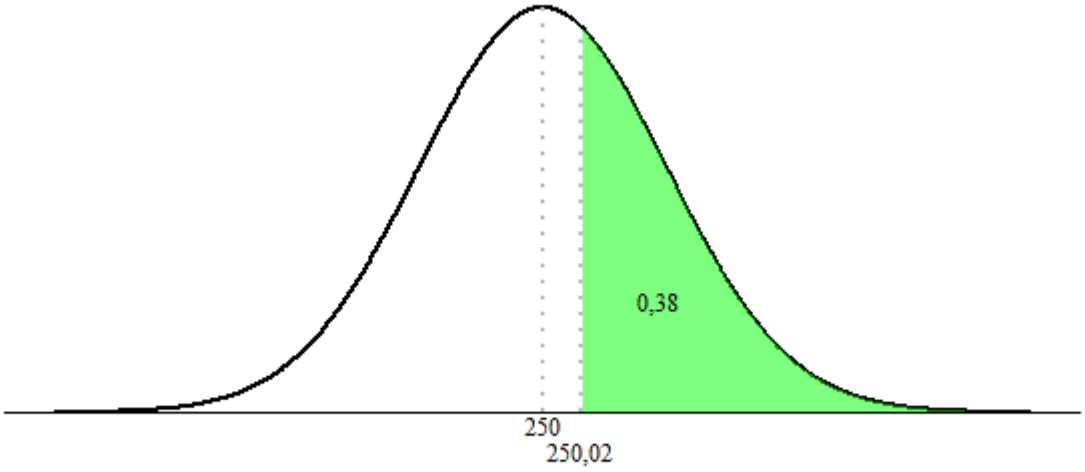


Figura 9.12 *Valor p. Prueba* $H_o: \mu = 250$ *vs* $H_a: \mu > 250$ *cuando* $\sigma = 0,2$. *Distribución Normal*

A medida que el estadístico de prueba se adentra en la región de rechazo, el valor p se hace más pequeño y aumenta el peso de la evidencia para rechazar la hipótesis nula. Haciendo un paralelismo con el procedimiento basado en la región de rechazo se puede establecer el criterio para rechazar o no a la hipótesis nula, de la siguiente manera:

- A partir del valor observado de estadístico, se rechaza H_o si el mismo pertenece a la región de rechazo.
- A partir del valor p, se rechaza H_o si valor $p \leq \alpha$.

Importante:

- Los paquetes de software estadístico, a partir de cargar los datos observados en la muestra, calculan automáticamente el valor p para cualquier prueba de hipótesis.
- Es importante destacar que el valor p es una de las formas de cuantificar la evidencia para rechazar H_o (cuanto menor es su valor, la evidencia en contra de H_o es mayor).
  Entonces, cuando se obtiene este valor, además de compararlo con α, se puede analizar el peso de la evidencia en contra de H_o. Por ejemplo, si valor $p < 0,001$, ya se sabe que la hipótesis nula se rechaza con mucha "confianza" de no equivocarse ya que el riesgo de cometer un error de tipo I es muy bajo.

h. Relación entre pruebas de hipótesis, intervalos de confianza y probabilidad de error de tipo II

¿Cuál es la probabilidad de cometer error de tipo I para una prueba estadística basada en un intervalo de confianza del 95 %?
En general, un valor hipotético (μ_o) para el promedio de la población (μ) [2] se puede rechazar con una probabilidad de error de tipo I (α), si y solo si μ_o no pertenece al intervalo de $(1-\alpha)\,100\,\%$ de confianza.

Para una prueba bilateral para μ, un valor hipotético (μ_o) no se rechaza a un nivel de significación $\alpha = 0,05$ si el valor observado del estadístico pivote $Z = \frac{\bar{Y}-\mu_o}{\frac{\sigma}{\sqrt{n}}}$ se encuentra en el intervalo: $-1,96 \leq \frac{\bar{y}_{obs}-\mu_o}{\frac{\sigma}{\sqrt{n}}} \leq 1,96$
Esto es equivalente a: $\bar{y}_{obs} - 1,96\,\frac{\sigma}{\sqrt{n}} \leq \mu_o \leq \bar{y}_{obs} + 1,96\,\frac{\sigma}{\sqrt{n}}$.

Observe que los extremos de la doble desigualdad coinciden con las expresiones de los extremos del intervalo de confianza de 95 % para μ cuando la desviación estándar poblacional, σ, es conocida. En este sentido, los intervalos de confianza y las pruebas de hipótesis dan resultados equivalentes.

Para una prueba unilateral a la izquierda, un valor hipotético (μ_o) no se rechaza a un nivel de significación $\alpha = 0,05$ si el valor observado del estadístico pivote Z resulta $\frac{\bar{y}_{obs}-\mu_o}{\frac{\sigma}{\sqrt{n}}} \geq -1,645$.
Si se despeja μ_o de esa desigualdad, se obtiene como resultado el intervalo de confianza del 95 % unilateral a la izquierda: $\mu_o \leq \bar{y}_{obs} + 1,645\,\frac{\sigma}{\sqrt{n}}$.

Para una prueba unilateral a la derecha, un valor hipotético (μ_o) no se rechaza a un nivel de significación $\alpha = 0,05$ si el valor observado del estadístico pivote Z resulta $\frac{\bar{y}_{obs}-\mu_o}{\frac{\sigma}{\sqrt{n}}} \leq 1,645$.
Si se despeja μ_o de esa desigualdad, se obtiene como resultado el intervalo de confianza de 95 % unilateral a la derecha: $\mu_o \geq \bar{y}_{obs} - 1,645\,\frac{\sigma}{\sqrt{n}}$.

En resumen, un intervalo de $(1-\alpha)\,100\,\%$ de confianza para la media poblacional μ está conformado por todos los valores de μ_o para los cuales no se rechaza la hipótesis $H_o: \mu = \mu_o$ en favor de la alternativa correspondiente, a un nivel de significación α, a partir del valor observado en la muestra.

[2] Lo afirmado en este texto para la media poblacional vale para cualquier otro parámetro de interés

Es importante tener en cuenta que, para que los resultados de pruebas de hipótesis e intervalos de confianza sean equivalentes, el valor de α debe ser el mismo en ambos casos y la hipótesis alternativa debe plantearse en el mismo sentido (bilateral, unilateral a la derecha o a la izquierda) que el intervalo de confianza.

¿Qué información adicional dan los intervalos de confianza?

- Cuando la hipótesis nula no se rechaza, los intervalos de confianza permiten dar una idea del riesgo de cometer un error tipo II. Un intervalo amplio indica un grado alto de incertidumbre y, por lo tanto, una probabilidad alta de cometer un error de tipo II.
- El intervalo de confianza no solamente sirve para saber si se rechaza o no a la hipótesis nula, si no que da una estimación del valor del parámetro.

i. Uso de las pruebas de hipótesis para la toma de decisiones sobre parámetros. Resultado significativo vs resultado relevante

En las páginas precedentes de este capítulo, una de las cuestiones principales que se desarrollan son los procedimientos para decidir si se rechaza una hipótesis o no. Retomando el ejemplo en el cual se plantea la hipótesis nula $H_o : \mu = 250$ contra $H_a : \mu > 250$, en este caso se analiza si hay evidencia en la muestra para poder concluir que la longitud promedio es mayor a 250 mm.

Como se menciona en la página 458, sobre el cálculo de la probabilidad de error tipo II, β, suponga que un cambio en el promedio a 250,15 mm se considera importante ya que, de ocurrir, significaría que el proceso está produciendo un alto porcentaje de barras fuera de especificaciones ($\mu_a = 250,15$).

Con respecto a las conclusiones que se toman al realizar la prueba de hipótesis se puede destacar:

- Si se rechaza H_o por cualquiera de las formas vistas (valor p, región de rechazo o intervalo de confianza), se obtiene un **resultado estadísticamente significativo**. Esto significa que se encontró en la muestra evidencia suficiente para rechazar lo postulado en la hipótesis nula, es decir, para afirmar que el valor del parámetro es mayor al postulado en H_o. Se concluye que es cierto lo planteado en la hipótesis alternativa, para el nivel de significación dado. Esto se relaciona con lo mencionado en el Capítulo 8, en ocasión de la interpretación de los intervalos de confianza (página 402) donde se remarca que con el intervalo se puede concluir que un parámetro es mayor, menor o distinto a un valor determinado, si dicho intervalo no incluye a ese valor (primer corolario). Ese valor determinado es el postulado en H_o.
- Si no se rechaza H_o, es decir, si el resultado no es estadísticamente significativo, es claro que esto no significa que la misma sea cierta, ya que se puede estar cometiendo un error de tipo II. Es decir, que no haya evidencia en la muestra en contra de que el valor del parámetro sea igual al postulado en H_o, no significa que sea igual a ese valor. Esto se relaciona con el segundo corolario de la página 402, en el que se afirma que, aunque el intervalo incluya al valor hipotético, no se puede concluir nunca que el parámetro es exactamente igual a dicho valor. Además, como se ve en la página 458, existe una probabilidad de cometer un error de tipo II para diferentes valores alternativos del parámetro. Esta probabilidad brinda información sobre el riesgo de concluir mal, al no descartar la hipótesis nula.
  En el ejemplo, para $\mu = 250,15; \quad \beta = 0,34$. Esto significa que si la longitud promedio poblacional aumentó realmente a 250,15, hay una probabilidad de 0,34 de no detectar ese cambio con la información brindada por la muestra (ya que no se rechaza H_o). Relacionándolo con el intervalo

de confianza, antes de tomar la muestra hay una probabilidad de que el mismo contenga a 250 cuando μ es igual a 250,15.

Así se rechace o no a la hipótesis nula, se puede profundizar el análisis teniendo en cuenta el contexto del problema, con la ayuda del intervalo de confianza.

- Si se rechaza H_o, es decir, si se concluye que el valor del parámetro es significativamente diferente al postulado en H_o, se puede analizar la importancia de esta "diferencia" en el contexto del problema.
  - Puede pasar que, por tener un tamaño de muestra muy grande, el estudio sea muy preciso y "detecte" cambios muy pequeños para el contexto. Esto significa que, puede rechazarse H_o y probarse que el valor del parámetro es diferente al postulado en H_o, pero el intervalo de confianza muestra que los posibles valores del parámetro, están muy cercanos al valor postulado. Es decir, con la información de la muestra se concluye que el parámetro no es igual a un determinado valor pero estaría muy cercano al mismo.
    En el ejemplo, si para las hipótesis mencionadas se obtiene el intervalo $IC_{95\%,\mu} = (250,002; 250,013)$, se observa que el mismo no incluye al valor 250, lo cual conduce al rechazo de H_o pero abarca valores que en el contexto del problema pueden considerarse muy cercanos (teniendo en cuenta que un cambio que realmente trae problemas al proceso ocurre cuando $\mu \geq 250,15$). Dicho de otra manera, a partir de lo analizado en la muestra se detectó un cambio en el promedio, μ no vale 250; pero en el contexto μ puede considerarse "prácticamente igual a 250". Se dice que este **resultado no es relevante o importante en el contexto**.
  - Por el contrario, si el intervalo quedara muy alejado, en contexto, del valor postulado, se concluiría no solamente que el valor del parámetro es distinto del postulado, si no que además tiene una diferencia importante en el contexto. Como por ejemplo, en el caso que el intervalo resulte $IC_{95\%,\mu} = (250,18; 250,33)$, lo que indica que el promedio poblacional está lo suficientemente alejado de 250 ($\mu > 250,15$) y significa que habría que hacer cambios o ajustes en el proceso.
- Si no se rechaza H_o, el análisis es el mismo que se describió en el Capítulo 8 (página 402), donde se mencionan diferentes casos en los que el intervalo contiene al valor postulado en H_o (en este ejemplo $\mu_o = 250$). En uno de los casos, el intervalo de confianza es $IC_{95\%,\mu} = (249,98; 250,03)$. Su amplitud es muy pequeña y permite concluir que el valor del parámetro es 250 o está muy próximo a este valor. En otro de los casos, el intervalo de confianza es $IC_{95\%,\mu} = (249,94; 250,22)$. Su amplitud es grande e incluye tanto al valor postulado en H_o como al valor 250,15, problemático para μ en el contexto de la situación planteada. Por ese motivo, aunque no se rechace H_o no se puede concluir que esta es cierta. En el primer caso (intervalo de amplitud pequeña), la probabilidad de error de tipo II es baja y en el segundo (intervalo de amplitud grande), es alta.
  La relación entre α, β y el tamaño de la muestra se mostró en la página 460, mediante una ecuación que involucra dichos valores. Recuerde que pensar que la amplitud del intervalo es pequeña o grande surge de analizar su magnitud en el contexto del problema.



Con lo visto hasta ahora, se pueden resolver las Actividades 1 a 4, Sección 9.6.

2. Prueba de hipótesis para la media poblacional, μ, cuando el desvío estándar poblacional, σ, es desconocido

En los capítulos 7 y 8 se estudió que cuando el parámetro de interés es la media poblacional, μ, pero no se conoce el valor de la desviación estándar poblacional, σ, se utiliza el estadístico pivote T, el cual tiene distribución t de Student con $(n-1)$ grados de libertad si $Y \sim N(\mu, \sigma)$ (Ver Capítulo 7, Sección 7.8.4): $T = \frac{\bar{Y}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$.

Se plantean las hipótesis nula y alternativa:
$H_o: \mu = \mu_o \quad H_a: \mu \neq \mu_o$ (caso bilateral).
$H_o: \mu = \mu_o \quad H_a: \mu > \mu_o$ (caso unilateral a la derecha).
$H_o: \mu = \mu_o \quad H_a: \mu < \mu_o$ (caso unilateral a la izquierda).

Luego se supone cierta H_o y se definen el estadístico de prueba y su distribución bajo H_o: $T = \frac{\bar{Y}-\mu_o}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$.

Definido el nivel de significación, α, la regla de decisión, en función de dicho estadístico, es la siguiente:

- Rechazar H_o en favor de H_a si $t_{obs} < t_{c_1}$ o $t_{obs} > t_{c_2}$ (prueba bilateral),
  donde $t_{obs} = \frac{\bar{y}_{obs}-\mu_o}{\frac{s_{obs}}{\sqrt{n}}}$ y t_{c_1} y t_{c_2} son los percentiles de orden $\frac{\alpha}{2}$ y $(1-\frac{\alpha}{2})$ de la distribución t-Student con $(n-1)$ grados de libertad.
- Rechazar H_o en favor de H_a si $t_{obs} > t_c$ (prueba unilateral a la derecha),
  donde t_c es el percentil de orden $(1-\alpha)$ de la distribución t-Student con $(n-1)$ grados de libertad.
- Rechazar H_o en favor de H_a si $t_{obs} < t_c$ (prueba unilateral a la izquierda),
  donde t_c es el percentil de orden α de la distribución t-Student con $(n-1)$ grados de libertad.

> Observe que, en este caso, no se puede plantear la regla de decisión en función del estadístico $\bar{Y}$ antes de tomar la muestra, ya que no se conoce el valor de la desviación muestral, s.

En las Figuras 9.13 y 9.14 se grafican las regiones de rechazo y de no rechazo de H_o para las hipótesis mencionadas anteriormente.

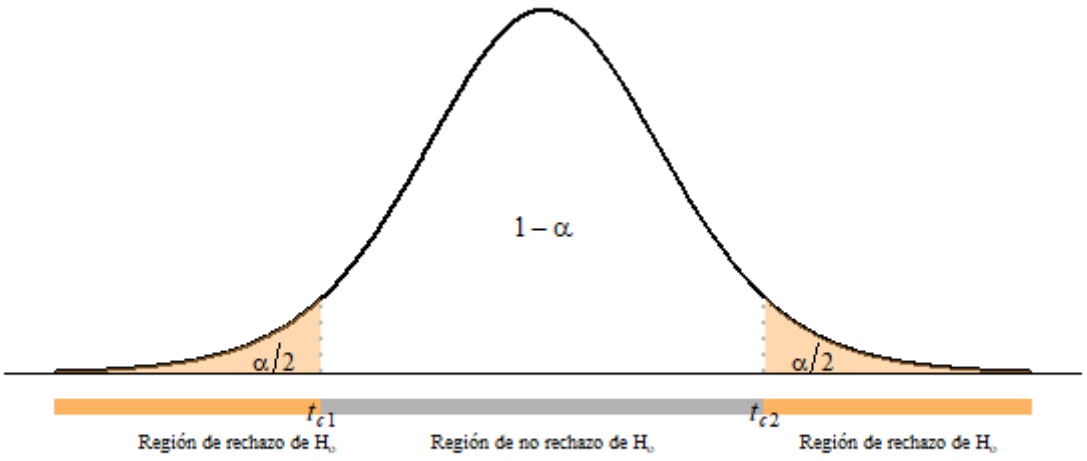


Figura 9.13 *Regiones de rechazo y no rechazo de H_o. Prueba bilateral para μ cuando σ es desconocido. Distribución t_{n-1}. El área de la región en naranja indica la probabilidad de cometer error de tipo I (α)*

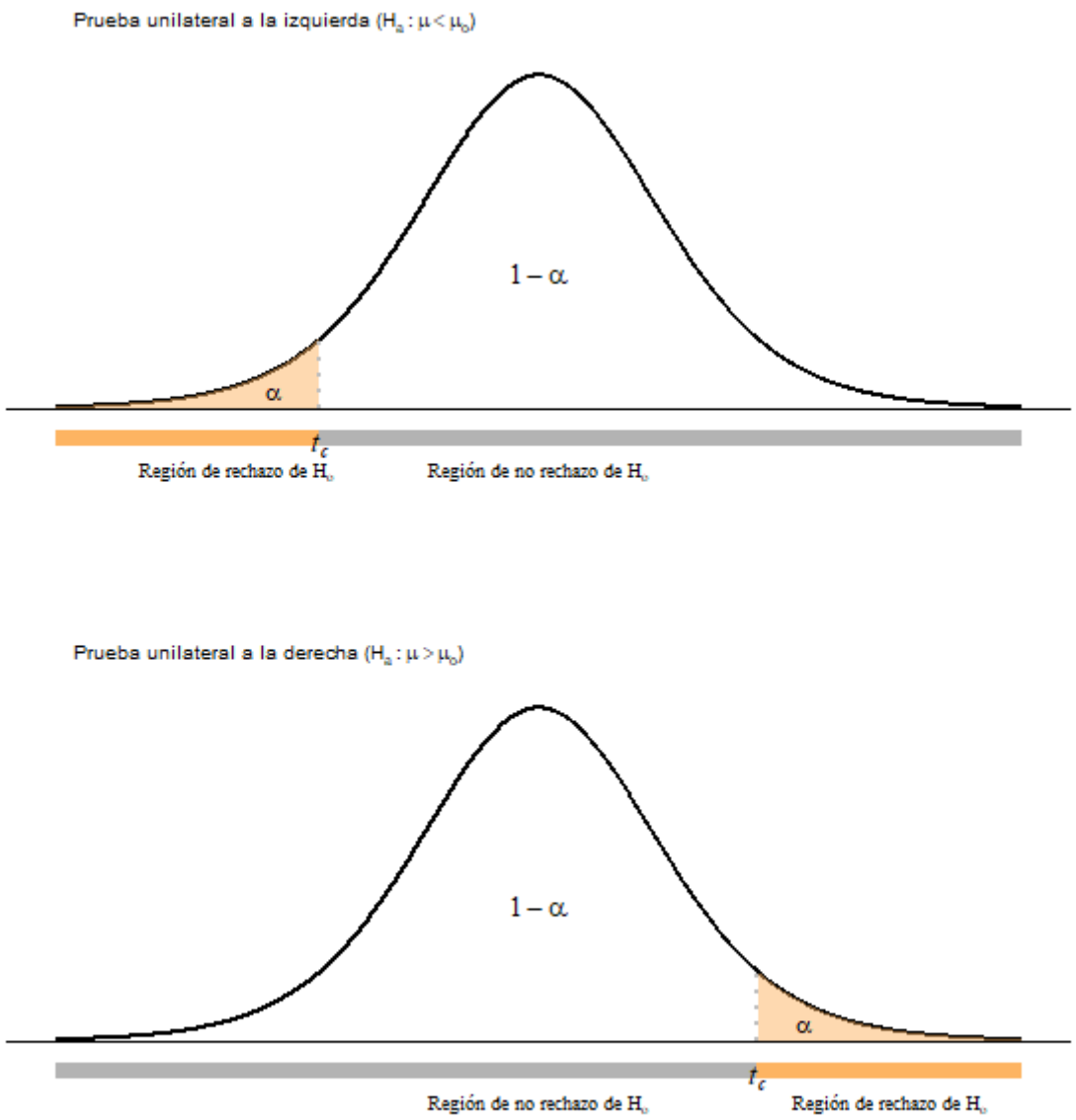


Figura 9.14 *Regiones de rechazo y no rechazo de H_o. Pruebas unilaterales para μ cuando σ es desconocido. Distribución t_{n-1}. El área de la región en naranja indica el máximo valor para la probabilidad de cometer error de tipo I (α)*

De forma análoga al caso en que la desviación estándar poblacional es conocida, cuando esta se desconoce también es posible tomar una decisión sobre las hipótesis planteadas utilizando el valor p o un intervalo de confianza. La diferencia principal es que, al no conocerse σ, se emplea la distribución t_{n-1} en los cálculos correspondientes.

Problema 1 (pág. 7): *Reconsidere la situación de las barras en la cual se quiere analizar, a partir de una muestra, si después de una serie de modificaciones implementadas en el proceso, la longitud promedio de las barras que se producen en el proceso, cambió. Suponga que ahora no se conoce el valor de la desviación estándar poblacional y que se tomó una muestra de tamaño n = 120. A partir de una muestra se obtuvo la longitud promedio de la muestra y la desviación estándar muestral, las cuales resultaron:* $\bar{y}_{obs} = 249,994\ mm$ *y* $s_{obs} = 0,3001\ mm$ *(Tabla 2.29, Capítulo 2).*

Las hipótesis nula y alternativa son:
$H_o: \mu = 250$
$H_a: \mu \neq 250.$

El estadístico de prueba es $T = \frac{\bar{Y}-250}{\frac{S}{\sqrt{120}}} \sim t_{119}$.

Se trabaja con $\alpha = 0,05$.

- *Si se define una región de rechazo (forma clásica).*
  La regla decisión es rechazar H_o *si* $t_{obs} < t_{119,\,0,025} = -1,98$ *o* $t_{obs} > t_{119,\,0,975} = 1,98$.

  El valor observado del estadístico de prueba es $t_{obs} = \frac{249,994-250}{\frac{0,3001}{\sqrt{120}}} = -0,219$.
  Como $-1,98 < t_{obs} < 1,98$, *no se puede rechazar* H_o *en favor de* H_a.

- *Si se obtiene el valor p.*
  valor $p = 2P(T < -0,219|H_o) = 0,414$.
  Como valor $p > \alpha$, *se obtiene la misma conclusión, es decir, no se rechaza* H_o *en favor de* H_a.
- *Si se construye un intervalo del 95 % de confianza para la longitud promedio de las barras,* μ.
  $IC_{95\%,\mu} = 249,994 \pm 1,98 \cdot \frac{0,3001}{\sqrt{120}}$
  $= 249,994 \pm 0,054 = (249,940; 250,048)$.
  Como el valor del parámetro postulado en la hipótesis nula ($\mu_o = 250$) *pertenece al intervalo de confianza obtenido, no se puede rechazar* H_o *en favor de* H_a.
  Si bien la conclusión es la misma que para los procedimientos anteriores, se puede analizar, además, el conjunto de posibles valores del parámetro que abarca el intervalo de confianza, en el contexto del problema: si todos los valores se consideran cercanos (en la práctica) al valor postulado en H_o, *podría decirse que se* ***"acepta la hipótesis nula"****.*

  A modo de ejemplo, suponga que un aumento o disminución de hasta 0,15 mm en la media poblacional no son importantes para el proceso. En este caso, el intervalo obtenido $IC_{95\%,\mu} = (249,94; 250,048)$ *indica que los valores posibles para la media poblacional se alejan como máximo en 0,06 mm, y por lo tanto, se podría "aceptar* H_o*" y considerar que, en la práctica, la longitud promedio poblacional* (μ) *es cercana a 250 mm.*

  En la Figura 9.15 se presentan las regiones de rechazo y de no rechazo de H_o, *junto con el valor observado. También se incluye información del valor p y del intervalo de confianza obtenidos.*

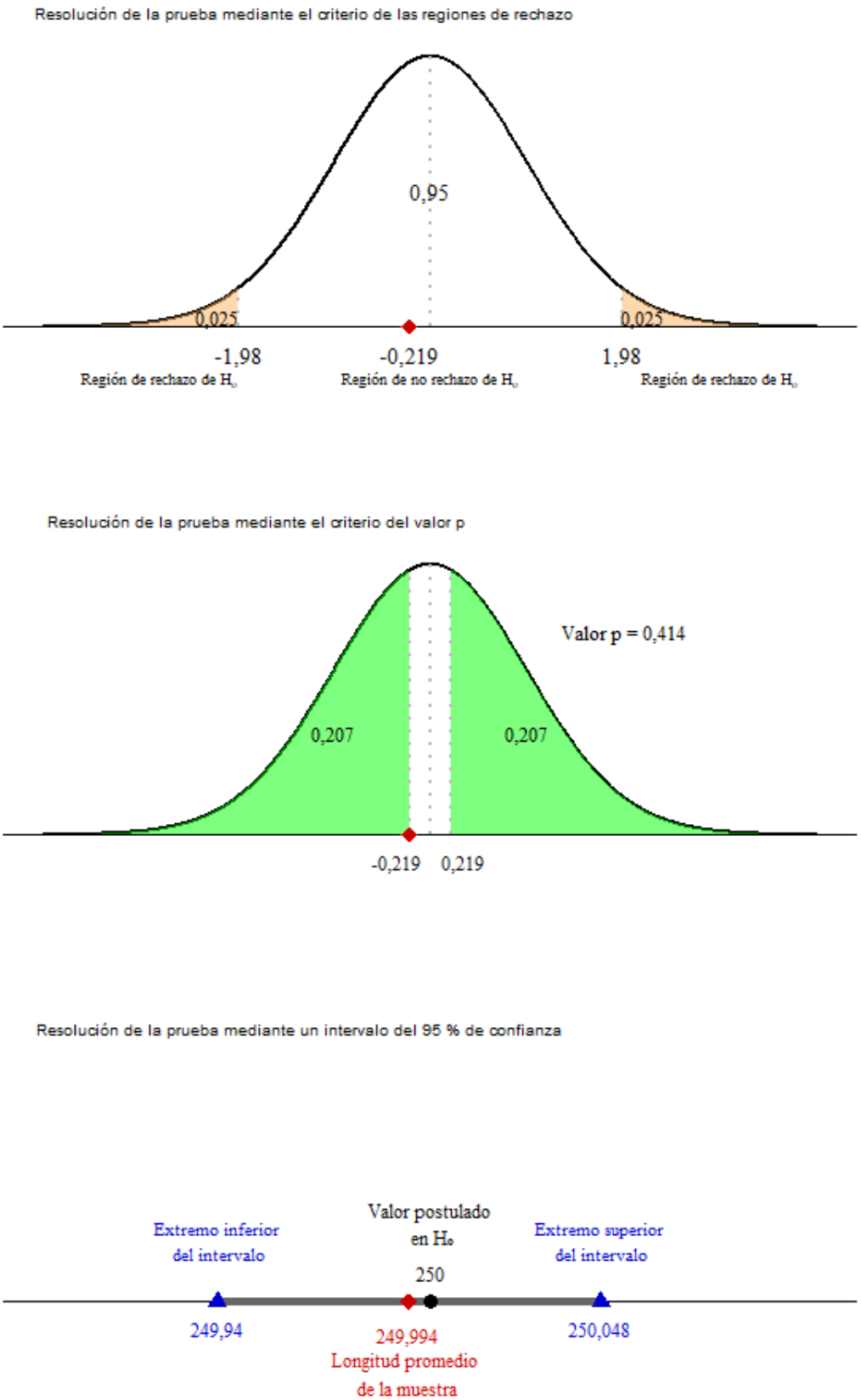


Figura 9.15 *Prueba* $H_o: \mu = 250$ *vs* $H_a: \mu \neq 250$ *cuando* σ *es desconocido. Distribución* t_{119}*. Resolución por diferentes caminos. El valor observado en la muestra se representa mediante un rombo rojo. El área de la región en naranja indica la probabilidad de cometer error de tipo I y la de la región en verde, el valor p*

Cuando no se conoce el valor de la desviación estándar poblacional (σ), no es posible calcular con exactitud la probabilidad de error de tipo II (β), aunque existen métodos para estimarla aproximadamente. En este libro, el cálculo de β, en pruebas de hipótesis sobre la media poblacional, solo se realiza en los casos en que se conoce la desviación estándar poblacional, σ.
Sin embargo, incluso cuando σ no se conoce, es posible formarse una idea general sobre si el riesgo de cometer un error de tipo II es alto o bajo, analizando la amplitud del intervalo de confianza en contexto. Una amplitud pequeña (en términos prácticos) sugiere que el riesgo β puede ser bajo. En cambio, una amplitud grande (en términos prácticos) indica que el riesgo β podría ser alto.

Como se mencionó al construir los intervalos de confianza para μ (página comentariot-normalidad), si la distribución de Y no es Normal pero es aproximadamente simétrica, puede emplearse la distribución t de Student para realizar pruebas de hipótesis respecto de μ, siempre que el tamaño de muestra no sea demasiado pequeño (se recomienda mayor a 10). Sin embargo, si la distribución de Y presenta una asimetría marcada, el uso de la t-student conduciría a resultados incorrectos independientemente del tamaño de muestra, ya que los riesgos reales asociados a las pruebas podrían superar considerablemente a los planteados originalmente.

Con lo visto hasta ahora, se pueden resolver las Actividades 5 a 9, Sección 9.6.

9.3.2 Pruebas de hipótesis para la variancia y el desvío estándar poblacionales

En los capítulos 7 y 8 se estudió que cuando el parámetro de interés es la variancia poblacional (σ^2), el estadístico correspondiente es la variancia muestral (S^2). Se define el estadístico pivote $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ que se comporta según tiene una distribución χ^2_{n-1} si $Y \sim N$.

Se plantean las hipótesis nula y alternativa:
$H_o: \sigma^2 = \sigma_o^2$ $H_a: \sigma^2 \neq \sigma_o^2$ (caso bilateral).
$H_o: \sigma^2 = \sigma_o^2$ $H_a: \sigma^2 > \sigma_o^2$ (caso unilateral a la derecha).
$H_o: \sigma^2 = \sigma_o^2$ $H_a: \sigma^2 < \sigma_o^2$ (caso unilateral a la izquierda).

Luego se supone cierta H_o y se definen el estadístico de prueba y su distribución bajo H_o de la siguiente manera:

$\chi^2 = \frac{(n-1)S^2}{\sigma_o^2} \sim \chi^2_{n-1}$.

Dado que el desvío estándar σ es una función estrictamente creciente de la variancia σ^2 (en el dominio positivo), cualquier hipótesis formulada en términos de σ puede reescribirse de manera equivalente en términos de σ^2, sin alterar la estructura del problema ni la regla de decisión. Es decir, trabajar con hipótesis sobre el desvío estándar o sobre la variancia resulta estadísticamente equivalente.
Sin embargo, debido a la forma en que está definido el estadístico de prueba, es necesario expresar la hipótesis nula en función de la variancia, ya que el estadístico pivote se construye directamente a partir del valor de σ^2 postulado en la hipótesis nula (σ_o^2).

Planteadas las hipótesis sobre la variancia, y fijado el nivel de significación α, la regla de decisión en función del estadístico de prueba es la siguiente:

- Rechazar H_o en favor de H_a si $\chi^2_{obs} < \chi^2_{c_1}$ o $\chi^2_{obs} > \chi^2_{c_2}$ (prueba bilateral), donde $\chi^2_{obs} = \frac{(n-1)s^2_{obs}}{\sigma^2_o}$ es el valor observado del estadístico de prueba y $\chi^2_{c_1}$ y $\chi^2_{c_2}$ son los percentiles de orden $\frac{\alpha}{2}$ y $(1-\frac{\alpha}{2})$ de la distribución χ^2 con $(n-1)$ grados de libertad.
- Rechazar H_o en favor de H_a si $\chi^2_{obs} > \chi^2_c$ (prueba unilateral a la derecha), donde χ^2_c es el percentil de orden $(1-\alpha)$ de la distribución χ^2 con $(n-1)$ grados de libertad.
- Rechazar H_o en favor de H_a si $\chi^2_{obs} < \chi^2_c$ (prueba unilateral a la izquierda), donde χ^2_c es el percentil de orden α de la distribución χ^2 con $(n-1)$ grados de libertad.

En la Figura 9.16 se representa gráficamente la regla de decisión para la prueba bilateral y en la Figura 9.17 se representan gráficamente las reglas para las pruebas unilaterales.

Análogamente a lo visto para las pruebas de hipótesis respecto de la media poblacional, también se puede decidir sobre las hipótesis planteadas a partir del valor p o de un intervalo de confianza. En este caso, los cálculos correspondientes se realizan a partir de la distribución χ^2_{n-1}.

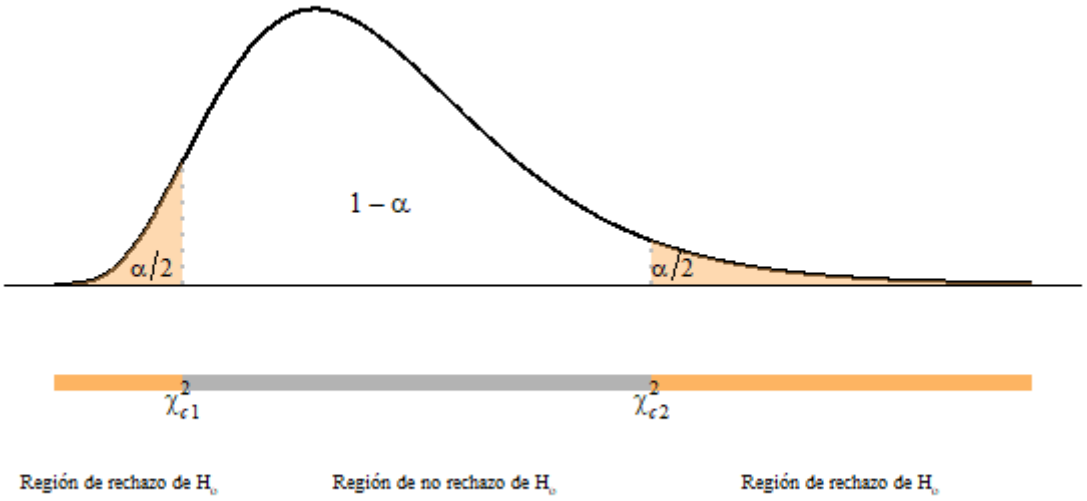


Figura 9.16 *Regiones de rechazo y no rechazo de H_o. Prueba bilateral para σ^2. Distribución χ^2_{n-1}. El área de la región en naranja representa la probabilidad de cometer error de tipo I (α)*

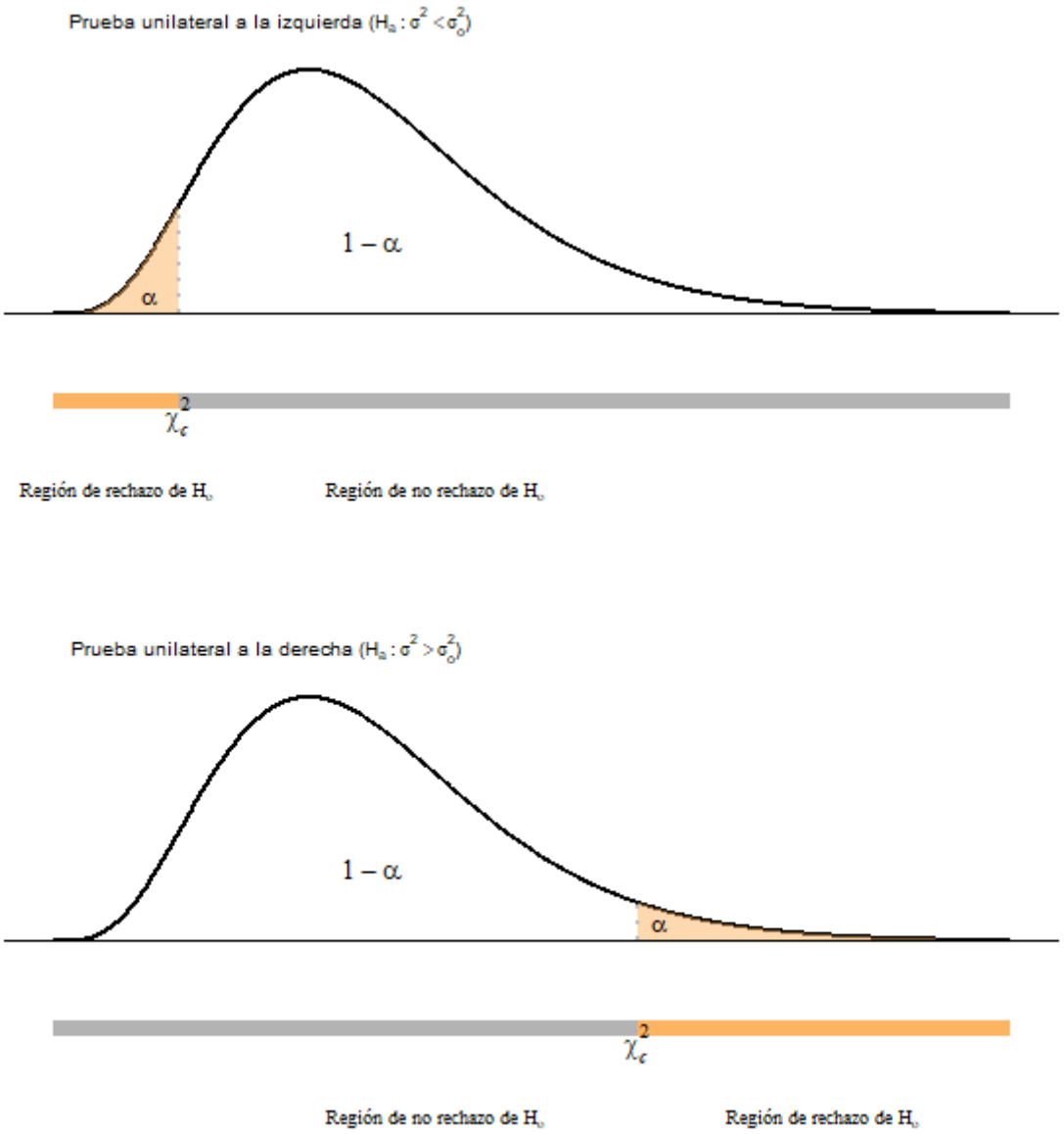


Figura 9.17 *Regiones de rechazo y no rechazo de H_o. Pruebas unilaterales para σ^2. Distribución χ^2_{n-1}. El área de la región en naranja representa el máximo valor de la probabilidad de cometer error de tipo I (α)*

Problema 1 (pág. 7): *Reconsidere la situación de la empresa metalúrgica interesada en estudiar el comportamiento de la longitud de las barras. Suponga que ahora se piensa que lo que pudo haber cambiado es la variancia poblacional de las longitudes, σ^2, la cual tenía originalmente un valor de $(0,2\ mm\ )^2 = 0,04\ mm^2$. Para comprobar esto se toma una muestra de 120 barras, en la que se obtuvo $s^2 = 0,091\ mm^2$.*

Si la longitud de las barras (Y) se puede considerar normalmente distribuida, entonces, el estadístico de prueba $X^2 = \frac{119S^2}{0,04} \sim \chi^2_{119}$.

Para un nivel de significación $\alpha = 0,05$, se obtienen los valores críticos de la distribución χ^2_{119} : $\chi^2_{119,\,0,025} = 90,699$ y $\chi^2_{119,\,0,975} = 151,084$.

El valor observado del estadístico, $x^2_{obs} = \frac{119 \cdot 0,091}{0,04} = 270,725$.

Como $x^2_{obs} > \chi^2_{119,\,0,975}$, se rechaza H_o en favor de H_a.

Se concluye, con un nivel de significación del 5 %, que la variancia poblacional cambió, que es diferente de $0,04\ mm^2$.
El valor p, se calcula de la siguiente manera:
valor $p = 2P(X^2 > 270,725|H_o\ cierta) = 1,5E-13$.
Dado que valor $p < \alpha$*, se obtiene la misma conclusión.*
También se puede tomar la decisión a partir del intervalo de confianza del 95 %:
$IC_{95\%,\sigma^2} = (0,072; 0,119)$
Observe que el intervalo está conformado en su totalidad por valores mayores que 0,04 (valor de la variancia postulado en la hipótesis nula), motivo por el cual se rechaza H_o.



Cuando la población de la cual se extrajo la muestra no puede considerarse normalmente distribuida, la prueba de hipótesis para este parámetro puede realizarse mediante el método de Bonett, el cual resulta robusto ante desviaciones de la normalidad.



En este libro no se aborda el cálculo de (β) en pruebas de hipótesis sobre la variancia poblacional; pero se sugiere analizar si el riesgo de cometer un error de tipo II es alto o bajo, analizando la amplitud del intervalo de confianza en contexto, como se mencionó para el caso de las pruebas de hipótesis para la media poblacional (μ), cuando la desviación estándar poblacional (σ) es desconocida.



Con lo visto hasta ahora, se pueden resolver las Actividades 10 a 12, Sección 9.6.

9.3.3 Prueba de hipótesis para la proporción de unidades de la población que cumplen cierta condición

Cuando el parámetro de interés es la proporción de unidades de la población que cumplen cierta condición (π), en la población se define una variable Y de la siguiente forma:

$Y = 1$ si se verifica la condición de interés e
$Y = 0$ si dicha condición no se verifica, donde $P(Y = 1) = \pi$.
Es decir, Y se comporta según el modelo Bernoulli, $Y \sim Be(\pi)$.

Como estadísticos se pueden definir dos, f_r y C, donde f_r es la proporción de éxitos en la muestra y C, la cantidad de unidades de la muestra que verifica la condición de interés. $f_r = \frac{C}{n}$ y $C \sim Bi(n, \pi)$.

Para desarrollar el intervalo de confianza exacto se presentaba el inconveniente de no poder obtener una expresión para el intervalo aleatorio ni una fórmula para el intervalo de confianza. Se presentó entonces un método numérico para obtener el intervalo exacto y también se presentaron fórmulas aproximadas basadas en la distribución Normal, dada la complicación para realizar el cálculo del intervalo exacto sin ayuda de software.

En el caso de las pruebas de hipótesis, los procedimientos vistos se basan en suponer cierta H_o. Esto hace que, al asignarle a π el valor π_o, se facilite el procedimiento, trabajando directamente con la distribución binomial sin necesidad de definir ninguna fórmula aproximada.

Los pasos para el procedimiento basado en una región de rechazo (forma clásica) son los siguientes:

- Se definen las hipótesis nula y alternativa
  $H_o: \pi = \pi_o \quad H_a: \pi \neq \pi_o$ (prueba bilateral).
  $H_o: \pi = \pi_o \quad H_a: \pi > \pi_o$ (prueba unilateral a la derecha).
  $H_o: \pi = \pi_o \quad H_a: \pi < \pi_o$ (prueba unilateral a la izquierda).

- Se define el estadístico de prueba y su distribución de probabilidades bajo H_o.
  En este caso, el estadístico C es la cantidad de unidades de la muestra que verifica la condición de interés: $C \sim Bi(n, \pi_o)$.

- Se define el nivel de significación α y se obtienen los valores críticos, definiendo la siguiente regla de decisión:
  - Rechazar H_o en favor de H_a si $c_{obs} \leq c_{c_1}$ o $c_{obs} \geq c_{c_2}$ (prueba bilateral),
    donde c_{c_1} es el mayor valor de C que acumula una probabilidad menor o igual que $\frac{\alpha}{2}$ y c_{c_2} es el menor valor de C que acumula una probabilidad mayor o igual que $1 - \frac{\alpha}{2}$, suponiendo la distribución de C bajo H_o.
  - Rechazar H_o en favor de H_a si $c_{obs} \geq c_c$ (prueba unilateral a la derecha),
    donde c_c es el menor valor de C que acumula una probabilidad mayor o igual que $1 - \alpha$.
  - Rechazar H_o en favor de H_a si $c_{obs} \leq c_c$ (prueba unilateral a la izquierda),
    donde c_c es el mayor valor de C que acumula una probabilidad menor o igual que α.

- A partir de la muestra, se obtiene el valor observado del estadístico y finalmente se toma la decisión.



Dado que el estadístico de esta prueba es una variable aleatoria discreta, es necesario considerar las siguientes particularidades:

- En la regla de decisión no es equivalente utilizar desigualdades estrictas ($<$ o $>$) que no estrictas ($\leq$ o $\geq$).
- La función de distribución acumulada presenta discontinuidades (saltos) en cada valor entero, lo que implica que el nivel de significación real (α real) asociado a la regla de decisión definida rara vez coincidirá con el nivel de significación deseado (α). Dado que α representa el valor máximo admisible para la probabilidad de error de tipo I, es fundamental que el nivel de significación real no lo exceda.

En las Figuras 9.18 y 9.19 se grafican las regiones de rechazo y de no rechazo de H_o para las hipótesis mencionadas anteriormente. Recuerde que el valor de α que se indica en las mismas resultará casi siempre menor que el deseado.

Todos los conceptos vistos en los casos anteriores son similares pero trabajando con la distribución Binomial asignándole a π el valor postulado en la hipótesis nula, π_o.

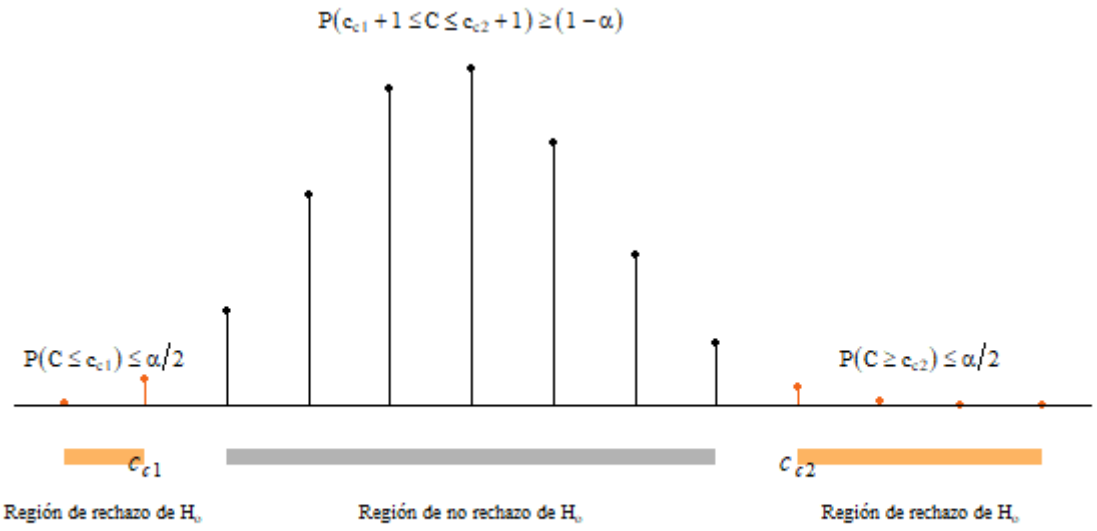


Figura 9.18 *Regiones de rechazo y no rechazo de H_o. Prueba bilateral para π. Distribución $Bi(n, \pi_o)$. La suma de las alturas de los bastones en naranja representa la probabilidad de cometer error de tipo I ($\le \alpha$)*

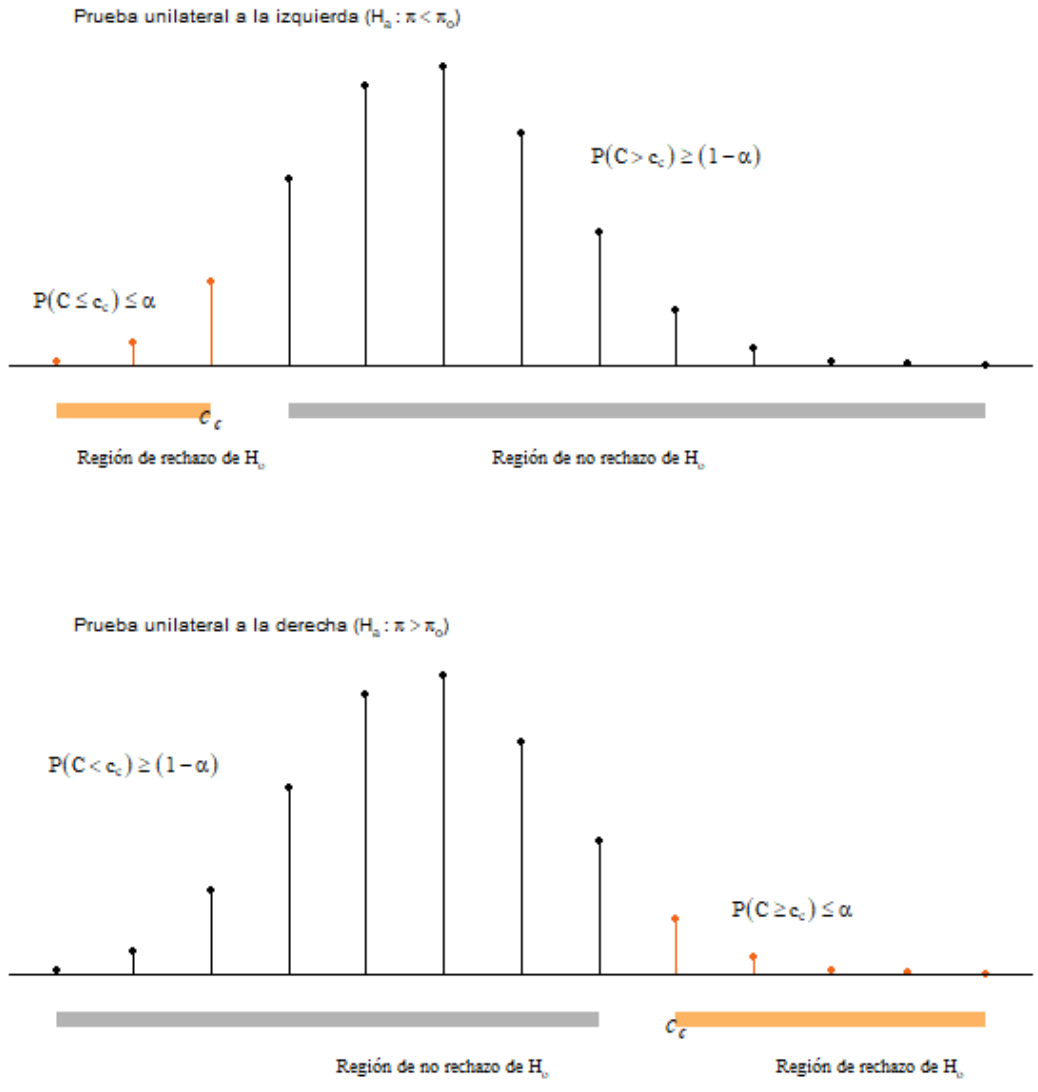


Figura 9.19 *Regiones de rechazo y no rechazo de H_o. Pruebas unilaterales para π. Distribución $Bi(n, \pi_o)$. La suma de las alturas de los bastones en naranja representa el máximo valor para la probabilidad de cometer error de tipo I ($\le \alpha$)*

Problema 1 (pág. 7): *Reconsidere la situación de la empresa metalúrgica productora de barras metálicas. Suponga que ahora interesa la proporción de barras que no cumplen las especificaciones en relación a la longitud, es decir, la proporción de barras con longitudes que no se encuentran entre 249,4 mm y 250,6 mm. Se quiere analizar si esta proporción es menor a 0,10, para poder considerar al proceso como de calidad aceptable. A partir de una muestra aleatoria de* $n = 120$ *barras se encontró que 114 cumplían con las especificaciones y 6 no.*

Se plantean las hipótesis: $H_o: \pi = 0,10 \quad H_a: \pi < 0,10$ *(prueba unilateral a la izquierda).*

El estadístico en este caso es C: cantidad de barras de la muestra que tienen longitudes fuera del intervalo de especificaciones, es decir, menores que 249,4 mm o mayores que 250,6 mm.

Bajo H_o, $C \sim Bi(120; 0,1)$.

Se plantea un nivel de significación $\alpha = 0,05$.

Entonces, se puede obtener el valor crítico buscando el mayor valor de C que acumula una probabilidad menor o igual que 0,05 en la distribución $Bi(120; 0,1)$. *Este valor es* $c_c = 6$, *que acumula una probabilidad igual a 0,038.*

La regla de decisión consiste en rechazar H_o *en favor de* H_a *si* $c_{obs} \leq 6$.

En la Figura 9.20 se visualiza la probabilidad acumulada hasta el valor 6 (0,038), que es el α *real, inferior al originalmente planteado de 0,05. Si se tomara como valor crítico al 7, el* α *real sería 0,078, y se correría un riesgo de error tipo I mayor al deseado.*

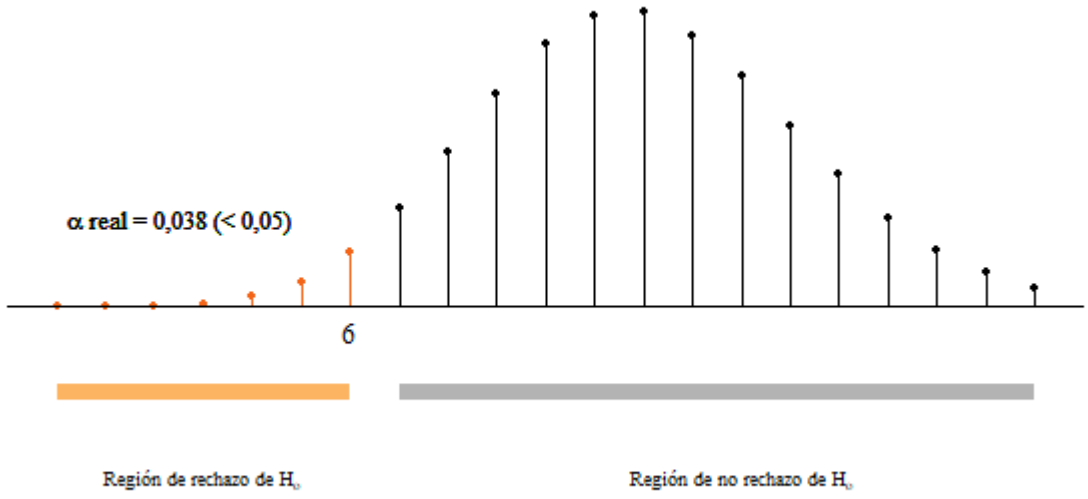


Figura 9.20 *Regiones de rechazo y no rechazo de* H_o. *Prueba* $H_o: \pi = 0,10 \quad vs \quad H_a: \pi < 0,10$. *Distribución* $Bi(n, \pi_o)$. *Nivel de significación real 0,038.*
La figura muestra únicamente los primeros 20 valores de la distribución

En esta muestra, el valor observado de C resultó $c_{obs} = 6$ ($c_{obs} \leq c_{crtico}$) *por lo que la decisión es rechazar* H_o *en favor de* H_a *y concluir, con un riesgo del 5 % (en realidad, del 3,82 %) que la proporción de barras que no cumplen con las especificaciones relativas a la longitud es menor que 0,10.*

Si se resolviera utilizando el valor p, siguiendo lo visto anteriormente quedaría:

valor $p = P(C \leq 6|H_o) = 0,0382 < \alpha$ *y se llegaría a la misma conclusión. Observe que en este caso particular, la probabilidad asociada a la región de rechazo coincide con la correspondiente al valor p, ya que el valor observado del estadístico coincide con el valor crítico.*

Cálculo de la probabilidad de error de tipo II, β

Al igual que se planteó respecto del cálculo de β para el caso de pruebas respecto de la media poblacional (μ) cuando se conoce el valor de la desviación estándar poblacional (σ) (página 458), para las pruebas respecto de la proporción poblacional (π) basadas en la distribución exacta también se puede obtener de manera sencilla la probabilidad de cometer error de tipo II. Para ello se requiere determinar un valor alternativo del parámetro, π_a, y calcular la probabilidad de no rechazar H_o, suponiendo que $\pi = \pi_a$, es decir, suponiendo que $C \sim Bi(n, \pi_a)$. A continuación se presenta un ejemplo.

Problema 1 (pág. 7): *Anteriormente se plantearon las hipótesis* $H_o: \pi = 0,10 \quad H_a: \pi < 0,10$
Si la hipótesis nula es cierta, el proceso no alcanza el nivel de calidad aceptable y por lo tanto debe implementarse un plan de mejoras; en cambio, si la hipótesis alternativa es verdadera, el proceso se considera de calidad aceptable y no requiere mejoras.
En este contexto, los posibles errores al tomar una decisión son:

- ***error de tipo I**: rechazar H_o cuando es cierta; es decir, considerar que el proceso es de calidad aceptable (y no implementar las mejoras) cuando en realidad no lo es;*
- ***error de tipo II**: no rechazar H_o cuando es falsa; es decir, considerar que el proceso no es de calidad aceptable (e implementar las mejoras) cuando en realidad estas no se requieren.*

El riesgo de error de tipo I para esta prueba está acotado por $\alpha = 0,05$. ¿Cuál es la cota superior para el riesgo de error de tipo II? Para responder a esta pregunta, se necesita determinar, en contexto, un valor alternativo del parámetro de interés.

Suponga que la empresa no desearía implementar el plan de mejoras si la proporción de barras que cumplen con las especificaciones fuese menor o igual que 0,05, ya que esto implicaría incurrir en altos costos cuando la calidad del proceso es adecuada. Entonces es de interés calcular la probabilidad de no rechazar H_o bajo ese supuesto, es decir, la probabilidad de no rechazar H_o para $\pi = 0,05$.
Manteniendo la regla de decisión asociada al mismo nivel de significación, $\alpha = 0,05$, que conduce a rechazar H_o si $C \leq 6$, y teniendo en cuenta que ahora $C \sim Bi(120; 0,05)$, el cálculo de β es:
$\beta_{\pi_a=0,05} = P(\text{no rechazar } H_o | H_o \text{ es falsa}) = P(C > 6 | \pi_a = 0,05) = 0,3936.$
La probabilidad de error de tipo II obtenida se presenta en la Figura 9.21.

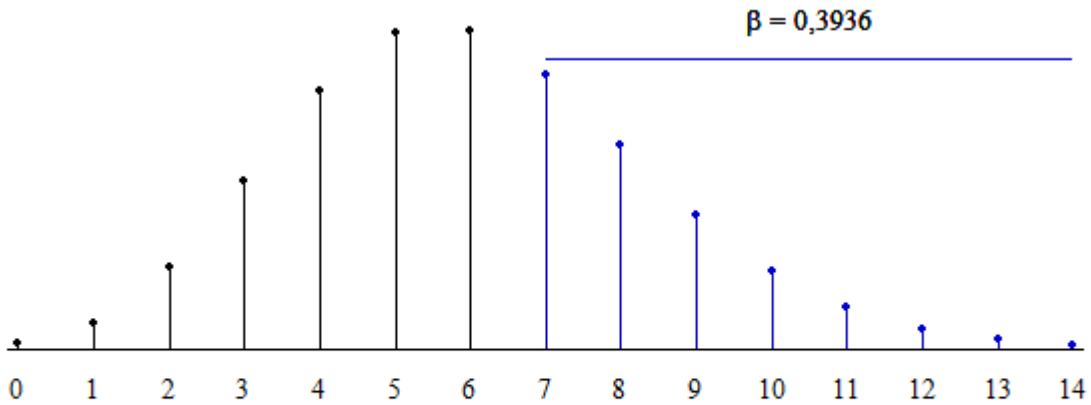


Figura 9.21 *Probabilidad de error de tipo II para $\pi_a = 0,05$ (en azul). Prueba $H_o: \pi = 0,10$ vs $H_a: \pi < 0,10$. Distribución $Bi(120; 0,05)$*

> *Esto quiere decir que, si la proporción de barras que no cumplen las especificaciones fuese de 0,05, existe una chance de aproximadamente 0,39 de no rechazar H_o (en otras palabras, de "no detectar" dicha disminución con la información de la muestra). Note que, si π fuese menor que 0,05, la probabilidad de error de tipo II sería menor, por lo que el valor obtenido constituye una cota superior para ese riesgo, para valores de $\pi_a \leq 0,05$ (Tabla 9.2, página 458).*

Relación entre la prueba de hipótesis y el intervalo de confianza exacto para π

En el Capítulo 8 se presenta el intervalo de confianza para la proporción basado en la distribución Binomial (distribución exacta del estadístico). Allí se menciona que no existe una fórmula para obtener los extremos del intervalo y que por eso se obtienen mediante un procedimiento computacional. Una vez vista la prueba de hipótesis para la proporción, se puede explicar más sencillamente este método computacional. Al obtener el valor observado del estadístico (c_{obs}), el intervalo de $(1-\alpha)\,100\,\%$ de confianza queda conformado por todos los valores de π_o para los cuales no se rechazaría la hipótesis nula $H_o: \pi = \pi_o$ en favor de la alternativa propuesta, para un nivel de significación α.
En el Capítulo 8, Sección 9.3.1 se describió esta relación para el caso de la media poblacional μ cuando se conoce la desviación estándar poblacional, σ.

En el ejemplo anterior, se obtuvo un valor de $c_{obs} = 6$. En ese caso, se había planteado $H_o: \pi = 0,10$ $H_a: \pi < 0,10$, y el valor p obtenido fue menor que α, por lo que se tomó la decisión de rechazar H_o.

Esto significa que el valor 0,1 no estará incluído en un intervalo de confianza del 95 % unilateral a la izquierda.

El intervalo exacto obtenido computacionalmente es (0 ; 0,0963).
Entonces, se puede afirmar que si se planteara una prueba con $H_o: \pi = \pi_o$ $H_a: \pi < \pi_o$ y en la muestra se observara un valor del estadístico igual a 6, para cualquier valor de $\pi_o > 0,0963$ se obtendría que valor $p < \alpha$, es decir, se rechazaría H_o en favor de H_a y para cualquier valor de $\pi_o < 0,0963$ se obtendría valor $p > \alpha$ (no se rechazaría H_o).



Con lo visto hasta ahora, se pueden resolver las Actividades 13 a 16, Sección 9.6.

Aplicación al muestreo de aceptación: Inspección de aceptación simple por atributo

En situaciones donde una empresa (consumidor) compra lotes de unidades a otra (productor) y esta última no realiza un control estadístico de procesos para garantizar los porcentajes máximos de unidades que no cumplen las especificaciones, la empresa compradora puede realizar una inspección por muestreo para decidir la aceptación o no de cada lote.

Suponga que las unidades se envían de productor a consumidor en lotes que varían en tamaño. Idealmente, cada lote no debería contener unidades "defectuosas", pero esto no siempre ocurre [3]. Se necesita evidencia, entonces, de que la proporción de unidades "no conformes a especificaciones" o "defectuosas" no es excesiva y, por lo tanto, se requiere inspeccionar el lote.

[3] El control estadístico de procesos brinda las herramientas para mejorar los procesos y reducir al mínimo posible la proporción de unidades defectuosas del mismo.

La inspección puede ser exhaustiva (es decir, aplicarse al 100 % del lote), aunque en la práctica se recurre con mayor frecuencia a la inspección por muestreo. Esto se debe a diversas razones, como, por ejemplo, que los ensayos pueden ser destructivos o que una inspección total implica costos y tiempos excesivos. Además, las inspecciones del 100 % de las unidades suelen ser tediosas, lo que incrementa la posibilidad de no detectar las defectuosas. Por el contrario, la inspección por muestreo presenta ventajas significativas. Una de las más importantes es que ejerce una presión más efectiva sobre la mejora continua de la calidad, ya que puede conducir al rechazo de lotes completos, mientras que en una inspección del 100 % solo se rechazan los productos individuales defectuosos.

El proveedor debería hacer un esfuerzo continuo por hacer las cosas bien desde el principio: no hay razón para aceptar a la larga una proporción fija de unidades defectuosas. Sin embargo, en muchos casos se sigue decidiendo la aceptación de lotes por muestreo (lo cual se denomina “plan de inspección de aceptación” o “plan de aceptación por muestreo”).

En la inspección de aceptación por atributo, una unidad defectuosa es aquella que no se apega a las especificaciones de calidad fijadas en una o más características. Un procedimiento corriente en la aceptación por muestreo es considerar por separado cada lote de producto sometido y basar la decisión de aceptarlo o rechazarlo, en la evidencia de una o más muestras escogidas aleatoriamente de cada uno de ellos. Así se derivan planes de aceptación por muestreo simples, dobles, múltiples y secuenciales, según la cantidad de muestras en que se base la decisión. En este texto, solo se describe el plan de aceptación por muestreo simple por atributo, en el cual se decide aceptar o rechazar el lote en base al número de unidades defectuosas encontrados en una única muestra aleatoria extraída de dicho lote.

En los acuerdos entre productor y consumidor, se puede considerar alguna forma de crédito por las unidades defectuosas encontradas por el consumidor. El rechazo de un lote no significa que este deba ser destruido: solo que debe ser sometido a una inspección más rigurosa (por parte del productor o del consumidor) con la idea de eliminar las unidades defectuosas.

Se define: N: tamaño del lote que va a ser evaluado, antes de pasar a producción.
π: proporción de unidades defectuosas del lote.
π_o: máxima proporción de unidades defectuosas aceptable para ese lote o “nivel de calidad aceptable” (AQL, en inglés).
C: cantidad de unidades defectuosas en una muestra de tamaño n, seleccionada de dicho lote.

Entonces, se plantean las siguientes hipótesis:
$H_o: \pi = \pi_o$ (el lote es bueno y se debería aceptar)
$H_a: \pi > \pi_o$ (el lote es malo y se debería rechazar)

La regla de decisión es la siguiente: Rechazar el lote si $C \geq c_c$, donde c_c es el valor crítico de la prueba de hipótesis.

> ! Observe que la calidad se define negativamente a través de la proporción de unidades defectuosas del lote; de ahí que un valor alto para este parámetro indique que el lote no es bueno y debe ser rechazado.

También es importante tener en cuenta un valor $\pi_a > \pi_o$ como la máxima proporción de unidades defectuosas tolerada del lote ($LTPD$, en inglés), a partir de la cual es importante no aceptar el lote.

En este contexto, el **error de tipo I** consiste en rechazar un lote "bueno", es decir, un lote en el cual la proporción de unidades defectuosas es menor o igual que AQL; mientras que el **error de tipo II** consiste en no rechazar un lote "malo", es decir, un lote en el cual la proporción de defectuosas es mayor que AQL, y especialmente mayor que $LTPD$.

Recordando las dos interpretaciones de la probabilidad presentadas en el Capítulo 4, sobre las probabilidades de error de tipo I y de tipo II, en este caso se puede decir lo siguiente:

- La primera interpretación es la que se viene utilizando en las pruebas de hipótesis: cada una de estas probabilidades representa el riesgo de cometer error de tipo I o de tipo II respectivamente, en una única realización de la prueba de hipótesis, es decir, al decidir sobre un único lote.
- La segunda interpretación supone que se realiza una gran cantidad de pruebas de hipótesis (una por cada lote que va a ingresar al proceso). En este contexto,

    - La probabilidad de cometer error de tipo I se interpreta como la proporción de lotes "buenos" (es decir, de lotes para los cuales $\pi \leq \pi_o$) que serán rechazados. El máximo valor para esta proporción, α, se obtiene para el caso $\pi = \pi_o = AQL$ y se denomina **"riesgo del productor"**.
    - La probabilidad de cometer error de tipo II, obtenida suponiendo que el lote tiene una cierta proporción π_a de unidades defectuosas, se interpreta como la proporción de lotes "malos" (es decir, de lotes para los cuales $\pi = \pi_a$) que serán aceptados.
      Recuerde que la probabilidad de cometer error de tipo II depende del valor real de la proporción de defectuosos del lote, π. Para cada valor $\pi_a > AQL$ puede calcularse una probabilidad de error de tipo II distinta.
      La probabilidad de error de tipo II, β, se denomina **"riesgo del consumidor"**. Entre los valores de β que resultan relevantes para el consumidor, el máximo se obtiene cuando $\pi_a = LTPD$.

Definidos $AQL(\pi_o)$ y $LTPD(\pi_a)$ en el contexto del problema, se definen el tamaño de la muestra y el valor crítico (que depende de α). Con esa información se puede obtener el riesgo β que interesa al consumidor.

Por tratarse de una prueba de hipótesis, un plan de muestreo puede caracterizarse mediante su curva OC, la cual describe el nivel de protección ofrecido por el plan frente a lotes de distintas calidades, mostrando la probabilidad de aceptar el lote en función de π.

Esta probabilidad se calcula de manera exacta utilizando la distribución Hipergeométrica (ya que el lote es finito, de tamaño N); sin embargo, cuando el tamaño del lote es grande, suele emplearse una aproximación mediante la distribución Binomial. Calculando la probabilidad $P(C < c_c|\pi)$ para distintos valores de π, se obtiene la curva OC, que se presenta en la Figura 9.22.

En el eje de abscisas se representan posibles valores de π, entre ellos AQL y $LTPD$; mientras que en el eje de ordenadas se presenta la probabilidad de no rechazar H_o, es decir, la probabilidad de aceptar un lote con una proporción de unidades defectuosas igual a π.

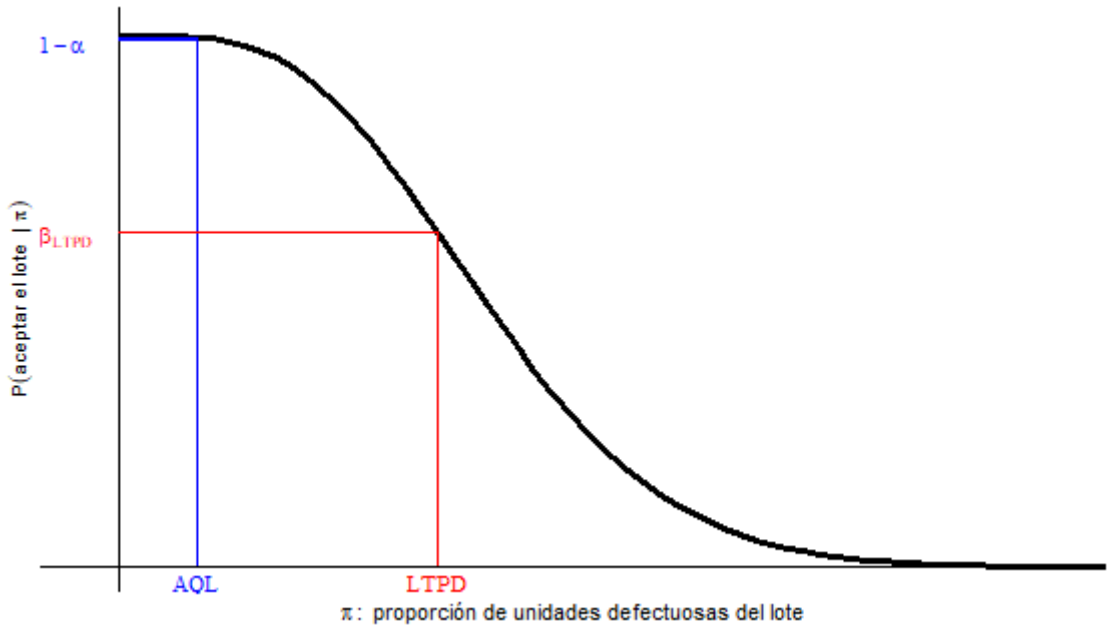


Figura 9.22 *Curva OC de un plan de aceptación simple por atributo. Dada una regla de decisión, esta curva muestra la probabilidad de aceptar un lote (no rechazar H_o), en función de π*

Observe que existen infinitos valores posibles de β; sin embargo, al consumidor le interesan aquellos correspondientes a valores de $\pi \geq LTPD$. Entre ellos, el valor máximo de β se alcanza cuando $\pi = LTPD$.
De manera análoga, el productor enfrenta infinitos valores posibles de la probabilidad de rechazo de un lote *bueno* ($\pi \leq AQL$), pero el valor máximo, α, se alcanza cuando $\pi = AQL$.

Con lo visto hasta ahora, se pueden resolver las Actividades 17 y 18, Sección 9.6.

9.3.4 Prueba de hipótesis para la mediana de la población a través de una proporción

La mediana poblacional, que es una medida de tendencia central no sensible a valores atípicos como la media, en muchas ocasiones resulta el parámetro de interés.

Recuerde que:

- cuando la distribución de la variable de interés es simétrica, los valores de la mediana poblacional y de la media poblacional coinciden, de modo que una prueba para la mediana es también una prueba para la media,
- cuando la distribución de la variable es asimétrica, mediana y media no coinciden y las conclusiones de una prueba para la mediana son válidas solo para ese parámetro.

Como es sabido, la mediana no se define con una fórmula que involucra a todos los valores de la variable, como ocurre con la media. Este parámetro se define a partir de una probabilidad, es decir, la mediana de una población es aquel valor que verifica que la probabilidad de que la variable tome

valores menores (o mayores) a él es 0,5. Entonces, una hipótesis que involucra a la mediana ($\tilde{\mu}$) puede reformularse como una que involucre a una proporción.

La forma de proceder es la siguiente: Planteadas las hipótesis nula y alternativa sobre la mediana,

$H_o: \tilde{\mu} = \tilde{\mu}_o \quad H_a: \tilde{\mu} \neq \tilde{\mu}_o,$

se define una nueva variable Y':

$$Y' = \begin{cases} 1 & \text{si } Y > \tilde{\mu}_o \\ 0 & \text{en caso contrario} \end{cases}$$

Y' se comporta según el modelo Bernoulli: $Y' \sim Be(\pi)$, donde π es la proporción de valores de la población que son mayores que $\tilde{\mu}_o$.

Dada la variable de interés Y, cuya mediana ($\tilde{\mu}$) es el parámetro de interés, $\tilde{\mu} = \tilde{\mu}_o \iff \pi = 0,5$, y por lo tanto, las hipótesis planteadas originalmente pueden ser reemplazadas por: $H_o: \pi = 0,5$ $H_a: \pi \neq 0,5$.

A partir de esto se resuelve esta nueva prueba hipótesis como ya se presentó en la Sección 9.3.3 y se concluye sobre π (proporción de valores mayores que $\tilde{\mu}_o$ en la población) y, consecuentemente sobre $\tilde{\mu}$ (mediana poblacional).

Note que, para resolver esta prueba, solo se necesita saber cuántas observaciones superaron el valor $\tilde{\mu}_o$ en la muestra. Es decir que el estadístico de prueba es C: Cantidad de unidades de la muestra que presentan valor de la variable superior a $\tilde{\mu}_o$ y su distribución de probabilidades bajo H_o es $C \sim Bi(n; 0,50)$.

Para el caso de pruebas unilaterales, las hipótesis que se plantean son las siguientes:

$H_o: \tilde{\mu} = \tilde{\mu}_o \quad H_a: \tilde{\mu} > \tilde{\mu}_o$, lo que equivale a plantear $H_o: \pi = 0,5 \quad H_a: \pi > 0,5$ (prueba unilateral a la derecha).

$H_o: \tilde{\mu} = \tilde{\mu}_o \quad H_a: \tilde{\mu} < \tilde{\mu}_o$, lo que equivale a plantear $H_o: \pi = 0,5 \quad H_a: \pi < 0,5$ (prueba unilateral a la izquierda).



En la definición de la variable Y', de π y del estadístico C, siempre se consideran los valores mayores que $\tilde{\mu}_o$ independientemente del tipo de prueba que se realice (bilateral, unilateral a la derecha o a la izquierda).

El procedimiento descripto para contrastar hipótesis sobre la mediana poblacional, basado en la proporción de observaciones mayores que un valor hipotético, $\tilde{\mu}_o$, se conoce como **prueba del signo**.

Esta es una prueba que no impone requerimientos sobre la forma de la distribución de la variable de interés y se basa exclusivamente en el conteo de signos (valores mayores o menores que $\tilde{\mu}_o$).

Problema 1 (pág. 7): *Suponga que en el caso de la empresa metalúrgica, se desea evaluar, a partir de la información de una muestra de 120 barras, si la mediana de las longitudes es superior a 250 mm. En la muestra, 60 barras tenían longitud superior a 250 mm (Tabla 2.15).*

Se plantea en primer lugar:

$H_o: \tilde{\mu} = 250$

$H_a: \tilde{\mu} > 250$

Luego se define:

$$Y' = \begin{cases} 1 & si\ Y > 250 \\ 0 & en\ caso\ contrario \end{cases}$$

donde Y : longitud, en mm.

π*: proporción de barras con longitud superior a 250 mm.*

Se reformulan las hipótesis:

$H_o: \pi = 0,5$

$H_a: \pi > 0,5$

Se define $\alpha = 0,1$ *en este caso.*

Estadístico de prueba, C: Cantidad de barras de la muestra que tienen longitud superior a 250 mm.

Bajo H_o, $C \sim Bi(120; 0,5)$.

El valor observado del estadístico en la muestra resultó $c_{obs} = 60$.

valor p $= P(C \geq 60 | \pi = 0,50) = 0,536\ (> 0,1)$

La decisión es no rechazar H_o. *Es decir que la muestra no presenta evidencia suficiente para afirmar que la mediana de las longitudes de las barras de la población es mayor a 250 mm.*

En la Figura 9.23 se representa la distribución del estadístico de prueba bajo H_o *y el valor p obtenido.*

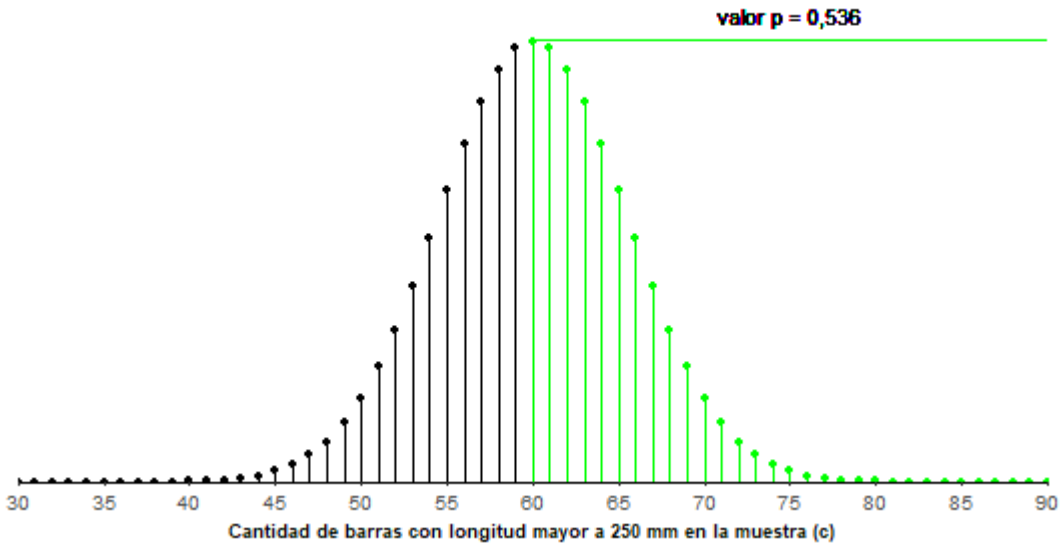


Figura 9.23 *Valor p (representado por la suma de las alturas de los bastones en verde). Prueba* $H_o: \tilde{\mu} = 250$ *vs* $H_a: \tilde{\mu} > 250$. *Distribución* $Bi(120; 0,50)$
(en la figura solo se representan valores del estadístico entre 30 y 90)

Problema 8 (pág. 8): *En la empresa interesada en adquirir los dispositivos electrónicos sospechan que la mediana del tiempo hasta la falla de los mismos es menor que 300 h. En la muestra de n = 45 dispositivos analizados, observaron que en 19 de ellos, el tiempo hasta la falla superaba las 300 h (Tomado a partir de la planilla de volcado que se encuentra en el archivo https://sites.google.com/view/libro-estadistica.*
Se plantea en primer lugar:
$H_o: \tilde{\mu} = 300$
$H_a: \tilde{\mu} < 300$
Luego se define:

$$Y' = \begin{cases} 1 & si\ Y > 300 \\ 0 & en\ caso\ contrario \end{cases}$$

donde Y : tiempo hasta la falla, en horas.
π: proporción de dispositivos con tiempo hasta la falla superior a 300 h.
Se reformulan las hipótesis:
$H_o: \pi = 0,5$
$H_a: \pi < 0,5$
Se define $\alpha = 0,05$ en este caso.
Estadístico de prueba, C: Cantidad de dispositivos de la muestra con tiempo hasta la falla mayor que 300 h.
Bajo H_o, $C \sim Bi(45; 0,5)$.
Dado el nivel de significación, el valor crítico resulta $c_c = 16$. Observe que $P(C \leq 16) = 0,036$ y que $P(C \leq 17) = 0,068$, de modo que se elige como valor crítico al 16, ya que con 17 se superaría el valor máximo pretendido para α.
La regla de decisión es: rechazar H_o en valor de H_a si $c_{obs} \leq 16$.
En este caso, el valor observado del estadístico en la muestra resultó $c_{obs} = 19$, por lo cual la decisión es no rechazar H_o.
Es decir, no hay suficiente evidencia para afirmar que la mediana de los tiempos hasta la falla de los dispositivos de ese tipo es menor que 300 h.
En la Figura 9.24 se representa la distribución del estadístico de prueba bajo H_o, junto con los valores crítico y observado.

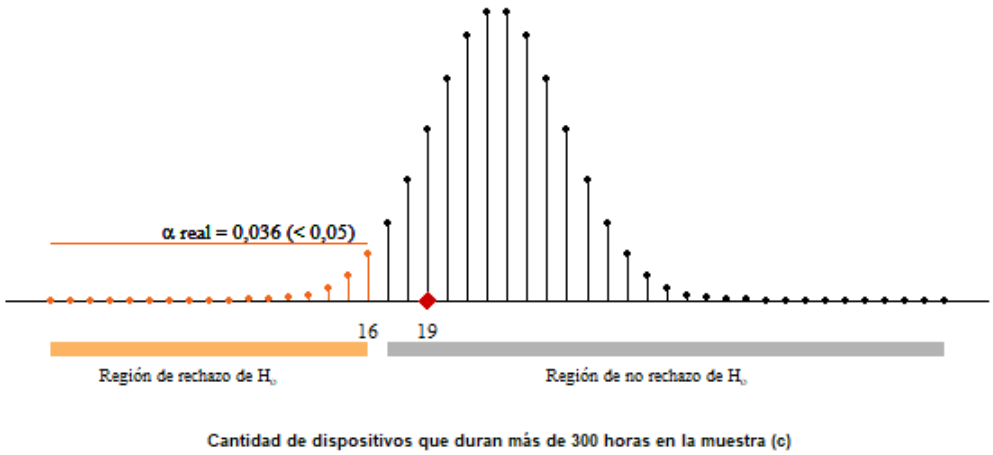


Figura 9.24 *Regiones de rechazo y de no rechazo de H_o. Prueba $H_o: \tilde{\mu} = 300$ vs $H_a: \tilde{\mu} < 300$. Distribución $Bi(45; 0,50)$. El valor observado en la muestra se representa mediante un rombo rojo. La suma de las alturas de los bastones en naranja representa el valor de la máxima probabilidad de error de tipo I, α real = 0,036*

Si bien en el Capítulo 8 no se presenta el intervalo de confianza para la mediana, existen métodos para obtenerlos.
Análogamente a lo visto para la media poblacional en la página 449, un intervalo de $(1-\alpha)\,100\,\%$ de confianza para la mediana poblacional, $\tilde{\mu}$, está conformado por todos los valores de $\tilde{\mu}_o$ para los cuales no se rechaza la hipótesis $H_o: \tilde{\mu} = \tilde{\mu}_o$ en favor de la alternativa correspondiente, a un nivel de significación α, a partir del valor observado en la muestra.



Como se mencionó en la página 482, cuando la distribución de la variable de interés es simétrica, la media y la mediana poblacionales coinciden. En ese caso, si bien las conclusiones respecto de uno de los parámetros son válidas para el otro, es importante destacar que las pruebas de hipótesis correspondientes presentan diferencias relevantes:

- **Prueba para la media**: resulta más adecuada porque utiliza toda la información de la muestra y, por lo tanto, es más potente. No obstante, puede verse seriamente afectada por la presencia de valores atípicos, especialmente cuando el tamaño muestral es pequeño.
- **Prueba para la mediana**: no se ve influida por valores atípicos ni requiere supuestos estrictos de normalidad, aunque suele ser menos potente cuando la normalidad es razonable.

Cuando la distribución es asimétrica, la media y la mediana difieren, por lo que la prueba de hipótesis debe formularse respecto del parámetro que sea de interés para el problema. En particular, si el interés está en la media, conviene aplicar métodos robustos [4] o trabajar con tamaños muestrales mayores.

Cabe destacar que la prueba para la mediana puede aplicarse también a variables medidas en escala ordinal, lo cual constituye una ventaja respecto de la prueba para la media, que se utiliza únicamente con variables medidas en escala de intervalo o de razón.



Con lo visto hasta ahora, se pueden resolver las Actividades 19 y 20, Sección 9.6.

9.4 Pruebas no paramétricas. Pruebas de bondad de ajuste

Entre las pruebas no paramétricas se encuentran las que analizan si una población tiene una distribución teórica específica, las cuales se denominan **pruebas de bondad de ajuste**. Estas pruebas constituyen una herramienta esencial en el análisis estadístico, ya que permiten evaluar la pertinencia de modelos teóricos y respaldar decisiones basadas en datos de manera objetiva y sistemática.

Las hipótesis que se plantean en una prueba de bondad de ajuste son las siguientes:

H_o : La distribución de la variable es una distribución específica.

H_a : La distribución de la variable es diferente de la postulada en H_o.



Note que el signo "=" está en H_o y en dicha hipótesis se describe un único escenario; mientras que en H_a se describen infinitos escenarios.

[4] Los métodos robustos son técnicas que buscan que las estimaciones y las pruebas sean menos sensibles a valores atípicos o distribuciones asimétricas. En este texto no se desarrollan dichos métodos.

Como en toda prueba de hipótesis, las hipótesis nula y alternativa se plantean para la población y se ponen a prueba a partir de información muestral. El análisis consiste en evaluar si la evidencia proporcionada por la muestra es suficiente para rechazar H_o en favor de H_a, dado un nivel de significación previamente fijado.

El principio fundamental de las pruebas de bondad de ajuste se basa en la comparación entre las frecuencias observadas y las frecuencias esperadas. Las frecuencias observadas provienen de los datos recolectados, mientras que las frecuencias esperadas se calculan a partir del modelo teórico planteado en H_o. Si las diferencias entre ambas resultan suficientemente grandes y no pueden atribuirse al azar, se considera que existe evidencia suficiente para rechazar la hipótesis nula; es decir, para concluir que el modelo propuesto no es adecuado, con un determinado nivel de significación.

Si la evidencia no resulta suficiente, no se puede rechazar la hipótesis nula, en la que se afirma que el modelo propuesto es adecuado. Sin embargo, es importante destacar que no rechazar H_o no implica que esta sea verdadera, ya que siempre existe la posibilidad de estar cometiendo un error de tipo II al tomar la decisión.

Una particularidad de las pruebas de bondad de ajuste es que, generalmente, lo que se pretende probar en la práctica es lo postulado en la hipótesis nula y no en la alternativa. Es por esto que, en este tipo de pruebas toma mayor relevancia la probabilidad de error de tipo II, β, cuyo valor es prácticamente imposible de determinar. Lo que siempre puede afirmarse en toda prueba de hipótesis es que, para un mismo nivel de significación, α, a mayor tamaño de muestra, menor es el riesgo de error de tipo II, β. Entonces, el tamaño de muestra es muy importante a la hora de concluir. Con pocos datos, sería muy arriesgado, al no rechazar H_o, afirmar que la variable se comporta según el modelo planteado. Por el contrario, con un tamaño de muestra grande, si bien nunca se puede asegurar que el modelo propuesto en H_o es el adecuado, no sería tan riesgoso suponerlo.

Existen distintos métodos para realizar pruebas de ajuste. Se pueden mencionar las pruebas de Chi-cuadrado, de Shapiro-Wilk, de Anderson–Darling y de Kolmogorov-Smirnov, entre otras alternativas disponibles en la literatura.

En cada una de ellas se define un estadístico de prueba y su distribución de probabilidades bajo la hipótesis nula. A partir de esta distribución y de la información muestral, se calcula el valor observado del estadístico y el correspondiente valor p, para luego decidir si se rechaza o no la hipótesis nula. En este libro no se desarrollan los estadísticos ni sus distribuciones; la decisión se toma a partir del cálculo del valor p con ayuda de un software, como por ejemplo, `R`.

En muchas situaciones, resulta de interés evaluar si la variable en estudio se comporta según el modelo Normal, ya sea porque es parte del objetivo del estudio estadístico o bien para decidir la herramienta de análisis inferencial a aplicar.

Las hipótesis nula y alternativa que se plantean son las siguientes:

$H_o: Y \sim N(\mu, \sigma)$

$H_a: Y$ tiene otra distribución diferente a la postulada en H_o.

Los valores de los parámetros pueden conocerse previamente (caso muy poco común). Si no se conocen, el software utiliza los propios valores de la muestra para estimarlos.

Luego, con los datos observados, el software obtiene el valor p y se puede concluir como en cualquier prueba de hipótesis.

En este texto se aplica la propuesta de Anderson-Darling para decidir sobre la hipótesis nula mencionada.



¿Por qué no es recomendable realizar pruebas de bondad de ajuste cuando el tamaño de muestra es pequeño?

El siguiente ejemplo se presenta para advertir sobre los riesgos de llevar a cabo pruebas de bondad de ajuste cuando el tamaño muestral es pequeño.

Ejemplo 9.1 Se consideran dos variables aleatorias: $X \sim N(100; 10)$ e $Y \sim \text{Exp}(0,3)$. De cada población se extrae una muestra de tamaño $n = 10$, otra de $n = 50$ y otra de $n = 250$.
Con la información de las seis muestras mencionadas se construyen diagramas de probabilidad Normal (Figuras 9.25 y 9.26) y se llevan a cabo pruebas de normalidad de Anderson-Darling.
Adicionalmente, para las tres muestras de la población con distribución Exponencial se construyen diagramas de probabilidad Exponencial (Figura 9.27) y se realizan pruebas de Anderson-Darling para bondad de ajuste Exponencial.

Las hipótesis nula y alternativa que se postulan al llevar a cabo las **pruebas de normalidad de Anderson-Darling** son las siguientes:

- Para la variable X (distribución Normal):

  $Ho : X \sim N(\mu, \sigma)$
  $H_a : X$; no se comporta según la distribución Normal

- Para la variable Y (distribución Exponencial):

  $H_o : Y \sim N(\mu, \sigma)$
  $H_a : Y$; no se comporta según la distribución Normal

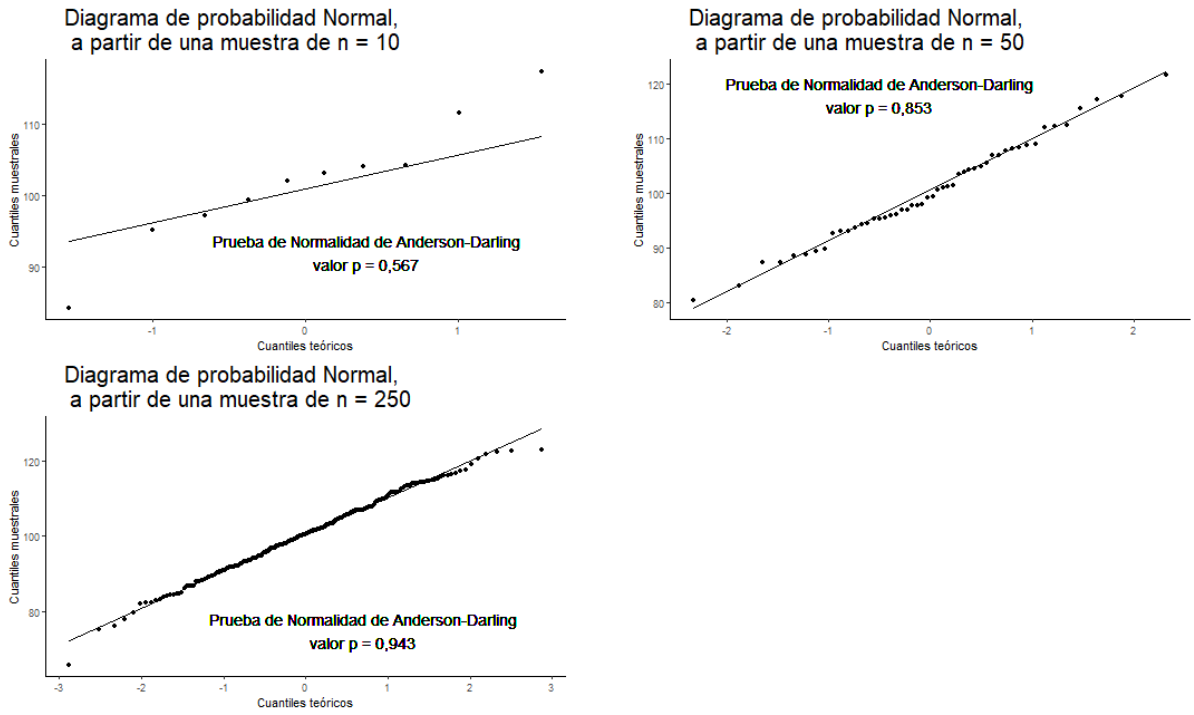


Figura 9.25 *Diagramas de probabilidad Normal para una muestra de n = 10, otra de n = 50 y otra de n = 250, tomadas de una población* $N(100; 10)$. *Valor p de la prueba de Anderson-Darling correspondiente*

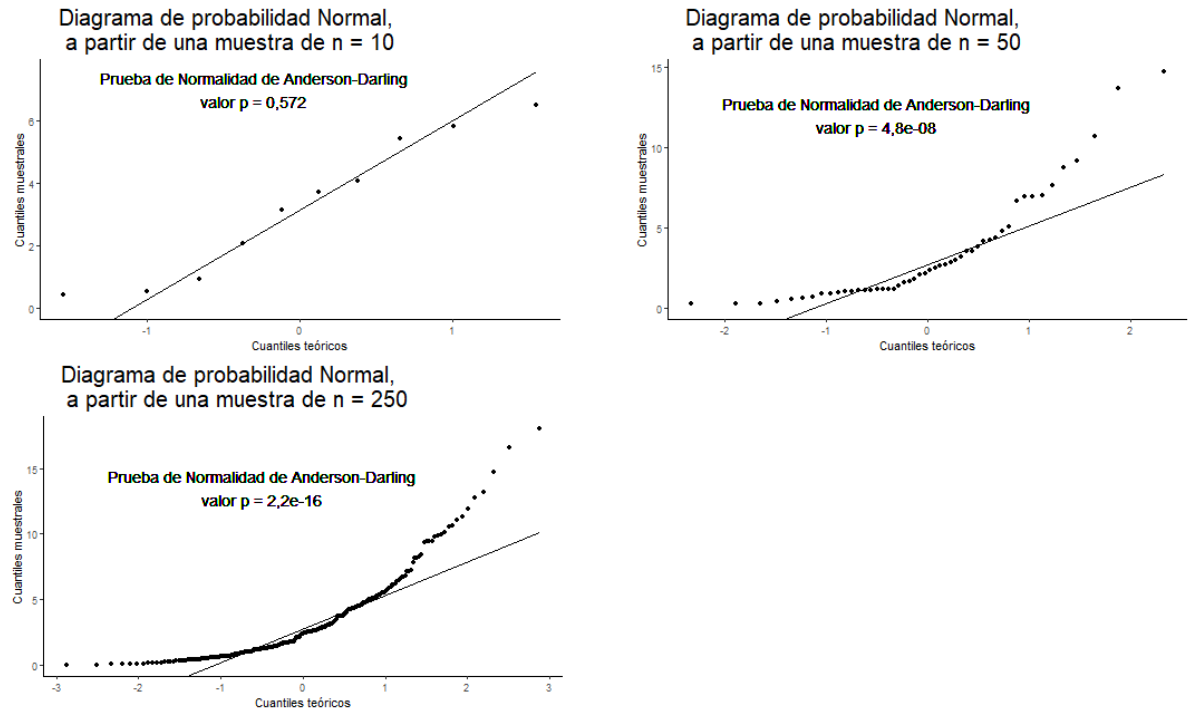


Figura 9.26 *Diagramas de probabilidad Normal para una muestra de n = 10, otra de n = 50 y otra de n = 250, tomadas de una población* $Exp(0,3)$. *Valor p de la prueba de Anderson-Darling correspondiente*

Dado que para la variable X la hipótesis nula es verdadera, el valor p obtenido en cada caso conduce **correctamente** al no rechazo de H_o (Figura 9.25).

Para la variable Y, en cambio, la hipótesis nula es falsa.

- para $n = 10$, el valor p conduce al no rechazo de H_o (error de tipo II).
- para $n = 50$ y $n = 250$, el valor p conduce al rechazo de H_o (decisión correcta, Figura 9.26)

Este ejemplo pone en evidencia que cuando el tamaño de muestra es pequeño,la prueba no tiene la potencia suficiente para detectar la falta de normalidad. No obstante, es importante recordar que aumentar el tamaño muestral reduce pero no elimina el riesgo de error tipo II.

Las hipótesis nula y alternativa que se postulan al llevar a cabo las **pruebas de bondad de ajuste exponencial de Anderson-Darling** son las siguientes:

$H_o : Y \sim Exp(\alpha)$

$H_a : Y$ no sigue la distribución Exponencial

Dado que para la variable Y la hipótesis nula es verdadera, el valor p obtenido en cada caso conduce **correctamente** al no rechazo de H_o (Figura 9.27).

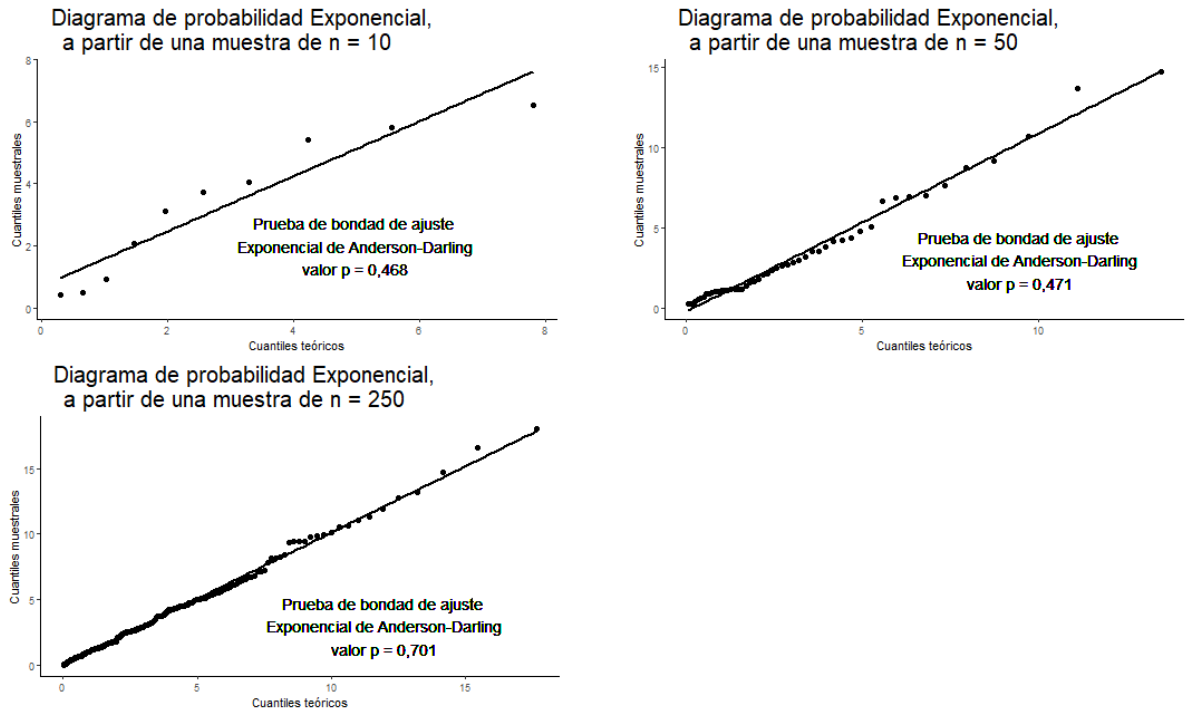


Figura 9.27 *Diagramas de probabilidad Exponencial para una muestra de n = 10, otra de n = 50 y otra de n = 250, tomadas de una población* $Exp(0,3)$. *Valor p de la prueba de Anderson-Darling correspondiente*

La hipótesis nula es verdadera. El valor p de cada una de las pruebas realizadas conduce al no rechazo de H_o (Figura 9.27) y por lo tanto, la decisión es correcta.

Observe que la muestra de tamaño $n = 10$ extraída de la distribución Exponencial conduce simultáneamente a:

- No rechazar $H_o : Y \sim \text{Exp}(\alpha)$ (decisión correcta)
- No rechazar $H_o : Y \sim N(\mu, \sigma)$ (decisión errónea)

Esto demuestra que con muestras pequeñas, diferentes distribuciones pueden resultar indistinguibles, lo que subraya la limitación fundamental de las pruebas de bondad de ajuste en estos escenarios.

En todas las pruebas llevadas a cabo en el ejemplo anterior se puso a prueba el modelo (Normal o Exponencial) suponiendo que los parámetros matemáticos de los mismos eran desconocidos al momento de seleccionar las muestras, situación que es la que ocurre más frecuentemente en la realidad.

Tal como se expuso en la Sección 8.5 del capítulo anterior, la resolución de problemas estadísticos requiere la combinación de herramientas de análisis descriptivo e inferencial.
Las pruebas de bondad de ajuste, presentadas en este capítulo, permiten analizar inferencialmente si la variable de interés se comporta de acuerdo con una distribución determinada, lo cual resulta de utilidad en dos situaciones:

- cuando dicho análisis constituye un objetivo del estudio estadístico;
- cuando es un requisito para la aplicación de otras herramientas inferenciales.

Problema 1 (pág. 7): *A partir del análisis descriptivo de las longitudes de las barras de acero seleccionadas (Capítulo 2) y de la inspección del diagrama de probabilidad Normal que se presenta en la Figura 9.28, se observa que los datos parecen ajustarse a una distribución Normal. Esta suposición se evaluó mediante la prueba de Anderson–Darling. El valor p obtenido es lo suficientemente alto como para no rechazar la hipótesis nula* $H_o : Y \sim N(\mu, \sigma)$*, donde Y representa la longitud. Por lo tanto, no se descarta que las longitudes de las barras se comporten según el modelo Normal.*

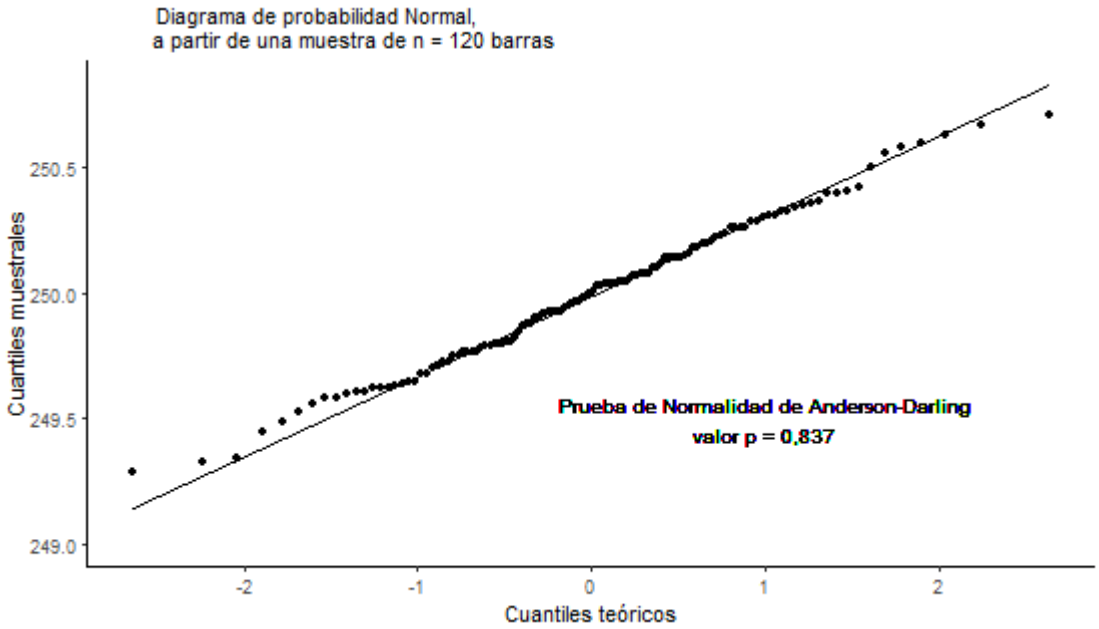


Figura 9.28 *Diagrama de probabilidad Normal construido con las longitudes de las n = 120 barras de acero seleccionadas y valor p de la prueba de Anderson-Darling correspondiente*

Problema 8 (pág. 8): *Para los tiempos hasta la falla de los n = 45 dispositivos electrónicos analizados, el diagrama de probabilidad Normal que se presenta en la Figura 9.29 muestra un patrón curvo, con puntos que se alejan de la recta teórica. Esto sugiere que la muestra no proviene de una población con distribución Normal. La prueba de Normalidad de Anderson–Darling, confirma esta sospecha arrojando un valor p bajo, que permite rechazar la hipótesis nula $H_o : Y \sim N(\mu, \sigma)$, donde Y representa al tiempo hasta la falla.*

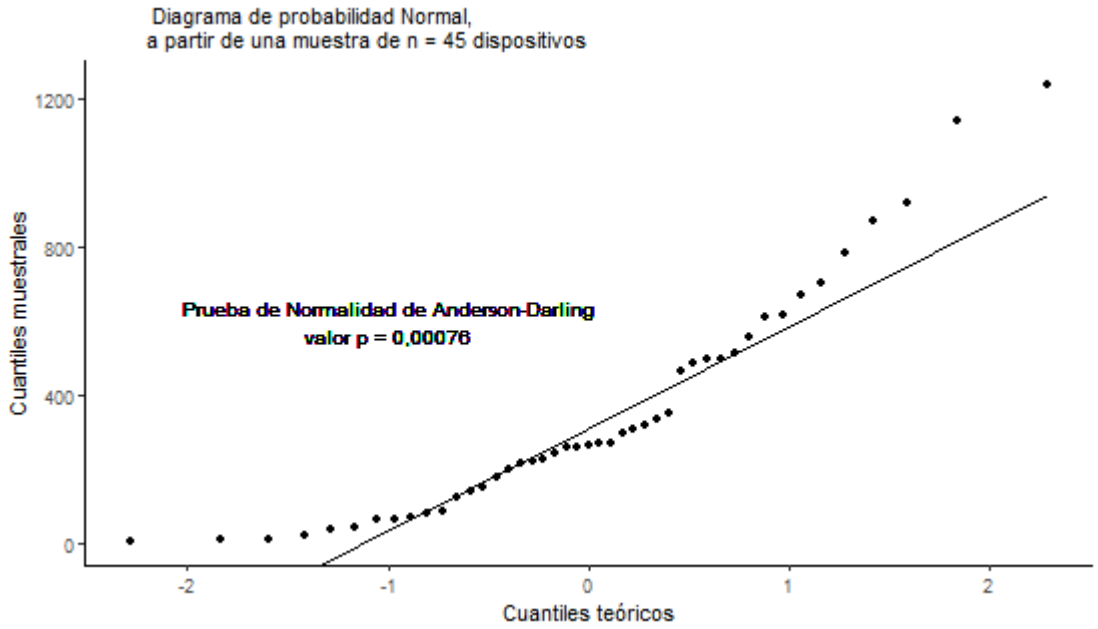


Figura 9.29 *Diagrama de probabilidad Normal construido con los tiempos hasta la falla de n = 45 dispositivos electrónicos analizados y valor p de la prueba de Anderson-Darling correspondiente*

Con la misma muestra se construyó un diagrama de probabilidad Exponencial y se aplicó la prueba de bondad de ajuste Exponencial de Anderson-Darling. El diagrama y el valor p de la prueba se presentan en la Figura 9.30.
El comportamiento de los puntos en el diagrama difiere notablemente del caso anterior, mostrando un ajuste más cercano a la recta teórica. El valor p obtenido es lo suficientemente alto como para no rechazar la hipótesis nula $H_o : Y \sim Exp(\alpha)$.

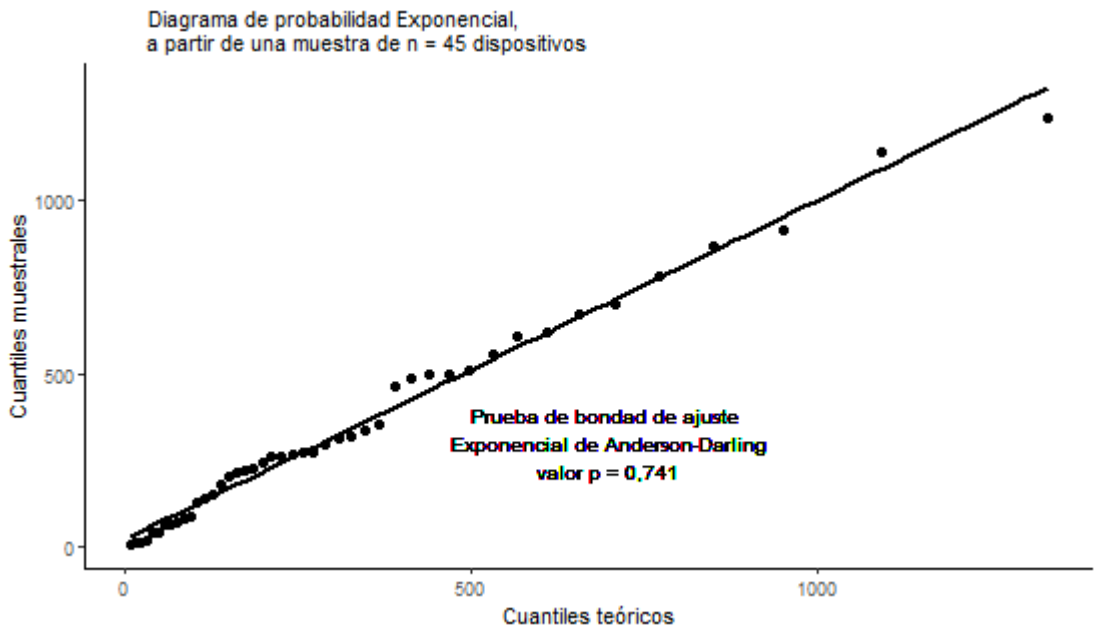


Figura 9.30 *Diagrama de probabilidad Exponencial construido con los tiempos hasta la falla de n = 45 dispositivos electrónicos analizados y valor p de la prueba de Anderson-Darling correspondiente*

Con lo visto hasta ahora, se pueden resolver las Actividades 21 a 34, Sección 9.6.

9.5 Síntesis

Además de los intervalos de confianza, que permiten realizar inferencias sobre parámetros, existe otra herramienta inferencial de gran importancia: las pruebas de hipótesis. Estas se utilizan principalmente cuando se desea contrastar una afirmación formulada acerca del valor de un parámetro de interés (pruebas paramétricas) y en otros casos, cuando las afirmaciones se refieren al modelo o distribución que sigue una variable, la existencia de relación entre dos variables u otras características de interés, etc. (pruebas no paramétricas).

Si bien el procedimiento de las pruebas de hipótesis difiere del utilizado en la construcción de los intervalos de confianza, ambos se apoyan en los mismos principios fundamentales. En particular, es esencial definir un estadístico adecuado y conocer su distribución muestral, la cual se determina bajo el supuesto de que la hipótesis nula es verdadera, para establecer márgenes de error, niveles de riesgo y criterios de decisión basados en la información observada.

En una prueba de hipótesis se formulan inicialmente las hipótesis nula y alternativa y luego, a partir de la información muestral, se decide si se rechaza o no la hipótesis nula. En caso de rechazarse, se prueba lo postulado en la hipótesis alternativa. Toda decisión implica un riesgo de error: si se rechaza la hipótesis nula cuando en realidad es verdadera, se comete un error de tipo I, cuya probabilidad suele fijarse previamente a la toma de la muestra. Si, en cambio, no se rechaza la hipótesis nula cuando es falsa, se incurre en un error de tipo II, cuya probabilidad puede ser elevada o incluso desconocida en algunos casos.

La conclusión se puede tomar utilizando los procedimientos propios de las pruebas de hipótesis y en el caso de pruebas paramétricas, también a través de un intervalo de confianza. Los procedimientos propios son dos: el criterio basado en una región de rechazo y el basado en el valor p.

Las herramientas inferenciales suponen ciertos requerimientos que deben verificarse previamente. En este sentido, las pruebas de bondad de ajuste, orientadas a evaluar el modelo de distribución de una variable, contribuyen a la verificación de algunos requerimientos. Además, resultan especialmente útiles cuando el análisis de la distribución forma parte del objetivo del estudio estadístico.

En la Tabla 9.3 se presentan las expresiones de los estadísticos de prueba considerados en este texto junto con sus respectivas distribuciones de probabilidades, suponiendo cierta la hipótesis nula.

Tabla 9.3 *Pruebas de hipótesis respecto de μ, σ^2, π y $\tilde{\mu}_o$. Expresión del estadístico de prueba y su distribución de probabilidades bajo H_o*

Parámetro	**Estadístico de prueba**	**Distribución de probabilidades bajo H_o**
μ **(se conoce σ)**	$\bar{Y}$ $Z = \frac{\bar{Y}-\mu_o}{\frac{\sigma}{\sqrt{n}}}$	$\bar{Y} \sim N(\mu_o, \frac{\sigma}{\sqrt{n}})$ $Z \sim N(0,1)$
μ **(no se conoce σ)**	$T = \frac{\bar{Y}-\mu_o}{\frac{S}{\sqrt{n}}}$	$T \sim t_{n-1}$
σ^2	$\chi^2 = \frac{(n-1)S^2}{\sigma_o^2}$	χ^2_{n-1}
π	C : cantidad de unidades de la muestra que cumplen cierta condición	$C \sim Bi(n, \pi_o)$
$\tilde{\mu}$ (se plantea una hipótesis respecto de π [5]	C : cantidad de unidades de la muestra con valor mayor que $\tilde{\mu}_o$	$C \sim Bi(n; 0,50)$

Observaciones:

- Los supuestos y condiciones de validez de los procedimientos presentados son los mismos que los indicados al pie de la Tabla 8.3 (página 424), para cada parámetro.
- Cuando se desea poner a prueba hipótesis sobre el desvío estándar poblacional, σ, estas deben formularse en términos de la variancia poblacional, σ^2.
- En el caso de la mediana, el procedimiento consiste en poner a prueba una hipótesis sobre la proporción de unidades de la población cuyos valores son mayores que el valor postulado para la mediana en la hipótesis nula.

[5] Se trata de la proporción de unidades de la población con valor mayor que el valor postulado para la mediana en la hipótesis nula.

9.6 Actividades propuestas

1. Reconsidere la situación planteada en las Actividades 1 y 2 (Capítulo 8, página 426), referidas a una empresa metalúrgica dedicada a la fabricación de recipientes sometidos a presión. La característica de interés es la resistencia al estallamiento (expresada en libras por pulgadas cuadradas) de las soldaduras utilizadas en dichos recipientes.
   Para un determinado tipo de recipiente se exige que la resistencia promedio de las soldaduras sea superior a 6800 lb/pulg^2. Con el fin de decidir si un cierto tipo de soldadura es apto para esta aplicación, se realizaron 10 soldaduras de ese tipo y se midió la resistencia al estallamiento de cada una, obteniéndose los siguientes valores:

   7200 - 6100 - 7300 - 7300 - 8000 - 7400 - 7300 - 7300 - 8000 - 6700

   Algunas medidas de resumen obtenidas a partir de estos datos se presentan en el enunciado de la Actividad 1.
   Suponga que las 10 observaciones se pueden considerar una muestra aleatoria de una población con distribución Normal y que se conoce que la desviación estándar de la resistencia al estallamiento para ese tipo de soldaduras, σ, es de 500 lb/pulg^2.
   a) Defina población y variable de interés. Indique el objetivo del estudio.
   b) Defina el parámetro de interés y plantee las hipótesis nula y alternativa correspondientes.
   c) Escriba, en términos del problema, en qué casos se cometería error de tipo I y de tipo II.
   d) Concluya a partir de la obtención de la región de rechazo (5 pasos). Trabaje con un nivel de significación del 5 %.
   e) Compare su conclusión con la obtenida en la Actividad 2. ¿Qué información adicional le brinda el intervalo de confianza?
   f) Antes de tomar la muestra, ¿cómo se interpreta el valor 0,05 correspondiente al nivel de significación?

2. Cuando está operando adecuadamente, una planta química tiene una media de producción diaria de 720 toneladas, con una desviación estándar de 15 toneladas. Se sospecha que en el último tiempo esta media pudo haber disminuido, por lo que se decidió registrar la producción en una muestra aleatoria de 30 días. Para analizar los datos se obtuvo lo siguiente:

```
One Sample z-test
z = -1.801, Std. Dev. Population = 15, p-value = 0.035
alternative hypothesis: true mean is less than 720
95 percent confidence interval:
-Inf 719.571
sample estimates:
mean of x
715.067
```

   a) Plantee el problema (mencione objetivo, población y población estadística, unidad elemental, variable en estudio y parámetro de interés).
   b) Establezca las hipótesis nula y alternativa asociadas a esta problemática, e interprete los posibles errores de tipo I y tipo II en el contexto del problema.
   c) ¿Cuál es la conclusión de la prueba de hipótesis? Obténgala luego de definir la región

de rechazo correspondiente y también a partir de analizar la información brindada por el intervalo de confianza.

d) ¿Se obtiene un resultado estadísticamente significativo?

3. Un torno automático produce un tipo de pieza cuya longitud (en centímetros) se distribuye según un modelo Normal con $\sigma = 0,1$. Cuando la herramienta de corte está correctamente ajustada, la longitud promedio de las piezas es de 2 cm; pero cuando la misma adopta una posición incorrecta, dicho promedio se altera (aumentando o disminuyendo), sin que se modifique la desviación estándar. En este último caso, es necesario corregir la posición de la herramienta. Una muestra de 10 piezas arrojó una longitud promedio de 2,04 cm.
   a) Plantee el problema (mencione objetivo, población y población estadística, unidad elemental, variable en estudio y parámetro de interés).
   b) Establezca las hipótesis nula y alternativa asociadas a esta problemática.
   c) Interprete los posibles errores de tipo I y tipo II en el contexto del problema.
   d) Si se establece en 0,1 la probabilidad de concluir erróneamente que el torno está desajustado, ¿cuál sería su conclusión? Determine la región de rechazo en función de los valores del estadístico $\bar{Y}$ y responda.
   e) Determine la región de rechazo en función de estadístico pivote y vuelva a realizar la prueba.
   f) Obtenga nuevamente las conclusiones a través de un intervalo de confianza. ¿Cuál es la información adicional que proporciona el mismo?
   g) Si el torno estuviera desajustado y la longitud promedio fuera de 1,95 cm, ¿cuál sería la probabilidad de no efectuar el ajuste?
   h) ¿Cuál debería ser el tamaño de la muestra si se desea reducir dicha probabilidad a 0,2?

4. Reconsidere las primeras tres actividades y realice lo siguiente:
   a) En la Actividad 1, calcule el valor *p* y concluya sobre las hipótesis planteadas.
   b) En la Actividad 2,
      1) concluya utilizando el valor *p* dado en la salida;
      2) indique cómo se obtuvo dicho valor (solo plantee formalmente).
   c) Si en la Actividad 3 se obtuviera el valor *p* de la prueba, ¿este resultaría menor o mayor que 0,10? Justifique su respuesta conceptualmente y luego, compruébela realizando el cálculo correspondiente.

5. Reconsidere la Actividad 4 (Capítulo 8, página 427), en la cual se quiere probar la sospecha de que en una determinada zona (denominada Zona 2), la concentración media de dióxido de carbono en el aire (en partes por millón en volumen)es menor que 420. Para verificar dicha sospecha, se seleccionan aleatoriamente 20 puntos en esa zona y se mide en cada uno la concentración de dióxido de carbono, obteniendo un promedio de 378 ppmv y una desviación estándar de 80 ppmv.
   a) Defina población y variable. Mencione el parámetro de interés.
   b) Plantee las hipótesis nula y alternativa correspondientes.
   c) Indique en qué casos se cometería error de tipo II. Comente cómo se puede reducir el riesgo de que esto ocurra.
   d) Explicite la distribución de probabilidades del estadístico de prueba $\bar{Y}$ y del estadístico pivote que debe utilizarse en este caso, suponiendo cierta H_o. Justifique.

e) Para un nivel de significación de 0,02, ¿qué puede concluir en relación a la sospecha? Defina la región de rechazo y concluya. Informe sus conclusiones.
f) Calcule el valor *p* de la prueba y compruebe su conclusión.
g) Relacione la conclusión obtenida en los ítems anteriores, con la que surge del intervalo de confianza calculado para la Actividad 4 (Capítulo 8).

6. En la Actividad 5 (Capítulo 8, página 427), se presenta una situación en la cual los ingenieros de una industria automotriz están investigando el tiempo (en segundos) que demanda cierto proceso de ensamblado. Por experiencia, conocen que esa variable se comporta según el modelo Normal y sospechan que el tiempo promedio para ese proceso es mayor que 90 s.
Para verificar su sospecha, seleccionan una muestra aleatoria simple de 20 realizaciones del proceso y registran los tiempos de cada una de ellas, los cuales son los siguientes:

93, 90 , 97 , 90 , 93 , 91 , 96 , 94 , 91 , 91 , 88 , 93 , 95 , 91 , 89 , 92 , 87, 88 , 90 , 86

La siguiente es la salida de R, obtenida con los 20 tiempos:

```
t.test(tiempo, alternative = "greater", mu = 90, conf.level = 0.99)
One Sample t-test
t = 1.9042, df = 19, p-value = 0.03607
alternative hypothesis: true mean is greater than 90
99 percent confidence interval:
89.583       Inf
sample estimates:
mean of x
91.25
```

a) Defina población, variable y parámetro de interés.
b) Plantee las hipótesis nula y alternativa que permitan decidir sobre la sospecha de los ingenieros.
c) Indique, en términos del problema, en qué caso se estaría cometiendo error de tipo I y en qué caso, error de tipo II.
d) Alguien afirma que como el tiempo promedio de la muestra resultó 91,25 s, la sospecha de los ingenieros es acertada. Usted, ¿está de acuerdo? Justifique su respuesta y si esta es negativa, analice adecuadamente los datos y dé su conclusión.
e) Analice el intervalo de confianza dado y comente qué información adicional le brinda, en relación a la decisión tomada.

7. Reconsidere la Actividad 7 (Capítulo 8, página 428) referida a un proceso de producción de cierto tipo de cojinetes. Para estos cojinetes, el diámetro interior (en centímetros) debe ser, en promedio, igual a 3. Por experiencia los fabricantes conocen que el diámetro se comporta según el modelo Normal, pero sospechan que el promedio no es el pretendido y deciden llevar a cabo un estudio estadístico.
En el marco del mismo, seleccionan en forma aleatoria, 12 cojinetes y en cada uno miden su diámetro interior, obteniendo los siguientes valores:

$3,01 - 3,05 - 2,99 - 3,03 - 3,00 - 3,02 - 3,05 - 2,99 - 3,04 - 2,97 - 3,02 - 3,01$

Con R obtienen lo siguiente:

```
mean(cojinetes)
3,015
sd(cojinetes)
0.02504
```

a) Defina población, variable y parámetro de interés.
b) Indique un posible objetivo del estudio estadístico. Expréselo en función del parámetro definido en el ítem anterior.
c) Plantee las hipótesis nula y alternativa que permitan decidir si el proceso está centrado en el valor que se pretende.
d) Indique, en términos del problema, en qué caso se estaría cometiendo error de tipo I y en qué caso, error de tipo II.
e) Defina el estadístico de la prueba e informe su distribución de probabilidades bajo H_o. Justifique.
f) Defina las regiones de rechazo y no rechazo de la hipótesis nula, suponiendo un nivel de significación de 0,05.
g) Analice los datos obtenidos e informe sus conclusiones.
h) De haber un error en la decisión tomada en este caso, ¿qué tipo de error podría ser? Justifique.

8. Una empresa produce piezas metálicas, en las cuales se controla su longitud (en centímetros) como una característica de calidad importante. Dichas longitudes pueden considerarse normalmente distribuidas con valor medio 3 y desviación estándar 0,03.
Para un cliente en particular, la longitud de las piezas, en cm, debe satisfacer la especificación: $(3 \pm 0,1)$ cm y pone como condición que el porcentaje de piezas fuera de especificación sea a lo sumo 0,27 %.
Se conoce, por experiencia, que los problemas que pueden presentarse en el proceso de producción afectan el valor medio de las longitudes de las piezas, pero no su desviación estándar. En el caso que se presente esta situación, debe investigarse en profundidad si corresponde o no realizar tareas de ajuste para evitar problemas con el cliente.
Diariamente se toma una muestra aleatoria simple de 40 piezas que se envían al Laboratorio de Pruebas Físicas. Allí se miden las longitudes de las piezas y se obtiene la longitud promedio de la muestra. Si ese valor no pertenece al intervalo $(3 \pm 0,01)$ cm, se informa al Sector Producción para que verifique si estos cambios pueden afectar los requerimientos del cliente.

a) En caso de haber un cambio en el parámetro de interés, ¿hasta qué valor podrá aumentar o disminuir este parámetro sin que se supere el máximo porcentaje de piezas fuera de especificaciones admitido por el cliente?
b) Defina población, variable y parámetro de interés. Luego plantee el problema de decisión que se enfrenta diariamente en la empresa.
c) ¿Cuáles son las hipótesis nula y alternativa que se ponen a prueba en el estudio muestral que se realiza diariamente?
d) ¿Con qué estadístico se lleva a cabo la prueba? ¿Cuál es su distribución muestral? Justifique la respuesta y si existen requerimientos que deben verificarse, indique cuáles son y cómo evaluaría su cumplimiento.
e) ¿Qué errores se pueden cometer al llevar adelante la prueba? ¿Qué consecuencias tiene

para la empresa cada uno de estos errores?

f) ¿Cuál es el nivel de significación de la prueba?

g) Evalúe el riesgo máximo de no detectar un corrimiento (cambio) en el promedio de las longitudes que afecte negativamente al proceso (un cambio que haga que no se cumpla lo que pretende el cliente).

h) Suponga que alguien informa que el día lunes de esta semana, el valor *p* de la prueba fue 0,028.

1) ¿El resultado de ese día fue estadísticamente significativo? Justifique su respuesta.
2) ¿Existe la posibilidad de haber cometido un error al tomar la decisión? ¿De qué tipo?
3) ¿Se informó a Producción ese día? Justifique su respuesta.
4) En caso de responder afirmativamente al ítem anterior, ¿qué pasos deberá seguir Producción para saber si se cumplen los requerimientos del cliente?

i) Considere los dos intervalos presentados en el enunciado: $3 \pm 0,1$ y $3 \pm 0,01$. ¿Cuál es la diferencia entre ellos?

9. (Actividad adaptada de Hildebrand y Ott (1997) [15])

En una fábrica de piezas fundidas para camiones interesa la longitud de las mismas (medida en pulgadas), como el factor más importante en la calidad. Idealmente, la pieza debería medir 8,05", ya que luego se le da un acabado para que quede con una longitud de 8". Aun si el proceso trabaja adecuadamente, hay una variación sustancial en la longitud de las piezas, debido a los cambios en la temperatura exterior, la humedad y a las variaciones en la calidad del acero que se utiliza como materia prima. La experiencia indica que, cuando el proceso de fundición trabaja bien, la desviación estándar de las longitudes de las piezas fundidas es aproximadamente igual a 0,18".

Los problemas que suelen presentarse en el proceso son relativos a una longitud media incorrecta, antes que un incremento de la variabilidad. Por consiguiente, se propone obtener muestras diarias de 16 piezas, y si la media muestral está muy alejada de 8,05", detener el proceso de fundición y llevar a cabo un largo procedimiento de reajuste (costoso).

El director del proceso de fundición quiere establecer límites de 7,915" y 8,185" para la media muestral, pero el gerente del proceso de acabado prefiere límites de 8" y 8,1".

Ambos están de acuerdo en que si la longitud promedio poblacional se corre (aumenta o disminuye) en 0,10" o más, sería grave para el proceso.

a) Responda las siguientes preguntas:

1) ¿Cuál es el problema que da origen al estudio por muestreo?
2) ¿Cuáles son las hipótesis estadísticas en juego?
3) ¿Qué consecuencias están asociadas con las posibles decisiones que se tomen?
4) ¿Qué errores pueden cometerse?
5) ¿Qué reglas de decisión propone cada uno de los gerentes?
6) ¿Qué utilidad aporta conocer el mínimo corrimiento en la longitud promedio poblacional que sería grave para el proceso? Explique.
7) ¿Cómo se puede evaluar si las reglas propuestas por los gerentes son adecuadas? Realice el análisis que crea conveniente e informe sus conclusiones sobre las mismas.
8) ¿Debió verificar algún requerimiento para realizar los cálculos en el ítem anterior? Menciónelos y justifique por qué.

b) El Gerente General decide finalmente que puede admitirse:
- una detención innecesaria del proceso de fundición, como máximo el 2 % de las veces.
- que el sector acabado reciba partidas que no cumplen lo establecido para la longitud promedio, no más del 5 % de las veces.

¿Considera que alguna de las reglas propuestas por los gerentes cumple estas condiciones? Si no es así, ¿puede usted hacer alguna propuesta que las cumpla?

10. Reconsidere la Actividad 9 (Capítulo 8, página 429). En dicha actividad se analizó un proceso de producción en el que se implementaron modificaciones destinadas a reducir la variabilidad de los diámetros (en milímetros) de cierto tipo de válvulas especiales. La variancia del proceso antes de las modificaciones era de 16 mm^2 y se estableció que dichas modificaciones se considerarían efectivas si lograban una reducción de al menos un 40 % en el desvío estándar. Luego de implementar los cambios, se tomó una muestra aleatoria de 60 válvulas, obteniéndose un desvío estándar muestral de 2,5 mm. Además, no se encontró evidencia para descartar la normalidad de los diámetros.
    a) A partir de la situación planteada, defina nuevamente población, variable y parámetros de interés y exprese al objetivo del estudio estadístico en términos de los parámetros mencionados.
    b) ¿Considera que las modificaciones llevadas a cabo en el proceso lograron reducir la variabilidad de los diámetros de las válvulas? Plantee las hipótesis nula y alternativa que le permitan dar una respuesta a esa pregunta.
    c) Indique cuál es el estadístico pivote y su distribución de probabilidades bajo H_o. Justifique.
    d) Construya la región de rechazo de H_o para un nivel de significación del 2 %. Luego analice los datos e informe sus conclusiones.
    e) Los resultados fueron:
    - significativos y relevantes.
    - significativos, pero no relevantes.
    - significativos, pero no se pudo concluir si fueron relevantes o no.
    - ni significativos ni relevantes.

    Indique cuál es la opción que corresponde en este caso. Si necesita realizar otro análisis para responder, hágalo y justifique claramente su respuesta.

11. Reconsidere la Actividad 10 (Capítulo 8, página 429), relativa a los errores de medición de una distancia conocida de 500 metros con un equipo exacto. Interesa analizar si el equipo es, además, preciso.
    Suponga que el equipo se considera preciso si la desviación estándar de los errores de las mediciones resulta menor que 0,07 m. A partir de 15 mediciones de la distancia conocida, realizadas con dicho equipo, se obtuvo lo siguiente salida de `R`:

```
VarCI(errores,method = ''classic'')
(0.0019;0.0086)
VarCI(errores,method = ''classic'',sides = ''left'')
(0.0021,Inf)
VarCI(errores,method = ''classic'',sides = ''right'')
(0;0.0074)
```

a) Realice el planteo completo del problema.
b) Plantee las hipótesis nula y alternativa que le permitan decidir sobre la precisión de este equipo de medición.
c) Observe el intervalo de confianza apropiado y concluya. Justifique.
d) De acuerdo al intervalo elegido, informe el nivel de significación de la prueba realizada.
e) Si hubiera obtenido el valor p, ¿este sería menor o mayor que el nivel de significación mencionado en el ítem anterior? Justifique.

12. Reconsidere la Actividad 16 (Capítulo 8, página 432), en la cual se hace referencia a un fabricante asegura que la capacidad media (en amperios-hora) de cierto tipo de baterías que produce la compañía es de al menos 140. Un grupo para la defensa del consumidor desea poner a prueba esta afirmación y para ello mide la capacidad de 20 baterías de ese tipo seleccionadas al azar, obteniendo los siguientes resultados:

```
mean(capacidad)
138.47
sd(capacidad)
2.659

t.test(capacidad, alternative = "two.sided", mu = 140,  conf.level = 0.99)
t = -2.5734, df = 19, p-value = 0.01861
alternative hypothesis: true mean is not equal to 140
99 percent confidence interval:
136.769      140.171

t.test(capacidad, alternative = "greater", mu = 140,  conf.level = 0.99)
t = -2.5734, df = 19, p-value = 0.9907
alternative hypothesis: true mean is greater than 140
99 percent confidence interval:
136.960      Inf

t.test(capacidad, alternative = "less", mu = 140,  conf.level = 0.99)
t = -2.5734, df = 19, p-value = 0.009305
alternative hypothesis: true mean is less than 140
99 percent confidence interval:
-Inf         139.980
```

En el Capítulo 8 se presentan algunos gráficos obtenidos con la información de las baterías analizadas.

a) Plantee el problema. Defina población, variable y parámetros de interés.
b) Escriba las hipótesis nula y alternativa que permitan poner a prueba la afirmación del fabricante.
c) Indique, en contexto, en qué casos se cometería error de tipo I y en qué casos, error de tipo II, según las hipótesis planteadas en el ítem anterior.
d) Para llevar adelante la prueba, ¿se debe verificar algún requerimiento sobre la distribución de la variable? Comente sobre esto a partir del análisis de los gráficos que se presentan en

la salida.

e) Suponga que se admite un máximo de 0,01 para la probabilidad de error de tipo I. Concluya sobre las hipótesis planteadas.

f) ¿Considera que tiene información para afirmar que la desviación estándar de la capacidad de las baterías es inferior a 3 Ah? Plantee las hipótesis correspondientes y póngalas a prueba. Concluya.

13. Considere la Actividad 11 (Capítulo 8, página 430), relativa a una empresa productora de placas de madera. En dicha empresa, históricamente la proporción de placas buenas era 0,90; pero los profesionales a cargo del proceso sospechan que ese valor cambió. Con el objetivo de confirmar esa sospecha, seleccionaron una muestra aleatoria simple de 200 placas y encontraron que 182 placas eran buenas.

a) Defina población, variable y parámetro de interés.

b) Plantee las hipótesis nula y alternativa que permitan verificar la sospecha mencionada.

c) Informe cuál es el estadístico de la prueba y cuál es su distribución de probabilidades bajo H_o.

d) Lleve adelante la prueba de hipótesis correspondiente para un nivel de significación de 0,05. Informe sus conclusiones.

e) ¿Cómo utilizaría un intervalo de confianza para decidir sobre H_o en este caso? Explique y comente sobre la información adicional que brinda un intervalo (sea que se rechace o no se rechace H_o)

14. En un proceso de producción, la proporción de artículos que no conforman las especificaciones es, en la actualidad, igual a 0,05. Con el objetivo de reducir dicha proporción, se resolvió evaluar, a escala piloto, una serie de modificaciones en el proceso.
Con ese fin, se extrajo una muestra aleatoria de 300 artículos producidos luego de realizadas las modificaciones a escala piloto y se registraron 6 artículos fuera de especificación. El intervalo del 95 % de confianza para la proporción de artículos del proceso que no conforman las especificaciones resultó $IC_{95\%,\pi} = (0,0074; 0,043)$.

a) Plantee el objetivo del estudio. Mencione variable, población física y parámetros de interés.

b) El problema planteado, ¿es un problema de decisión estadística? Justifique su respuesta.

c) Plantee las hipótesis nula y alternativa correspondientes y defina el estadístico de prueba. ¿Cómo se distribuye este estadístico? Justifique.

d) ¿Puede considerarse, en función del intervalo obtenido, que con las modificaciones realizadas en el proceso se obtiene una reducción estadísticamente significativa en la proporción de artículos defectuosos del mismo? Justifique su respuesta.

e) Se ha realizado un estudio de costos de dichas modificaciones, y se resolvió que su implementación en la planta solo se justificaría si la proporción de artículos defectuosos se reduce como mínimo a la mitad. ¿Recomendaría implementar las modificaciones en la planta? Justifique su respuesta.

f) Resuelva la prueba de hipótesis planteada por la forma clásica y el valor *p* respetando el mismo nivel de significación que en el ítem d. Compare sus respuestas.

15. Reconsidere la Actividad 12 (Capítulo 8, página 430) relativa a una maderera minorista que recibe embarques de sus proveedores. Cada embarque contiene una gran cantidad de hojas de madera y se considera rentable si el porcentaje de hojas de segunda calidad es inferior a 15 %.

Para un embarque particular, un supervisor selecciona una muestra aleatoria de 144 hojas y encuentra que 18 hojas no pueden venderse como de primera calidad.
Se obtuvieron las siguientes salidas de R:

```
binom.test(x=18, n=144, p=0.15, alternative = "less", conf.level =0.95)
Exact binomial test
data:  18 and 144
number of successes = 18, number of trials = 144, p-value = 0.2388
alternative hypothesis: true probability of success is less than 0.15
95 percent confidence interval:
 0.0000       0.1797
sample estimates:
probability of success
0.125
```

a) Defina población, variable y parámetro de interés.
b) Plantee las hipótesis nula y alternativa que hagan referencia a si el embarque que acaba de llegar es rentable o no.
c) Analice los datos e informe sus conclusiones.
d) Elabore una regla de decisión, utilizando la forma clásica, que permita al supervisor decidir sobre las hipótesis planteadas en el ítem b, para cualquier embarque que llegue a la empresa, a partir de una muestra aleatoria de tamaño $n = 144$ hojas.

16. Considere la Actividad 13 (Capítulo 8, página 431). El fabricante de autopartes al que se hace referencia en dicha actividad utiliza un conjunto de pruebas para evaluar a sus productos y ha estipulado que todas las autopartes deben pasar las pruebas antes de ser enviadas al cliente. Si la proporción de autopartes que no pasan dichas pruebas supera 0,025, el fabricante considerará realizar modificaciones en el proceso de producción. En una muestra aleatoria de 500 autopartes, 15 fallan en alguna de las pruebas, obteniendo el siguiente intervalo de confianza para la proporción de autopartes que no pasan la prueba:

```
BinomCI(15, 500, method = "clopper-pearson")
 est     lwr.ci    upr.ci
0.03     0.017      0.049
BinomCI(15, 500, method = "clopper-pearson", sides = "left")
 est     lwr.ci    upr.ci
0.03     0.019         1
BinomCI(15, 500, method = "clopper-pearson", sides = "right")
 est     lwr.ci    upr.ci
0.03       0        0.046
```

a) Realice el planteo del problema.
b) Plantee las hipótesis nula y alternativa que le permitan decidir si se deben realizar modificaciones en el proceso de producción.
c) Tome el intervalo de confianza que crea conveniente y decida. Exprese sus conclusiones en términos del problema.
d) Indique si las siguientes afirmaciones son verdaderas o falsas. Justifique.

- De acuerdo al intervalo elegido, se puede considerar que el nivel de significación de esta prueba es 0,05.
- El resultado fue estadísticamente significativo.
- Si se hubiera obtenido el valor p, este resultaría menor que el nivel de significación de la prueba.

17. Un cliente recibe habitualmente lotes de su proveedor, los cuales tienen una gran cantidad de piezas y pretende que estos lotes tengan como máximo el 1 % de piezas defectuosas. En la recepción selecciona una muestra aleatoria de 50 piezas de cada lote y lo acepta si el número de piezas defectuosas es a lo sumo 2. En caso contrario, se lo devuelve al proveedor.
    a) Si los lotes contienen un 1 % de piezas defectuosas (cumplen con lo pretendido), ¿cuál es la proporción de lotes que serán devueltos?
    b) ¿Qué nombre recibe la probabilidad obtenida en el punto anterior? Indique para quién implica un riesgo esta probabilidad.
    c) Si el valor de $LTPD$ se fija en 0,015, obtenga la probabilidad de que un lote con dicha proporción de piezas defectuosas sea aceptado. Indique para quién implica un riesgo esta probabilidad.
    d) ¿Qué significa que $LTPD$ vale 0,015 en términos del problema?

18. Una empresa utiliza en su proceso de producción tornillos especiales de acero. Los compra a un proveedor, que los envía en lotes con una gran cantidad de unidades. Para la recepción de los lotes se acordó el siguiente procedimiento: el comprador toma una muestra de 20 tornillos y evalúa si cada uno de ellos cumple o no con las especificaciones fijadas respecto de la longitud. Si el número de tornillos que no cumplen las especificaciones es menor o igual que 1, el lote se pasa a producción; si no, se devuelve al comprador. Para esta regla de decisión, el riesgo del proveedor es aproximadamente 0,017.
    a) Considere los siguientes posibles valores para AQL: $0,005 - 0,01 - 0,05 - 0,10$
    Indique cuál de estos valores es el AQL que corresponde a este caso. Justifique.
    b) Plantee las hipótesis nula y alternativa correspondientes, según el valor elegido en el ítem anterior.
    c) Si el valor de $LTPD$ se ha fijado en 0,03, calcule el riesgo del comprador e indique de qué maneras este puede reducirse.

19. Una fábrica está estudiando lanzar al mercado un nuevo tipo de galletitas finas. Este producto resulta altamente costoso pero la empresa sostiene que si tiene el éxito esperado, su producción va a reportar altas utilidades. Un equipo interdisciplinario evaluó que valdría la pena realizar una campaña promocional en el mercado local si por lo menos el 50 % de los consumidores evalúa al producto con una calificación mayor a 7.
    En un primer estudio, se tomó una muestra aleatoria simple de 20 consumidores potenciales y se registró la calificación dada por cada uno de ellos. Los resultados obtenidos fueron:

    8 - 10 - 4 - 10 - 2 - 9 - 8 - 7 - 8 - 8 - 9 - 5 - 8 - 8 - 10 - 9 - 9 - 10 - 8 - 10

    a) Defina población y población estadística, variable y parámetro de interés.

b) Plantee las hipótesis nula y alternativa asociadas.
Sugerencia: Dado que la variable solo toma valores enteros, para no generar inconvenientes en la obtención del valor observado del estadístico se sugiere plantear en la hipótesis nula un valor no entero.

c) Lleve adelante la prueba de hipótesis e informe sus conclusiones en contexto.

20. Para una empresa que realiza ventas telefónicas, una variable de interés es el tiempo (en segundos) que un cliente espera en línea al representante de ventas. Se han introducido algunos cambios en la organización del servicio con el objetivo de lograr que en por lo menos la mitad de las llamadas, los clientes esperen como máximo 12,5 s en línea antes de ser atendidos. Se lleva adelante un estudio con el objetivo de verificar si se cumple la meta de la empresa. Se selecciona una muestra aleatoria de 21 llamadas de clientes y se registra para cada una el tiempo que el cliente estuvo en línea esperando al representante de ventas.
Los datos fueron analizados por dos personas que no concordaron en sus conclusiones:
   - El analista A observó que en 15 de las 21 llamadas, los clientes esperaron menos de 12,5 s por lo que considera que la meta está cumplida.
   - El analista B analizó los datos con R y concluyó que no se cumplía la meta de la empresa ya que el tiempo promedio de espera en línea podía ser mayor que 13 s.

   Con R se obtuvo lo siguiente:

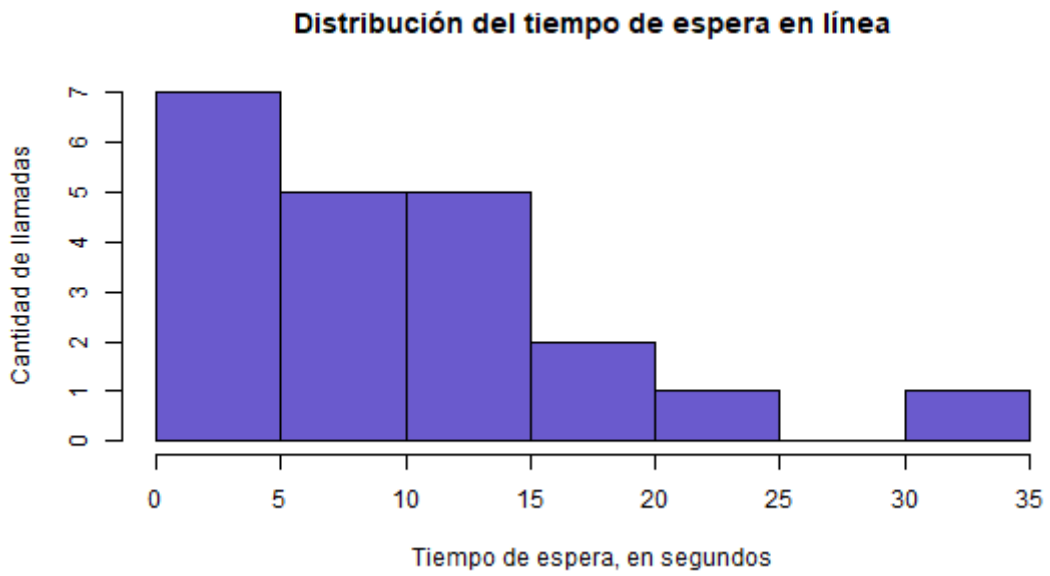
```
summary(tiempo)
Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
 3.0   5.0      8.0     10.9     14.0     35.0

MeanCI(tiempo, sd=NULL, method ="classic", conf.level=0.95)
mean    lwr.ci    upr.ci
10.905   7.261    14.549

MedianCI(tiempo, conf.level=0.95)
median lwr.ci upr.ci
8        5     14
attr(,"conf.level")
0.973

MedianCI(tiempo, conf.level = 0.95, sides = c("right"))
median lwr.ci upr.ci
8      -Inf     11
attr(,"conf.level")
0.961

MedianCI(tiempo, conf.level = 0.95, sides = c("left"))
median lwr.ci upr.ci
8        5    Inf
attr(,"conf.level")
0.961
```



Ambos analistas lo consultan a Ud. acerca de sus conclusiones contradictorias.
¿Considera Ud. que alguno de ellos tiene razón?

- En caso afirmativo, explique su razonamiento.
- En caso negativo, elabore sus propias conclusiones respecto del cumplimiento de la meta de la empresa, en relación al tiempo de espera en línea.

21. Reconsidere el proceso de producción de los cojinetes mencionado en la Actividad 7. Se desea verificar si la desviación estándar del diámetro de los mismos es superior a 0,02. Con la información de la muestra se obtiene la siguiente salida de R:

```
ad.test(cojinetes)
Anderson-Darling normality test
A = 0.29021, p-value = 0.5487

VarTest(cojinetes, alternative = "less", sigma.squared = 0.0004,
conf.level = 0.95)
One Sample Chi-Square test on variance
data:  cojinetes
X-squared = 17.25, df = 11, p-value = 0.8993
alternative hypothesis: true variance is less than 4e-04
95 percent confidence interval:
0.0000     0.0015
sample estimates:
variance of x
0.0006

VarTest(cojinetes, alternative = "greater",sigma.squared = 0.0004,
conf.level= 0.95)
One Sample Chi-Square test on variance
data:  cojinetes
```

```
X-squared = 17.25, df = 11, p-value = 0.1007
alternative hypothesis: true variance is greater than 4e-04
95 percent confidence interval:
0.0004     Inf
sample estimates:
variance of x
0.0006

VarTest(cojinetes, alternative = "two.sided",sigma.squared = 0.0004,
conf.level = 0.95)
One Sample Chi-Square test on variance
data:  cojinetes
X-squared = 17.25, df = 11, p-value = 0.1286
alternative hypothesis: true variance is not equal to 4e-04
95 percent confidence interval:
0.0003     0.0018
sample estimates:
variance of x
0.0006
```

a) Plantee las hipótesis nula y alternativa que le permitan responder a la inquietud planteada.
b) Indique cuál es el estadístico de la prueba y su distribución de probabilidades bajo H_o. Justifique.
c) ¿Se cumplen los requerimientos sobre la distribución de la variable, para poder llevar a cabo la prueba con la distribución $\chi 2$? Responda a esta pregunta planteando previamente las hipótesis nula y alternativa correspondientes a la prueba de bondad de ajuste e informando su decisión.
d) Concluya sobre la prueba de hipótesis planteada respecto del parámetro de interés. Indique con qué nivel de significación tomó su decisión.
e) Comente sobre la información adicional que le brinda el intervalo de confianza correspondiente.
f) De haber un error en este caso, ¿qué tipo de error podría ser? Justifique.

22. Una empresa cuenta con un proveedor al que compra habitualmente piezas metálicas. El comportamiento en probabilidad del diámetro (en milímetros) de dichas piezas se puede describir adecuadamente por una distribución Normal de parámetros $\mu = 24,5$ y $\sigma = 3,55$. Un nuevo proveedor ofrece las piezas metálicas a menor costo y afirma que las piezas que él comercializa tienen, en relación al diámetro, el mismo comportamiento que las del proveedor actual. Para poner a prueba esta afirmación, se tomaron al azar 35 piezas del nuevo proveedor y se registró el diámetro de cada una.
    a) Indique el objetivo del estudio estadístico planteado y proponga la línea de análisis a seguir con los datos, para lograr el objetivo planteado.
    b) Plantee las hipótesis nula y alternativa requeridas para poner a prueba la afirmación del nuevo proveedor.
    c) Los datos obtenidos se procesaron con R y se obtuvo la siguiente salida:

```
ad.test(diametro)
Anderson-Darling normality test
data:  diametro
A = 0.52305, p-value = 0.1707

t.test(diametro, alternative = "two.sided", mu = 24.5, conf.level = 0.95)
One Sample t-test
data:  diametro
t = -0.718, df = 34, p-value = 0.4776
alternative hypothesis: true mean is not equal to 24.5
95 percent confidence interval:
22.311     25.546
sample estimates:
mean of x
23.929

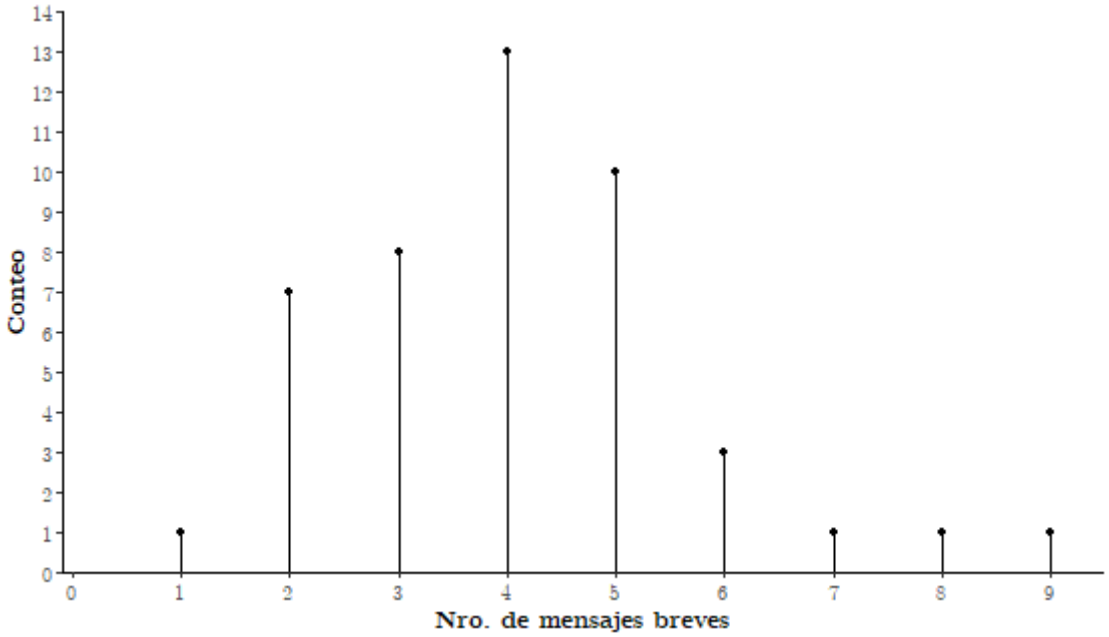
VarTest(diametro,alternative="two.sided",sigma.squared = 12.6025,
conf.level = 0.95)
One Sample Chi-Square test on variance
data:  diametro
X-squared = 59.795, df = 34, p-value = 0.005493
alternative hypothesis: true variance is not equal to 12.6025
95 percent confidence interval:
14.501     38.047
sample estimates:
variance of x
22.1639
```

Analice los resultados obtenidos en función de lo planteado en el ítem a. Informe sus conclusiones.

d) ¿Los datos resultaron significativos al 5 %?

e) ¿Aconsejaría usted el cambio de proveedor? Justifique su respuesta. Si considera que para concluir se requiere información adicional, indique cuál.

23. Reconsidere la Actividad 29 (Capítulo 2, página 120), en la cual se evalúa al sistema de atención telefónica de los clientes de una empresa de servicios informáticos. El mismo está habilitado de lunes a viernes, de 9 a 17 horas. Dos operadores capacitados atienden las llamadas y buscan solucionar los problemas y reclamos ellos mismos y derivan, si es necesario, la consulta a un especialista, que actúa como supervisor. Cuando un cliente no puede comunicarse telefónicamente con la empresa porque los operadores están ocupados, el sistema le permite dejar un mensaje breve. Este mensaje es luego analizado directamente por el supervisor, que devuelve la llamada al cliente y trata de solucionar su problema. En la empresa consideran que, si el número promedio diario de mensajes breves es mayor que 5, debe incorporarse un nuevo operador al servicio, para no restar tiempo al supervisor, que tiene asignadas además otras tareas. En los últimos 45 días hábiles se registró el número de mensajes breves (ver datos en página



120) y con la información recabada se obtuvieron las siguientes salidas de R:

```
summary(mensajes)
Min.   1st Qu.  Median    Mean    3rd Qu.    Max.
1.000   3.000   4.000     4.067   5.000    9.000
sd(mensajes)
1.629
```

a) Realice el planteo completo del problema.
b) Plantee las hipótesis nula y alternativa que considere pertinentes en este caso.
c) Analice exhaustivamente la información obtenida y concluya sobre la hipótesis nula. ¿Qué decisión deben tomar en la empresa en relación a la contratación de un nuevo operador?

24. Una empresa envasa un producto alimenticio en recipientes individuales destinados a un cliente mayorista. Una característica crítica es el peso neto de llenado, medido en kilogramos, el cual debe estar en el intervalo $(1,94; 2,06)$. La empresa garantiza que como mínimo el 99 % de los envases producidos tienen un peso dentro del intervalo mencionado.
Los encargados del proceso de llenado llevaron adelante un estudio estadístico para analizar si se cumple con lo que la empresa garantiza. En el marco de dicho estudio, tomaron una muestra aleatoria de $n = 150$ envases y se obtuvo la salida que se presenta a continuación.

```
summary(llenado)
Min      Q1     Median    Mean       Q3     Max
1.897   1.976   2.006     2.006     2.033   2.108

Anderson-Darling normality test
A = 0.11676, p-value = 0.9904

MeanCI(llenado, sd=NULL, method ="classic", conf.level =0.99)
mean     lwr.ci    upr.ci
2.006     1.997     2.015
```

```
VarCI(llenado, method ="classic", conf.level =0.99)
var   lwr.ci    upr.ci
0.0019  0.0014    0.0026

sum(llenado>=1.94 & llenado<=2.06)
121
```

a) Si el peso neto de llenado se puede considerar normalmente distribuido, ¿cuánto deberían valer el promedio y la desviación estándar para que la empresa cumpla con lo que garantiza? Justifique.

b) Realice el planteo completo del problema. Incluya población, variable, parámetros de interés y objetivo del estudio estadístico. Plantee las hipótesis nula y alternativa que permitan decidir sobre cada uno de los parámetros de interés.

c) ¿Podría indicar si la muestra proviene de una población con distribución Normal? Plantee las hipótesis correspondientes y concluya.

d) Analice exhaustivamente los datos e informe sus conclusiones.

e) ¿Hay evidencia para afirmar que la proporción de envases con peso fuera de especificación es mayor a la que garantiza la empresa?

f) ¿Cómo utilizaría un intervalo de tolerancia en esta situación?

25. Para evaluar el desempeño de un instrumento de medición se realizaron 15 mediciones repetidas de una magnitud $\delta = 9$ unidades y se construyó el intervalo de confianza del 95 % para el promedio poblacional, que resultó $(8,96; 9,01)$. El instrumento utilizado se considera preciso si el parámetro correspondiente resulta menor que 0,08 unidades.

a) Defina población, variable y parámetros de interés.

b) Plantee las hipótesis nula y alternativa que permitan analizar la exactitud del instrumento.

c) ¿En qué casos se comete error de tipo I? ¿y error de tipo II? Exprese de acuerdo al contexto del problema.

d) Pruebe las hipótesis planteadas en el ítem b. Informe con qué nivel de significación trabaja.

e) ¿Puede afirmar que el equipo es exacto? ¿Necesita información adicional para ello? Explique.

f) Con la información obtenida, ¿puede Ud. informar sobre la precisión del instrumento? Plantee las hipótesis correspondientes de tal manera que si se rechaza H_o se concluya que el equipo es preciso.

g) ¿En qué casos se comete error de tipo I? ¿y error de tipo II? Exprese de acuerdo al contexto del problema.

h) Defina un nivel de significación y concluya.

26. En una empresa metalúrgica, uno de los procesos que se lleva a cabo es la producción de una pieza especial para camiones. La longitud de las piezas, en mm, debe cumplir con las especificaciones: 250 ± 10. El objetivo de la empresa es que, como mínimo, el 99 % de las piezas verifiquen las especificaciones requeridas. Por estudios anteriores se conocía que la longitud de las piezas podía considerarse distribuida según una ley Normal con media 249 mm y desviación estándar 8 mm. En la empresa no estaban conformes con el comportamiento de la longitud de las piezas e implementaron una serie de modificaciones buscando mejorar el proceso. Finalizada una

primera etapa de modificaciones y con el proceso estabilizado, en la empresa quieren conocer si se cumplen los objetivos pautados en relación a la longitud de las piezas y para ello toman una muestra de 60 piezas y a cada una le miden su longitud a través de un sistema automático. La longitud promedio de la muestra resultó 251,14 mm y la desviación estándar, 3,02 mm. Además se llevó a cabo una prueba de bondad de ajuste para evaluar la normalidad de las longitudes de las piezas y el resultado no fue estadísticamente significativo.

a) Si la normalidad de las longitudes es razonable, ¿cuánto deberían valer la longitud promedio y la desviación estándar para que se cumplan las pretensiones de la empresa metalúrgica? Justifique.

b) Realice el planteo completo del problema. Incluya población, variable, parámetros de interés y objetivos del estudio estadístico. Plantee las hipótesis nula y alternativa que permita decidir sobre cada uno de los parámetros de interés.

c) Comente qué análisis descriptivo realizaría con los datos de las longitudes de la muestra.

d) ¿Qué significa que la prueba de bondad de ajuste arrojó un resultado no significativo? Plantee las hipótesis nula y alternativa correspondientes e informe su conclusión. Indique un posible valor para el valor *p* de dicha prueba.

e) ¿Qué implicancias tiene su respuesta al ítem anterior para la validez de las pruebas de hipótesis sobre los parámetros de interés? Explique.

f) Defina las regiones de rechazo correspondientes a las hipótesis planteadas en el ítem b y obtenga las conclusiones con la información muestral. Indique si los resultados obtenidos son estadísticamente significativos.

g) Suponga que a partir de los datos obtenidos se construyó el siguiente intervalo de tolerancia:
$IT_{P=0,99;95\%} = (241,875; 260,405)$.

Interprételo en el contexto del problema.

27. El formaldehído es un compuesto químico utilizado para la formulación de los adhesivos que se emplean en la fabricación de tableros derivados de la madera. Las resinas de urea/formaldehído y melamina/formaldehído presentan un pequeño porcentaje de formaldehído libre que va a emitir el tablero con el tiempo. Este compuesto es considerado nocivo e irritante y es clasificado además como cancerígeno. Es por ello que el control de este compuesto en los tableros está contemplado tanto en normas internas como en la legislación vigente. En la Argentina, rigen las normas nacionales IRAM que fijan niveles máximos aceptables de emisión, la cual debe ser $\leq 3,5$ mg/m^{2}h. Una empresa fabrica tableros, que va a comercializar en Argentina y quiere saber si estos cumplen con las especificaciones en relación a la emisión. A tal fin toman una muestra aleatoria de 200 tableros y en cada uno se evalúa la emisión de formaldehido, obteniendo el siguiente resultado:

```
MeanCI(emision, method ="classic", conf.level =0.95)
mean   lwr.ci    upr.ci
2.81   2.757      2.863

sum(emision > 3.5)
7
```

```
normtol.int(emision, alpha = 0.02, P = 0.995, side = 1)
alpha    P    x.bar   1-sided.lower 1-sided.upper
0.02   0.90    2.81        2.242        3.378
```

a) Un encargado considera, observando el intervalo del 95 % de confianza para la media poblacional construido con los datos registrados, que la operación se realiza adecuadamente. El ingeniero a cargo de la investigación afirma en cambio, que ese no es el análisis apropiado. Dé su opinión al respecto. Justifique claramente.

b) ¿Qué información brinda el intervalo de tolerancia en relación a los objetivos planteados?

c) Suponga que el proceso de producción de los tableros debe ser revisado en forma completa si la proporción de tableros con emisión mayor a la admitida por la norma IRAM es mayor a 0,10.

1) Plantee el problema de decisión y escriba las hipótesis nula y alternativa correspondientes.
2) Defina un estadístico de prueba e indique su distribución de probabilidad bajo H_o. Establezca una regla de decisión suponiendo que la máxima probabilidad de que ocurra una falsa alarma es 0,02.
3) Con la información de la muestra de 200 tableros, lleve adelante la prueba y concluya.
4) El resultado obtenido, ¿fue estadísticamente significativo? Justifique.

Las siguientes actividades se refieren a los problemas presentados en la Sección 1.4 (página 7). Para algunos de ellos se fueron realizando tareas de las distintas etapas del ciclo PPDAC a lo largo de todo el texto. Se sugiere revisar todo lo realizado, completar las tareas que falten y luego realizar un breve informe con las principales conclusiones en contexto. Los datos se presentan en el archivo https://sites.google.com/view/libro-estadistica.

28. Considere el Problema 1 que se desarrolla en la empresa metalúrgica que fabrica barras metálicas. Suponga que se quiere decidir si la proporción de barras que cumplen con las especificaciones para la longitud supera 0,98. Realice el planteo completo del problema y siga adelante todas las etapas hasta obtener las conclusiones.

29. Reconsidere el Problema 1. Suponga que la longitud de las barras se comporta según el modelo Normal.

a) Indique qué valor deben tener la longitud promedio, μ, y la desviación estándar, σ, si se pretende que como mínimo el 99,73 % de las barras cumplan con las especificaciones impuestas por la industria automotriz.

b) Sobre la media poblacional, plantee las hipótesis nula y alternativa de tal manera que si se rechaza la hipótesis nula se informe que el promedio no es el apropiado.

c) Sobre la desviación estándar poblacional, plantee las hipótesis nula y alternativa de tal manera que si se rechaza la hipótesis nula se informe que el valor de este parámetro es adecuado (según las pretensiones de la metalúrgica).

30. Considere el Problema 4 que tiene lugar en la empresa dedicada a la comercialización de productos en línea. En dicha empresa están preocupados por el incremento en las devoluciones, lo cual aumenta sus costos operativos y hace que su reputación como vendedores disminuya notoria-

mente. La Gerencia propone determinar el motivo principal de cada una de estas devoluciones, para detectar las causas más frecuentes e implementar luego acciones para mejorar el servicio y no seguir perdiendo clientes. En la Gerencia sospechan que una de las posibles causas podría ser el retraso en la entrega por problemas en el transporte y consideran que, si esta causa representa más del 30 % de las devoluciones, se justifica la contratación de una nueva empresa prestadora. Plantee las hipótesis que le permitan decidir si se debe contratar a una nueva empresa prestadora y concluya.

31. Considere el Problema 5 que se refiere a una empresa que fabrica cables eléctricos de diferente tipo. La empresa está interesada en estudiar el comportamiento del número de imperfecciones en rollos de cable unipolar de 2,5 mm, ya que tiene la posibilidad de abastecer a un cliente muy importante, que pretende que la proporción de rollos en los que se observan 2 imperfecciones o menos sea mayor o igual que 0,90.
¿Se podrá abastecer a dicho cliente? LLeve a cabo el planteo estadístico completo del problema y siga con las etapas del Ciclo PPDAC hasta informar las conclusiones correspondientes a los fabricantes de cables. Luego escriba un párrafo con dichas conclusiones, sin utilizar lenguaje estadístico.

32. Considere el Problema 6, relativo a un comitente que quiere conocer ciertas características de un tramo de ruta de 5 km ya construido. A partir de los criterios de calidad explicitados en el enunciado del problema, plantee las hipótesis correspondientes y concluya, utilizando la información recabada en el estudio. Informe si la ruta cumple con las exigencias de calidad en cuanto a espesor y resistencia.

33. Reconsidere el Problema 7. Suponga que Ud. es la persona a cargo de la evaluación del multímetro y que, con el fin de determinar si el procedimiento de medición en el que se utiliza dicho instrumento cumple los requisitos de exactitud y precisión, realizó 15 mediciones de una tensión patrón de 12 V y 15 mediciones de una resistencia patrón de 100 Ω. Suponga además lo siguiente:
    - Para la tensión se admite que el error sistemático sea como máximo 0,02 V(en valor absoluto). Además, el procedimiento se considera preciso si el parámetro correspondiente es menor que 0,01 V.
    - Para la resistencia se admite que el error sistemático sea como máximo 0,5 Ω (en valor absoluto). Además, el procedimiento se considera preciso si el parámetro correspondiente es menor que 0,2 Ω.

    a) Realice el planteo del problema y comente algunas cuestiones que se tuvieron en cuenta en la etapa de planificación.
    b) En relación a las mediciones de tensión:
        1) Plantee las hipótesis nula y alternativa que le permitan decidir sobre la exactitud del procedimiento de medición. Concluya.
        2) Plantee las hipótesis nula y alternativa que le permitan decidir sobre la precisión de dicho procedimiento. Concluya.
    c) En relación a las mediciones de resistencia:
        1) Plantee las hipótesis nula y alternativa que le permitan decidir sobre la exactitud y sobre la precisión del procedimiento de medición. Concluya.
        2) Suponga ahora que el procedimiento de medición de resistencia se considera exacto si

el error sistemático no supera 0,1 Ω en valor absoluto. Analice si cambia la conclusión anterior. Comente explícitamente la decisión adoptada respecto de la hipótesis nula y discuta la información adicional que aporta el intervalo de confianza para la media poblacional al evaluar el cumplimiento de este nuevo criterio de exactitud.

34. Considere el Problema 8, relativo al tiempo hasta la primera falla de dispositivos electrónicos. Se pretende que la mayoría de los dispositivos falle por primera vez después de las 150 horas. Suponga que Ud. es el usuario que debe decidir si comprar o no un cierto tipo de dispositivo. Plantee el problema en términos estadísticos y siga adelante con el proceso de resolución hasta informar sus conclusiones.

9.7 Estadística con R

9.7.1 Pruebas de hipótesis respecto de parámetros

Al igual que para la construcción de los intervalos de confianza, para llevar a cabo las pruebas de hipótesis se requiere instalar la librería `DescTools`.

Algunas cuestiones generales:

- Para indicar si la prueba de hipótesis es unilateral o bilateral, se debe utilizar la opción `alternative`: `alternative = c("two.sided")` corresponde al caso bilateral, aunque no hace falta indicarlo, ya que es la opción que se considera por omisión.
  `alternative = c("less")` indica que interesa una prueba unilateral a la izquierda, mientras que `alternative = c("greater")` indica que interesa una prueba unilateral a la derecha.

- Si se incluye la opción `conf.level` y se define un nivel de confianza, además de la prueba de hipótesis, R brinda el intervalo correspondiente al parámetro de interés. Esto es opcional, aunque siempre conveniente.

A continuación se detallan las particularidades para cada parámetro de interés:

1. Pruebas de hipótesis para la media poblacional

Suponga que se plantean una hipótesis nula y una alternativa respecto de la media poblacional y se cuenta con un conjunto de datos (*dataset*) denominado `data` que incluye una variable denominada `variable`, para llevar adelante la prueba de hipótesis correspondiente.

El primer paso es guardar los valores observados de `variable` en un vector numérico, llamado `y`. Esto se hace mediante el código:

```
y <- data$variable
```

- Si se conoce el valor de la desviación estándar poblacional σ,
  la función `ZTest` de la librería `DescTools` permite llevar a cabo la prueba de hipótesis para la media poblacional de la siguiente forma:

  ```
  ZTest(y, sd_pop = a, mu = b, alternative = c("two.sided", "less", "greater"))
  ```

  En este caso, `y` es el vector numérico donde están los valores obtenidos en la muestra. En `sd_pop` se indica el valor del desvío estándar poblacional conocido ($\sigma = a$ en este caso). En `mu` se indica el valor de la media poblacional postulado en la hipótesis nula ($mu = b$ en este caso) y en la opción `alternative` se indica si interesa una prueba bilateral o unilateral, como ya se explicó en párrafos anteriores.
- Si no se conoce el valor de la desviación estándar poblacional σ,
  la función `t.test` de la librería `DescTools` permite llevar a cabo la prueba de hipótesis para la media poblacional de manera análoga:

  ```
  t.test(y, mu = b, alternative = c("two.sided", "less", "greater"))
  ```

También se pueden indicar las medidas de resumen correspondientes usando la función:

```
TTestA(mx, sx, nx, mu = b)
```

En este caso, `mx` es el valor de la media de la muestra, `sx` es el valor del desvío estándar de la muestra, `nx` es el tamaño de la muestra y en `mu` se indica el valor de la media poblacional postulado en la hipótesis nula (*mu* = *b* en este caso).



Problema 1 (pág. 7): *Suponga que, para la producción de barras, se quiere probar que la longitud promedio poblacional es mayor que* 250 *mm.*

- *Si desviación estándar poblacional se puede considerar conocida y su valor es* $0,2$ *mm* ($\sigma = 0,2$)*, se lleva a cabo la prueba basada en el estadístico Z, como se detalla a continuación:*

```
ZTest(longitud, sd_pop = 0.2, mu = 250, alternative = c("greater"), conf.level = 0.95)
```

Los resultados obtenidos son:

```
One Sample z-test

data:  longitud
z = -0.31038, Std. Dev. Population = 0.2, p-value = 0.6219
alternative hypothesis: true mean is greater than 250
95 percent confidence interval:
 249.9643      Inf
sample estimates:
mean of x
 249.9943
```

- *Si no se conoce el valor de la desviación estándar poblacional, se lleva a cabo la prueba basada en el estadístico T, como se detalla a continuación:*

```
t.test(longitud, mu = 250, alternative = c("greater"), conf.level = 0.95)
```

Los resultados obtenidos son:

```
One Sample t-test

data:  longitud
t = -0.20633, df = 119, p-value = 0.5816
alternative hypothesis: true mean is greater than 250
95 percent confidence interval:
 249.9488      Inf
sample estimates:
mean of x
 249.9943
```

También se puede llevar a cabo la misma prueba, utilizando las medidas de resumen:

```
TTestA(249.9943, 0.3, 120, mu = 250, alternative = c("greater"), conf.level = 0.95)
```

Los resultados obtenidos son:

```
  One Sample t-test

data:  x
t = -0.20813, df = 119, p-value = 0.5823
alternative hypothesis: true mean is greater than 250
95 percent confidence interval:
 249.9489      Inf
sample estimates:
mean of x
 249.9943
```

En cada uno de los tres casos presentados se llevó a cabo una prueba unilateral a la derecha y se construyó el intervalo del 95 *% de confianza para la longitud media poblacional de las barras. En el primer caso, con el estadístico Z, suponiendo conocido el valor del desvío estándar poblacional,* σ *y en los otros dos, con el estadístico T.*

2. Pruebas de hipótesis para la variancia y el desvío estándar poblacionales

Suponga que se plantean una hipótesis nula y una alternativa respecto de la variancia poblacional y se cuenta con un conjunto de datos (*dataset*) denominado `data` que incluye una variable denominada `variable`, para llevar adelante la prueba de hipótesis correspondiente.

El primer paso es guardar los valores observados de `variable` en un vector numérico, llamado `y`. Esto se hace mediante el código:

```
y <- data$variable
```

la función `VarTest` de la librería `DescTools` permite llevar a cabo la prueba de hipótesis para la variancia poblacional, basada en la estadística χ^2, lo cual es válido cuando la variable de interés se distribuye normalmente.

```
VarTest(y, sigma.squared = b)
```

Nuevamente, `y` es el vector numérico donde están los valores obtenidos en la muestra. En el argumento `sigma.squared` se indica el valor de la variancia postulado en H_o.



Problema 1 (pág. 7): *Suponga que, para la producción de barras, se quiere probar que la variancia poblacional es menor que* $0,1\ mm^2$.
Como la normalidad de la variable de interés es razonable, se puede llevar adelante la prueba utilizando la estadística χ^2*, como se detalla a continuación:*

```
VarTest(longitud, sigma.squared = 0.1, alternative = c("less"), conf.level = 0.99)
```

Los resultados obtenidos son:

```
One Sample Chi-Square test on variance

data:  longitud
X-squared = 107.71, df = 119, p-value = 0.238
alternative hypothesis: true variance is less than 0.1
99 percent confidence interval:
 0.0000000 0.1251361
sample estimates:
variance of x
   0.09051216
```

O también puede realizarse una prueba de hipótesis bilateral:

```
VarTest(longitud, sigma.squared = 0.1, conf.level = 0.99)
```

Los resultados obtenidos son:

```
One Sample Chi-Square test on variance

data:  longitud
X-squared = 107.71, df = 119, p-value = 0.5331
alternative hypothesis: true variance is not equal to 0.1
99 percent confidence interval:
 0.0662903 0.1297421
sample estimates:
variance of x
   0.09051216
```

En los casos presentados se realizó, respectivamente, una prueba unilateral a la izquierda y una prueba bilateral, y se construyeron los correspondientes intervalos de confianza del 99% para la varianza poblacional de la longitud de las barras.

3. Prueba de hipótesis para la proporción poblacional

Suponga que se plantean una hipótesis nula y una alternativa respecto de la proporción poblacional de unidades que cumplen cierta condición y se cuenta con la cantidad de unidades de la muestra que verifican dicha condición.

La función `binom.test` de la librería `DescTools` permite llevar a cabo la prueba de hipótesis exacta para este parámetro, basada en el modelo Binomial.

```
binom.test(x, n, p = b)
```

En `x` se indica la cantidad de unidades de la muestra que verifican la condición de interés, en `n` se indica el tamaño de la muestra y en `p` se indica el valor postulado para el parámetro en la hipótesis nula. Si se cuenta con datos correspondientes a una variable cuantitativa `y` y la condición es que los

valores de la variable sean mayores o menores que cierto valor, se puede contar la cantidad de unidades que la cumplen utilizando la función `sum` de la siguiente manera:

```
x<-sum(y < a)
x<-sum(y > b)
x<-sum(y < a|y > b)
```

En el primer caso, se cuenta la cantidad de unidades de la muestra que tienen valor de la variable menor que a; en el segundo caso, la cantidad de unidades que tienen valor mayor que b y en el último caso, la cantidad de unidades que tienen valor menor que a o mayor que b.



Problema 1 (pág. 7): *Suponga que, para la producción de barras, interesa la proporción de unidades que tienen longitud menor que* $249,4$ *o mayor que* $250,6$ *mm, es decir la propoción de unidades con longitud fuera de especificaciones. Se quiere probar que dicha proporción es mayor que* $0,07$.
Los datos de las longitudes de la muestra de $n = 120$ *barras se encuentra en una variable denominada* `longitud`. *Con la función* `sum` *se cuenta la cantidad de barras de la muestra con longitud fuera de especificaciones y se guarda ese valor en la variable* `x`:

```
x<-sum(longitud < 249.4|longitud > 250.6)
```

La función indica que hay 6 barras en esas condiciones. Con esa información, se lleva a cabo la prueba de la siguiente manera:

```
binom.test(x, 120, p = 0.07, alternative = c("greater"), conf.level = 0.90)
```

Los resultados obtenidos son:

```
Exact binomial test

data:  6 and 120
number of successes = 6, number of trials = 120, p-value = 0.8523
alternative hypothesis: true probability of success is greater than 0.07
90 percent confidence interval:
 0.02647208 1.00000000
sample estimates:
probability of success
                  0.05
```

Suponga ahora que se quiere probar que la proporción de barras con longitud fuera de especificaciones es mayor que 0,01.

```
binom.test(x, 120, p = 0.01, alternative = c("greater"), conf.level = 0.90)
```

Los resultados obtenidos son:

```
Exact binomial test

data:  6 and 120
number of successes = 6, number of trials = 120, p-value = 0.001384
alternative hypothesis: true probability of success is greater than 0.01
90 percent confidence interval:
 0.02647208 1.00000000
sample estimates:
probability of success
                  0.05
```

En los dos casos presentados, se llevaron a cabo pruebas unilaterales a la derecha y se construyeron los correspondientes intervalos del 90% *de confianza para la proporción de barras con longitud fuera de especificaciones.*

4. Prueba de hipótesis para la mediana poblacional

Suponga que se plantean una hipótesis nula y una alternativa respecto de la mediana poblacional y que estas se van a expresar en función de la proporción de unidades de la población para las cuales la variable de interés toma valores mayores que $\tilde{\mu}_o$.

La prueba correspondiente se denomina **prueba del signo**.

La función `signTest` de la librería `EnvStats` permite llevar a cabo la prueba de hipótesis y adicionalmente brinda un intervalo de confianza.

```
signTest(y, ci.method = "exact", min.coverage = a, mu = b)
```

`y` es el vector numérico donde están los valores obtenidos en la muestra.

En el argumento `ci.method = "exact"` se indica que la distribución de probabilidades que se supone para el estadístico de la prueba es la Binomial; en el argumento `min.coverage` se define el menor nivel de confianza pretendido y en el argumento `mu` se indica el valor de la mediana postulado en H_o.

La función `MedianCI` de la librería `DescTools` también brinda un intervalo de confianza basado en la distribución Binomial.



Problema 1 (pág. 7): *Suponga que, para la producción de barras, se quiere probar que la mediana poblacional es mayor que* 250 *mm. Se lleva adelante la prueba del signo, basada en la proporción de barras de la población con longitud mayor que* 250.

```
signTest(longitud, mu = 250, alternative = c("greater"), ci.method = "exact",
min.coverage = 0.95)
```

Los resultados obtenidos son:

```
$statistic
# Obs > median
            60
$parameters
NULL

$p.value
# Obs > median
           0.5
$estimate
median
250.01

$null.value
median
   250

$alternative
[1] "greater"

$method
[1] "Sign test"

$data.name
[1] "longitud"

$interval
$name
[1] "Confidence"

$parameter
[1] "median"

$limit.ranks
[1] 51 NA

$limits
   LCL    UCL
249.93    Inf

$type
[1] "lower"

$method
[1] "exact"
```

```
$conf.level
[1] 0.9507382

$sample.size
[1] 119
```

En este caso se realizó una prueba unilateral a la derecha y se obtuvo el correspondiente intervalo de confianza unilateral para la mediana de la longitud de las barras. En la salida se informa que en la muestra se encontraron 60 barras con longitud superior a 250 mm y que el valor p de la prueba es 0,5. Se indica también que la mediana muestral resultó 250,01 mientras que el valor postulado en H_o para el parámetro es 250. El intervalo de confianza correspondiente es $(249,93\ ;\ +inf)$, para el cual el nivel de confianza real es $95,07\,\%$. El tamaño de la muestra que se indica en la salida es 119, ya que se descarta una barra cuya longitud coincide exactamente con el valor postulado en la hipótesis nula.
También se puede construir un intervalo de confianza para la mediana poblacional, con la función `MedianCI`

```
MedianCI(longitud, sides = "left", conf.level = 0.95)
```

Los resultados obtenidos son:

```
 median  lwr.ci  upr.ci
250.005 249.930     Inf
attr(,"conf.level")
[1] 0.9587963
```

En este caso, se brinda el intervalo unilateral a la derecha y se informa que su confianza real es $95,88\,\%$.



Problema 8 (pág. 8): *En la empresa interesada en adquirir los dispositivos electrónicos sospechan que la mediana del tiempo hasta la falla de los mismos es menor que 300 horas. Se lleva adelante la prueba del signo.*

```
signTest(duracion, mu = 300, alternative = c("less"), ci.method = "exact",
min.coverage = 0.95)
$statistic
# Obs > median
            19

$parameters
NULL

$p.value
# Obs > median
      0.185649

$estimate
median
```

```
264.82

$null.value
median
   300

$alternative
[1] "less"

$method
[1] "Sign test"

$data.name
[1] "duracion"

$interval
$name
[1] "Confidence"

$parameter
[1] "median"

$limit.ranks
[1] NA 29

$limits
   LCL    UCL
  -Inf 334.54

$type
[1] "upper"

$method
[1] "exact"

$conf.level
[1] 0.9637729

$sample.size
[1] 45
```

En este caso se realizó una prueba unilateral a la izquierda y se obtuvo el correspondiente intervalo de confianza unilateral para la mediana del tiempo hasta la falla de los dispositivos. En la salida se informa que en la muestra se encontraron 19 dispositivos con tiempo hasta la falla superior a 300 horas y que el valor p de la prueba es aproximadamente 0,18. Se indica también que la mediana muestral resultó 264,82 mientras que el valor postulado en H_o para el parámetro es 300. El intervalo de confianza correspondiente es

$(-inf\ ;\ 334,54)$, *para el cual el nivel de confianza real es* $96,38\,\%$. *El tamaño de la muestra que se indica en la salida es 45, ya que para ningún dispositivo, el tiempo hasta la falla resultó exactamente igual a 300 horas (valor postulado en la hipótesis nula).*

```
MedianCI(datos$observados, conf.level = 0.95, sides = "right")
median lwr.ci upr.ci
264.82   -Inf 318.36
attr(,"conf.level")
[1] 0.9637729
```

En este caso, se brinda el intervalo unilateral a la izquierda y se informa que su confianza real es $96,38\,\%$.

9.7.2 Pruebas de bondad de ajuste

En la Sección 8.5 se presentó al diagrama de probabilidad Normal, como una herramienta gráfica que orienta sobre el posible modelo o distribución de probabilidades de la población de la cual se extrajo la muestra. En la Sección 8.8.1 presentó el código para construirlo.

Este diagrama también se pueden construir para otros modelos como el Exponencial, por ejemplo.

Además de observar el diagrama, se puede llevar adelante una prueba de bondad de ajuste. Las hipótesis que se plantean son las siguientes:

$H_o)Y \sim$ modelo
$H_a)Y$ no se comporta según el modelo

Si se quiere llevar adelante una prueba de bondad de ajuste para evaluar la normalidad, se cuenta con la función `ad.test` de la librería `nortest`. Al usar esta función se lleva a cabo una prueba de normalidad de Anderson Darling. `R` informa el valor del estadístico y el valor p correspondiente.

```
ad.test(y)
```

La librería permite realizar otros tests de Normalidad; pero el de Anderson Darling es el recomendado.



Problema 1 (pág. 7): *Suponga que se quiere probar si la longitud de las barras se comporta según el modelo Normal. Con los datos de la muestra de* $n = 120$ *barras, se obtiene lo siguiente:*

```
ad.test(longitud)
Anderson-Darling normality test
data:  longitud
A = 0.21792, p-value = 0.8372
```

El valor p brindado por la salida no permite descartar la hipótesis de normalidad para la longitud de las barras producidas por la empresa metalúrgica. En la Figura 9.28 (página 491) se presenta el diagrama de probabilidad Normal construido con los datos de la muestra, y los resultados de la prueba de normalidad de Anderson-Darling.

En el caso del modelo Exponencial, en la Sección 8.8.1 (Capítulo 8, página 438) se presenta el código para construir un diagrama de probabilidad correspondiente.

Si se desea evaluar si este modelo proporciona un ajuste razonable, puede utilizarse la función `exp_test_ad` de la librería `agop`. Esta función implementa la prueba de bondad de ajuste de Anderson–Darling para la distribución Exponencial.

El resultado de la función incluye el valor del estadístico de prueba y el valor *p* asociado.



Problema 8 (pág. 8): *Suponga que se quiere probar si el tiempo hasta la falla de los dispositivos se comporta según el modelo Exponencial. Con los datos de la muestra de* $n = 45$ *dispositivos, se obtiene lo siguiente:*

```
library(agop)
exp_test_ad(tiempo_hasta_falla)
Anderson-Darling exponentiality test
data:  tiempo_hasta_falla
W = 0.34667, p-value = 0.7406
```

El valor p brindado por la salida no permite descartar la hipótesis de distribución Exponencial para el tiempo hasta la falla de los dispositivos de interés. En la Figura 9.30 (página 493) se presenta el diagrama de probabilidad Exponencial construido con los datos de la muestra, y los resultados de la prueba de bondad de ajuste Exponencial de Anderson-Darling.



Bibliografía

[1] Chris J Wild and Maxine Pfannkuch. Statistical thinking in empirical enquiry. *International statistical review*, 67(3):223–248, 1999.

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[3] Gnu general public license, version 3. `http://www.gnu.org/licenses/gpl.html`, Junio 2007.

[4] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2019.

[5] The R Foundation. The Comprehensive R Archive Network (CRAN), 2026.

[6] Hadley Wickham. Tidy data. *The Journal of Statistical Software*, 59, 2014.

[7] Hadley Wickham and Jennifer Bryan. *readxl: Read Excel Files*, 2019. R package version 1.3.1.

[8] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[9] Hadley Wickham, Jim Hester, and Romain Francois. *readr: Read Rectangular Text Data*, 2018. R package version 1.3.1.

[10] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[11] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. R package version 1.0.5.

[12] Matthew Bognar. *Probability Distributions, versión 5.8 [Aplicación móvil]*, 2024. Department of Statistics and Actuarial Science, University of Iowa.

[13] Douglas G. Bonett. Approximate confidence interval for standard deviation of nonnormal distributions. *Computational Statistics & Data Analysis*, 50:775–782, 2006.

[14] Gerald J. Hahn, William Q. Meeker, and Luis A. Escobar. *Statistical Intervals: A Guide for Practitioners and Researchers*. Wiley, 2nd edition, 2017.

[15] David K. Hildebrand and R. L. Ott. *Estadística aplicada a la administración y a la economía*. Addison Wesley Iberoamericana, México, 3 edition, 1997.