Centro Universitario de Estudios Medioambientales.
Seminarios de la reunión semanal del CUEM.
Fecha: 2026-05-18
Expositor: Alfredo Rigalli
Tema: Big data 4. Buscando significancia real

No existe investigación si no existen datos. Un dato es un valor numérico, cualidad, resultado de análisis y prácticamente cualquier cosa surgida de la aplicación del proceso de medición y observación de un sistema. Recordemos que los pilares de un dato son tres. El contexto: un dato es acompañado de metadatos. La objetividad: un dato debe ser reproducible. El registro: el dato debe perdurar en algún soporte. Un dato es dato si permite u obliga a tomar una decisión. Los datos y los metadatos forman bases de datos. Las bases de datos pueden pertenecer a small data, si ocupan poco espacio de memoria en una PC. Medium data si ocupan más lugar y cuando el espacio ocupado llega a los Terabytes podemos decir que se trata de bigdata.
Para llegar a conclusiones o tomar decisiones a partir de datos debemos aplicar herramientas de la estadística. Habitualmente decimos que los datos nos indican algo significativo o diferente, en tal caso el pvalue<0.05 (o el valor que defina el investigador, lo habitual es 0,05).
Cuando aumenta el número de datos que se someten a algún proceso de inferencia estadística, la probabilidad de hallar diferencias estadísticamente significativas aumenta, aun cuando esta diferencia es muy pequeña. Si el número de datos es relativamente grande, podemos hallar que la inferencia estadística nos diga que algo es diferente de otra cosa, sin que esa diferencia sea relevante o importante para la toma de decisión. La toma de una u otra decisión se debe hacer si hay diferencias estadísticas entre los datos y esa diferencia es relevante o importante.
Para evaluar la relevancia de la diferencia existen diferentes procedimientos.
1- de de Cohen: El valor d puede tomar diferente valores. Si es menor de 0.2 podemos decir que es irrelevante, si es mayor que 0.2 y menor que 0.5 es un efecto menor, si es mayor que 0,5 pero menor que 0,8 es un efecto medio y si es mayor que 0,8 es un efecto importante o muy relevante.
2- overlapping o superposición de distribuciones. Si la superposición es mayor al 95%, la diferencia es irrelevante.
3- Sentido común.
Ejemplo 1: Comparamos los ingresos mensuales de dinero de dos grupos de personas nos da: p=0,001, d=0,95, overlapping= 50% Podemos decir que hay diferencias significativas entre los grupos y esa diferencia parece ser importante. Sin embargo cuando miro los valores veo que un grupo gana 800000$ por mes en promedio y el otro gana en promedio $820000, con dispersiones similares, EL SENTIDO COMÚN me indica que aunque la diferencia es estadísticamente significativa, esa diferencia no incidirá produciendo diferentes niveles de vida o de gastos en los grupos.
Ejemplo 2: Comparo el porcentaje de afectados por una dada enfermedad entre dos grupos de personas. Obtengo p=0,12, d=0,18 y overlapping=98% , estadísticamente la diferencia de enfermos entre dos grupos es no significativa e irrelevante. Sin embargo, si un grupo tiene 1% y el otro 2% de enfermos, el SENTIDO COMÚN me dice que en el grupo 2 hay el doble de enfermos, lo cual no es poco. Hay un 100% más de enfermos.