Badler Clara
Alsina Sara
Pagano Ariel
Puigsubirá Cristina
Vitelleschi María Susana
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad Ciencias Económicas y Estadística Universidad Nacional de Rosario.[1]

UNA PROPUESTA DE EVALUACION PARA CONJUNTOS DE DATOS CON INFORMACIÓN FALTANTE EN LA ENCUESTA PERMANENTE DE HOGARES[1]

1. INTRODUCCION

El análisis de conjuntos de datos con información faltante requiere tratamiento de la misma, el que generalmente se realiza condicionado al planteo de ciertos supuestos. La metodología más generalizada supone que: "la información perdida tiene las mismas características que la observada".

Para establecer una forma operacional desde el punto de vista estadístico de evaluación de dicho supuesto, es necesario traducir el comportamiento en términos de distribuciones de probabilidad y trabajar con una o más variables relacionadas con la variable con pérdida y observadas en forma completa, y proponer pruebas estadísticas. Las pruebas a distribución libre que presentan características prácticas tanto para investigadores como usuarios son una alternativa útil para tal fin.

En este trabajo se propone la aplicación de las pruebas de Wilcoxon y Kolmogorov para evaluar la hipótesis de no existencia de diferencias entre los datos observados y los perdidos y su aplicación a información de la EPH, correspondiente al Aglomerado Gran Rosario, trabajándose con el monto del ingreso total familiar y la edad de los desocupados.

2. METODOLOGÍA

A partir de observaciones perdidas en una determinada variable se identifican dos grupos a los cuales se denomina: "grupo de información completa" (GIC) y "grupo de información faltante" (GIF), según la presencia o no de información en ella. Se considera información adicional en otra variable relacionada que se observa de manera completa para todas las unidades. Se obtiene una descripción de ésta última, comparando los grupos GIC y GIF mediante medidas resúmenes: la media (Ma), el modo (Mo), la mediana (Mna), el desvío standard (Sd) y el coeficiente de asimetría (As). Se aplican las pruebas de Wilcoxon y de Kolmogorov para evaluar las diferencias en las distribuciones de los grupos.

[1] Proyectos: PICTNº 0200095-01996 de la ANPyT y PID Nº 19/E045 de la SECyT

2.1. Prueba de Wilcoxon (W)

Permite probar la hipótesis de que dos muestras aleatorias pueden ser pensadas como una sola muestra de una población.

Dado las observaciones $X_1, \ldots, X_m$ e $Y_1, \ldots, Y_n$ se plantean los siguientes supuestos:

. $X_1, \ldots, X_m$ constituyen una muestra aleatoria de la población 1 (iid) y las $Y_1, \ldots, Y_n$ constituyen una muestra aleatoria de la población 2 (iid).

. Las X's y las Y's son mutuamente independientes, o sea que se plantea el supuesto de independencia dentro y entre las muestras.

. Las poblaciones 1 y 2 son continuas.

Siendo F y G las funciones de distribución correspondientes a las poblaciones 1 y 2 la hipótesis nula es:

$$H_0: F(t) = G(t) \qquad \forall\, t \qquad (1)$$

O sea que X e Y tienen la misma distribución de probabilidad pero la distribución común no está especificada.

La hipótesis alternativa en este caso especifica que Y tiende a ser mayor (o menor) que X, usándose para describirla, el modelo de traslación:

$$G(t) = F(t - \Delta) \qquad \forall\, t \qquad (2)$$

Este modelo expresa que la población 2 es la misma que la población 1 excepto por un desplazamiento Δ, parámetro que se denomina "cambio de ubicación".

Usando este modelo, (1) se reduce a:

$$H_0: \Delta = 0$$

Para obtener la estadística W, se ordena la muestra combinada de m + n valores de X e Y de menor a mayor. Si S_1 es el rango de Y_1 , . . . y S_n el rango de Y_n en este orden, W es la suma de los rangos asignados a los valores de Y, esto es:

$$W = \sum_{j=1}^{n} S_j$$

Bajo una hipótesis alternativa bilateral, con un nivel de significación α, se rechaza H_0 si:

$$W \geq W_{\alpha/2} \text{ o si } W \leq n(m+n-1) - W_{\alpha/2} \qquad (3)$$

La aproximación para muestras grandes se basa en la normalidad asintótica de W convenientemente estandarizada. Cuando H_0 es verdadera, la estadística W estandarizada es:

$$W^* = \frac{W - E_0(W)}{Var_0(W)} = \frac{W - \{n(m+n+1)/2\}}{\{m\,n\,(m+n+1)/12\}^{1/2}}$$

W^* posee, cuando el mínimo valor entre m y n tiende a infinito, una distribución asintótica N(0,1). La aproximación por teoría normal al procedimiento (3) es:

$$\text{Rechazar } H_0 \text{ si: } \left| W^* \right| \geq Z_{\alpha/2}$$

En el caso que existan unidades con el mismo valor de la variable, llamados ligas, se proporciona a las mismas el promedio de aquellos rangos que les hubiera correspondido. En estos casos, al calcular W^*, la esperanza de W bajo la hipótesis nula no se ve afectada pero sí su variancia.

2.2. Prueba de Kolmogorov (D)

Permite evaluar el ajuste de diferentes funciones de distribución a una muestra. Las observaciones $X_1,..., X_n$ constituyen una muestra aleatoria de una población con función de distribución F(x). A partir de la observación de las distribuciones empíricas existen elementos para creer que F(x) es una función de distribución que puede especificarse aunque no completamente. En este caso, bajo H_0 se postula que F es miembro de alguna familia paramétrica con uno o más parámetros no especificados.

La hipótesis nula es:

$$H_0 : F(x) = \hat{F}_0(x) \qquad \forall\, x$$

versus

$$H_1 : F(x) \neq \hat{F}_0(x) \qquad \text{para al menos una } x$$

donde $\hat{F}_0(x)$ es el estimador de la función F(x) utilizando, para sus parámetros desconocidos, la estimación máximo verosímil.

La estadística D se calcula como:

$$D = \sup_{-\infty < x < +\infty} \left\{ \left| Fn(x) - \hat{F}_0(x) \right| \right\}$$

donde Fn(x) es la función de distribución muestral y $\sup_{-\infty < x < +\infty}$ denota al valor máximo sobre las x de las diferencias en valor absoluto entre $F_n(x)$ y $\hat{F}_0(x)$.

Con un nivel de significación α, se rechaza H_0 si: $D \geq d_\alpha$

Con un nivel de significación del 5% y un total de observaciones mayor que 30, se rechaza H_0 si:

$$D > \frac{1.36}{\sqrt{n}}$$

En este trabajo, se postula la hipótesis respecto a la distribución Normal de parámetros $\mu = \mu_0$ (esperanza) y $\sigma^2 = \sigma^2_0$ (variancia) y Gamma de parámetros $\alpha=\alpha_0$ (forma) y $\beta=\beta_0$ (escala).

3. MATERIAL

La metodología propuesta es aplicada a información proveniente de las ondas Octubre 97 y Octubre 98 de la EPH Aglomerado Gran Rosario. El grupo de análisis es la sub-base de "Desocupados", definida a partir de los criterios de clasificación de la condición laboral implícita en la encuesta.

Una variable de importancia en la encuesta es el "Monto del Ingreso Total Familiar" (ITF), la cual asigna a cada miembro del hogar el monto total de los ingresos del hogar. Al analizar esta variable se identifican observaciones con valores iguales a cero, que a través del monitoreo se comprueba que no siempre corresponden a ingresos nulos, ya que de acuerdo con los criterios de codificación, valores de ITF nulos pueden presentarse cuando:

. los ingresos individuales de cada miembro del hogar son iguales a cero

. los ingresos, de al menos uno de los miembros del hogar, no son declarados o registrados

Dado esta característica en la información, se conviene en considerar a los valores nulos de ITF como perdidos y al conjunto de unidades con dicha característica como "grupo de información faltante" (GIF); el resto de las unidades integran el "grupo de información completa" (GIC).

Una variable relacionada con la variable de análisis y sobre la cual se intentan detectar diferencias de comportamiento entre los grupos GIC y GIF, es la EDAD (en años cumplidos), la cual se observa completamente a través de las unidades de ambos.

4. RESULTADOS

A partir de la consideración de valores nulos en ITF como faltantes, se definen los grupos GIF y GIC en cada onda cuya composición se detalla en el Cuadro 1.

Cuadro 1. Distribución de los desocupados según grupo y onda

	Grupo de Información		
Onda	GIF	GIC	Total de Desocupados
Octubre 97	33 (16.1%)	172 (83.9%)	205 (100%)
Octubre 98	40 (23.5%)	170 (76.5%)	210 (100%)

Considerando la EDAD de las unidades de cada grupo, el conjunto de medidas descriptivas permite analizar las diferencias de comportamiento de la misma en ambos grupos (Cuadro 2).

Cuadro 2. Medidas descriptivas de la variable EDAD según grupo y onda

Onda	Octubre 97		Octubre 98	
Medida	GIF	GIC	GIF	GIC
Ma	37.9	31.7	34.8	33.9
Mo	19 y 43	19 y 21	19	18
Mna	43	29	29	33
Sd	16.5	13.3	15.6	15.1
As	0.16	0.7	0.59	0.62

La distribución de la EDAD es presentada para ambas ondas en los dos grupos, notándose en Octubre 97 para el grupo GIF, la ausencia de observaciones entre los 29 y 42 años.

Gráfico 1. Distribución de la EDAD del grupo GIF. Onda Octubre 1997

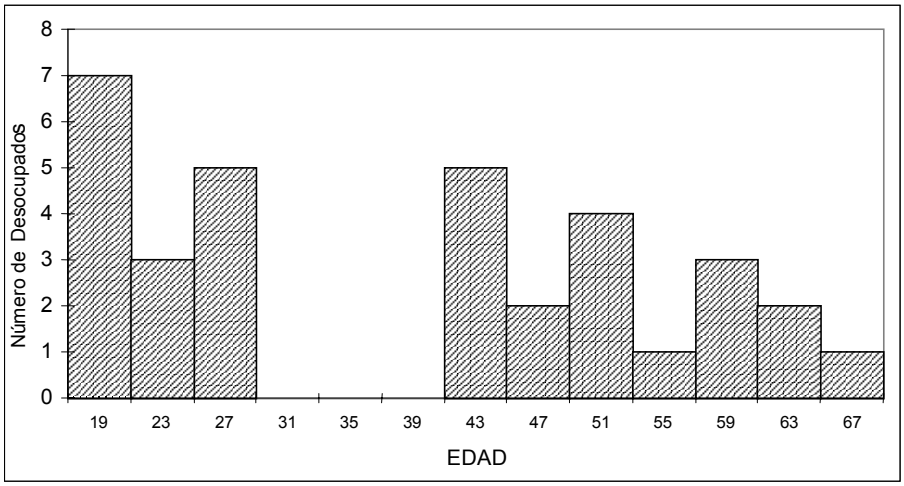
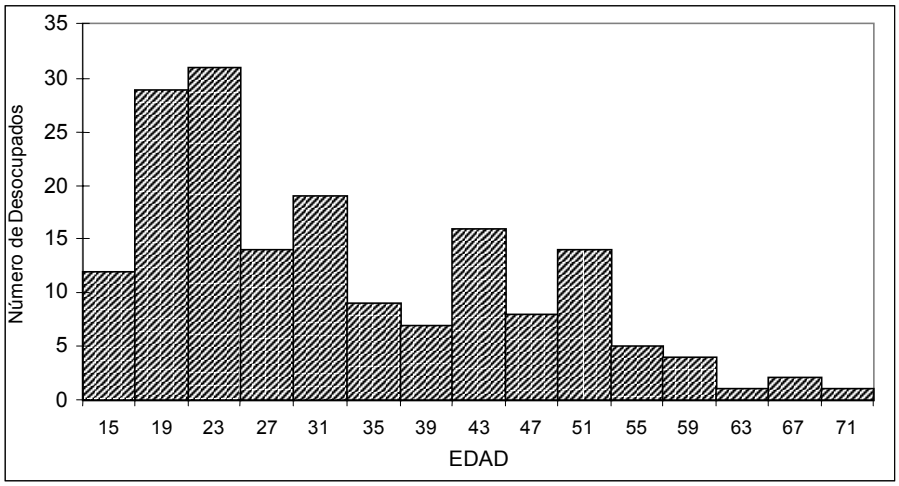
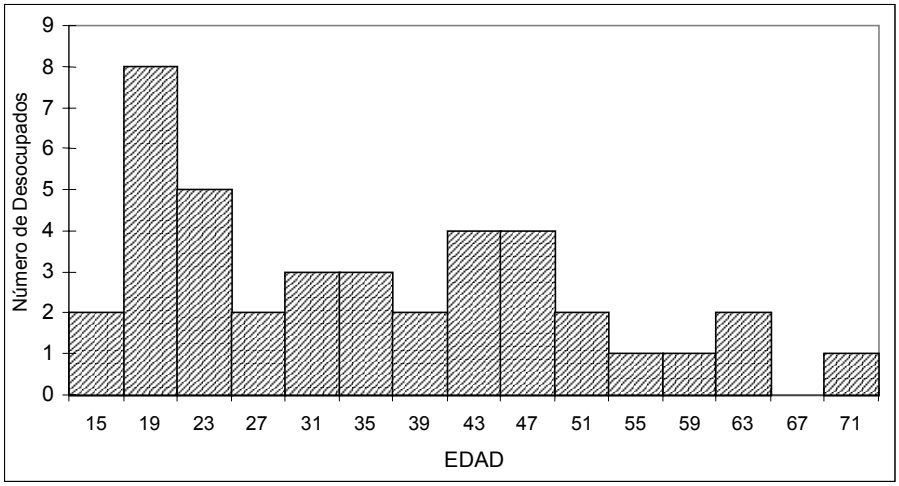


Gráfico 2. Distribución de la EDAD del grupo GIC. Onda Octubre 1997.



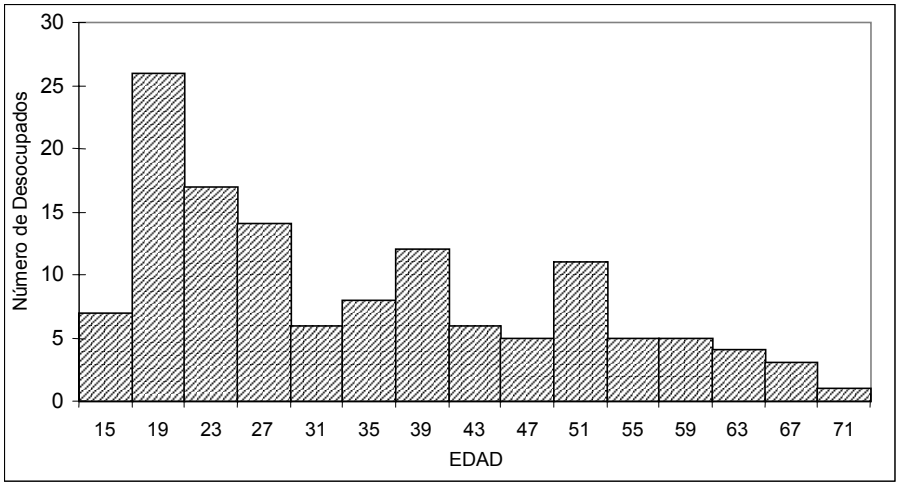
De la comparación entre las distribuciones (Gráfico 1 y 2) y las medidas descriptivas obtenidas, se observa que el comportamiento de la EDAD en Octubre 97 varía entre grupos, destacándose que el grupo GIC presenta una mayor asimetría hacia la derecha

Gráfico 3. Distribución de la EDAD del grupo GIF. Onda Octubre 1998.



Para la onda Octubre 98 (Gráfico 3 y 4), la distribución de la EDAD presenta asimetría positiva tanto para el grupo GIF como para el grupo GIC.

Gráfico 4. Distribución de la EDAD en el grupo GIC. Onda Octubre 1998.



De la comparación de las distribuciones se aprecia que para la onda Octubre 98, la EDAD posee similar ubicación y forma en los grupos GIF y GIC.

Se calcula la estadística W* y con un nivel de significación del 10% se rechaza la hipótesis nula para la onda Octubre 97 y no para Octubre 98, reafirmándose que para Octubre 97 la distribución de la EDAD del grupo GIF se encuentra desplazada con respecto al grupo GIC, mientras que para la onda Octubre 98 esto no ocurre (Cuadro 3).

Cuadro 3. Aplicación de la prueba de Wilcoxon

Onda	W*	Prob. Asociada
Octubre 97	1.8	0.07
Octubre 98	0.255	0.8

A partir de la observación de las distribuciones de la variable EDAD para cada grupo se realiza la aplicación de la prueba de Kolmogorov postulando dos hipótesis diferentes con las funciones de distribución Normal y Gamma.

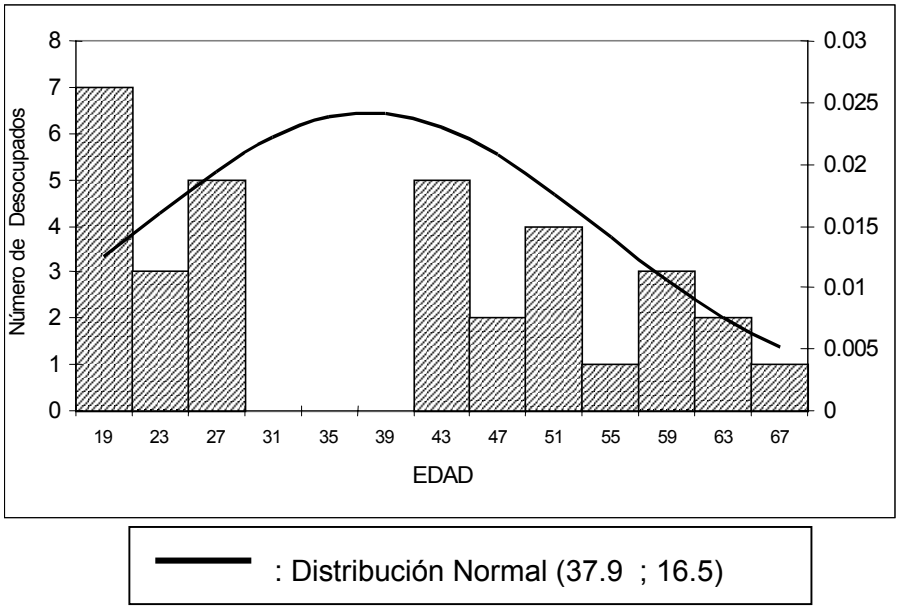
Cuadro 4. Valores de la estadística D según onda y grupo de información.

Onda / Distribución	Octubre 97		Octubre 98	
	GIF	GIC	GIF	GIC
Normal	0.159 (NR)	0.115 (R)	0.09 (NR)	0.135 (R)
Gamma	0.174(NR)	0.08 (NR)	0.062 (NR)	0.093 (NR)
Valores Críticos				
$1.36/\sqrt{n}$	0.237	0.104	0.215	0.104

Los resultados del test llevan a concluir que para la onda Octubre 97, la distribución de la EDAD del grupo GIF puede provenir tanto de una distribución Normal como de una Gamma, presentando menor discrepancia entre la distribución empírica y la propuesta, una distribución Normal con parámetros (37.88; 16.5). Para el grupo GIC, la distribución de la EDAD provendría de una distribución Gamma con parámetros (6.06; 5.24) ya que la hipótesis de normalidad resulta significativa.

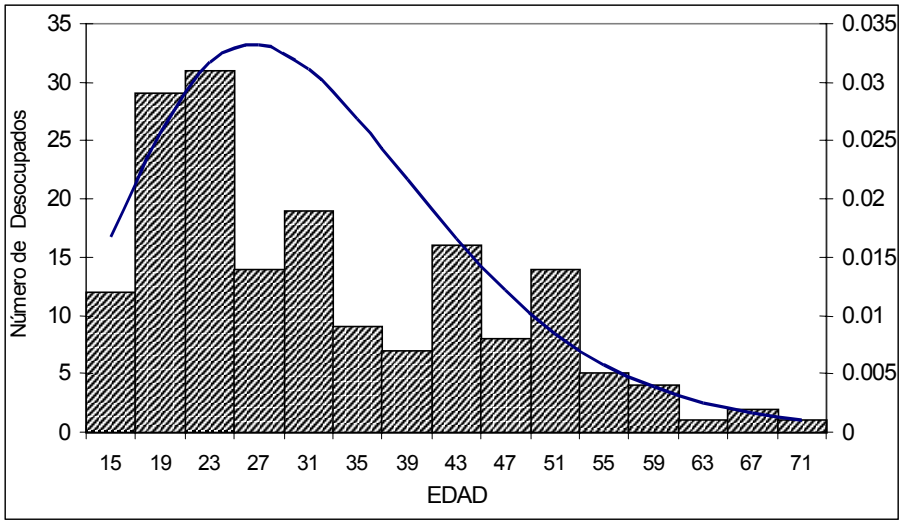
Para la onda Octubre 98, la distribución de la EDAD del grupo GIF puede provenir tanto de una distribución Normal como de una Gamma, presentando menor discrepancia esta última con parámetros (5.1; 6.8). Para el grupo GIC la EDAD puede considerarse proveniente de una Gamma de parámetros (5.26; 6.4) ya que la hipótesis de normalidad resultó significativa. Para cada caso se presenta el ajuste propuesto (Gráficos 5 a 8)

Gráfico 5. EDAD del grupo GIF. Onda Octubre 1997



: Distribución Normal (37.9 ; 16.5)

Gráfico 6. EDAD del grupo GIC. Onda Octubre 1997



: Distribución Gamma (6.06 ; 5.24)

Gráfico 7. EDAD del grupo GIF. Onda Octubre 1998

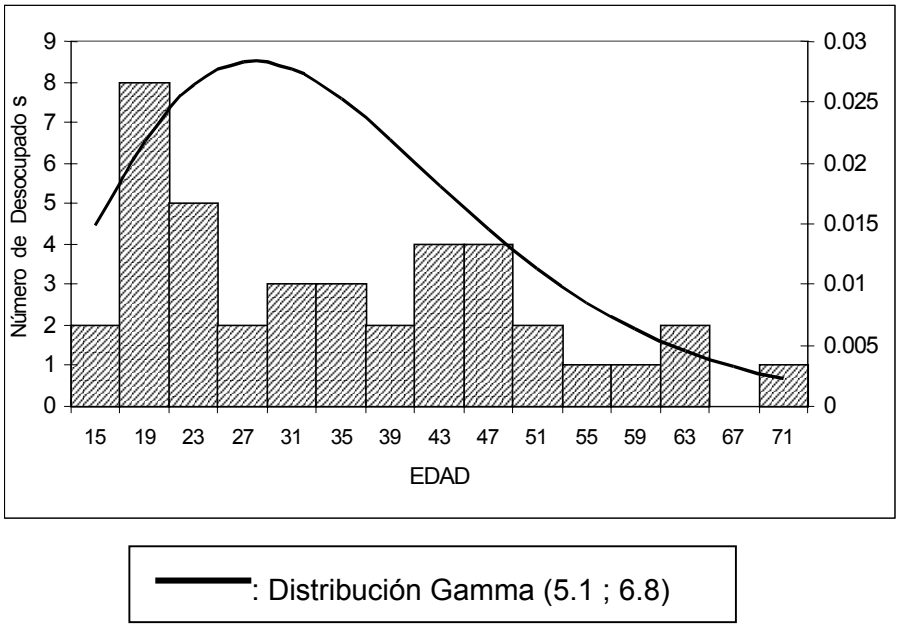
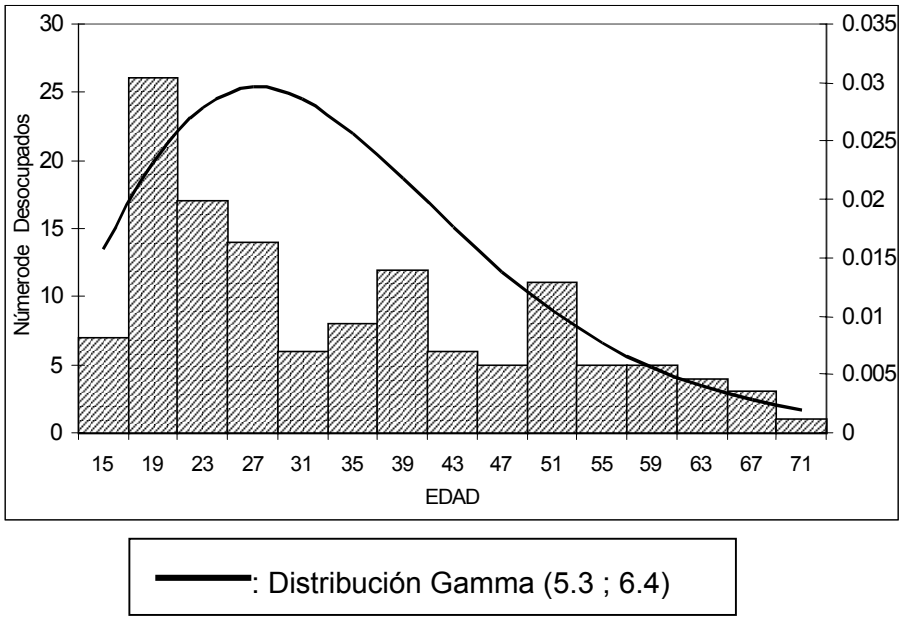


Gráfico 8. EDAD del grupo GIC. Onda Octubre 1998



La similitud entre los parámetros de cada distribución ajustada hace que la distribución de la EDAD en los grupos GIF y GIC para la onda Octubre 98 sean similares en forma y ubicación.

De los resultados obtenidos se concluye que para la onda Octubre 97, la distribución de la EDAD en el grupo GIF se encuentra desplazada con respecto al grupo GIC y las funciones de distribución propuestas difieren. Por lo tanto, la EDAD de los desocupados pertenecientes al grupo GIF no se comporta de igual forma con respecto al grupo GIC, o sea que las unidades del grupo GIF y GIC no provendrían de la misma población.

Para la onda Octubre 98, la distribución de la EDAD del grupo GIF no se encuentra desplazada con respecto al grupo GIC y las funciones de distribución propuestas resultan iguales. Para esta onda, la EDAD de los desocupados pertenecientes al grupo GIF se comporta de manera semejante al grupo GIC. Es decir que las unidades de ambos grupos provendrían de la misma población.

5. DISCUSIÓN

La aplicación de herramientas basadas en métodos a distribución libre para evaluar conjuntos de datos con información perdida permite, de manera ágil y sin la necesidad de realizar supuestos distribucionales, acceder a una forma de evaluación diferente del comportamiento de los grupos con información completa y faltante.

Se posibilita la comparación tanto de la ubicación como la postulación de posibles funciones de distribución de las cuales provendrían, orientando las conclusiones hacia la posibilidad de considerar la aleatoriedad de la pérdida.

El usuario de estas herramientas deberá evaluar la conveniencia de las mismas según la aplicación que se realice y ser cauteloso a la hora de extraer conclusiones ya que, en algunos casos, es posible encontrar evidencias poco claras de desplazamiento o más de una función de distribución que resulte no significativa.

6. BIBLIOGRAFÍA

HOLLANDER, M. AND WOLFE, D. A. (1999) “Nonparametric Statistical Methods”. 2° Edición. John Wiley & Sons.

LITTLE, R. J. AND RUBIN, D. B.(1987) “Statistical Analysis with Missing Data”. John Wiley & Sons.

BROS, L. ; DE LEEUW, E. ; HOOP, H.; KURVER, G. (1995) “Nonrespondents in a mail survey; who are they?”. International Perspectives on Nonresponse Laaksonen. Ed. Statistics Finland.

HAUPTMANN, P. (1995) ”Nonresponse: Who responds and who does not in an enterprise panel survey”. International Perspectives on Nonresponse Laaksonen. Ed. Statistics Finland.