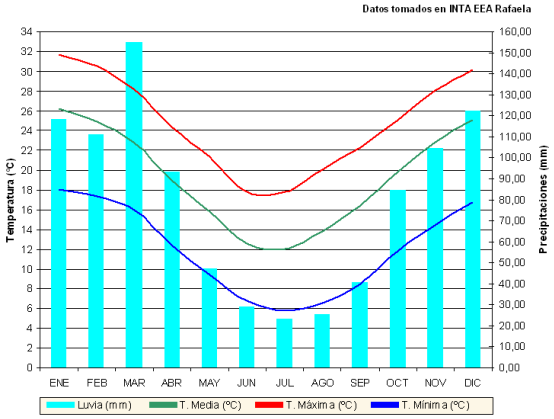
Estadística

Conceptos y herramientas



Deben tenerse en cuenta todas las formas de información pertinente, incluidos los conocimientos, las innovaciones y las prácticas de las comunidades científicas, indígenas y locales. Principio 11

Deben intervenir todos los sectores de la sociedad y las disciplinas científicas pertinentes. Principio 12

La información independientemente de lo costosa que haya sido crearla, puede ser replicada y compartida a un costo mínimo o nulo. -- Thomas Jefferson

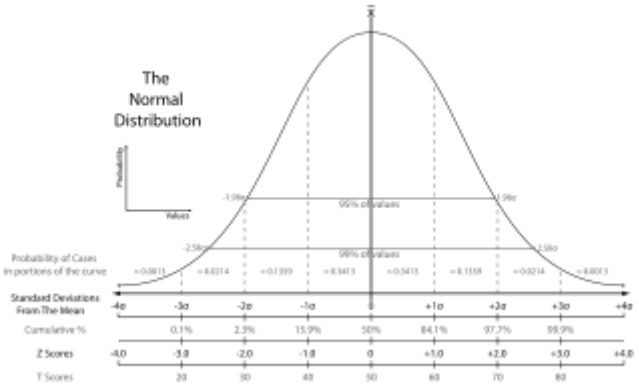
[Enero 2010]
Métodos de levantamiento y análisis de datos
Posgrado en Gestión de Áreas Protegidas y Desarrollo Ecorregional
UCI
Compilador J. Fallas
Jfallsa56@gmail.com

Tabla de contenido

1.ESTADÍSTICA

"El corazón jamás habla, pero hay que escucharlo para entender."
Proverbio Chino

La estadística es una ciencia con base matemática referente a la recolección, análisis e interpretación de datos, que busca explicar condiciones regulares en fenómenos de tipo aleatorio.



Distribución normal

Es transversal a una amplia variedad de disciplinas, desde la física hasta las ciencias sociales, desde las ciencias de la salud hasta el control de calidad. Se usa para la toma de decisiones en áreas de negocios o instituciones gubernamentales.

La estadística se divide en dos ramas:

La estadística descriptiva, que se dedica a los métodos de recolección, descripción, visualización y resumen de datos originados a partir de los fenómenos en estudio. Los datos pueden ser resumidos numérica o gráficamente. Ejemplos básicos de parámetros estadísticos son: la media y la desviación estándar. Algunos ejemplos gráficos son: histograma, pirámide poblacional, clústers, etc.

La inferencia estadística, que se dedica a la generación de los modelos, inferencias y predicciones asociadas a los fenómenos en cuestión teniendo en cuenta la aleatoriedad de las observaciones. Se usa para modelar patrones en los datos y extraer inferencias acerca de la población bajo estudio. Estas inferencias pueden tomar la forma de respuestas a preguntas si/no (prueba de hipótesis), estimaciones de características numéricas (estimación), pronósticos de futuras observaciones, descripciones de asociación (correlación) o modelamiento de relaciones entre variables (análisis de regresión). Otras técnicas de modelamiento incluyen anova, series de tiempo y minería de datos.

Ambas ramas (descriptiva e inferencial) comprenden la estadística aplicada. Hay también una disciplina llamada estadística matemática, la cual se refiere a las bases teóricas de la materia. La palabra «estadísticas» también se refiere al resultado de aplicar un algoritmo estadístico a un conjunto de datos, como en estadísticas económicas, estadísticas criminales, etc.

1.1 Etimología

La palabra «estadística» procede del latín *statísticum collégium* ('consejo de Estado') y de su derivado italiano *statista* ('hombre de Estado' o 'político'). El término alemán *statistik*, que fue primeramente introducido por Gottfried Achenwall (1749), designaba originalmente el análisis de datos del Estado, es decir, «la ciencia del Estado» (también llamada «aritmética política» de su traducción directa del inglés). No fue hasta el siglo XIX cuando el término «estadística» adquirió el significado de recolectar y clasificar datos. Este concepto fue introducido por el inglés John Sinclair.

En su origen, por tanto, la estadística estuvo asociada a datos, a ser utilizados por el gobierno y cuerpos administrativos (a menudo centralizados). La colección de datos acerca de estados y localidades continúa ampliamente a través de los servicios de estadística nacionales e internacionales. En particular, los censos suministran información regular acerca de la población.

Desde los comienzos de la civilización han existido maneras sencillas de estadística, pues ya se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para contar el número de personas, animales o ciertas cosas. Hacia el año 3000 a. C. los babilónicos usaban ya pequeñas tablillas de arcilla para recopilar datos en tablas sobre la producción agrícola y de los géneros vendidos o cambiados mediante trueque. Los egipcios analizaban los datos de la población y la renta del país mucho antes de construir las pirámides en el siglo XI a. C. Los libros bíblicos de *Números* y *Crónicas* incluyen, en algunas partes, trabajos de estadística. El primero contiene dos censos de la población de Israel y el segundo describe el bienestar material de las diversas tribus judías. En China existían registros numéricos similares con anterioridad al año 2000 a. C. Los griegos clásicos realizaban censos cuya información se utilizaba hacia el 594 a. C. para cobrar impuestos.

1.2 Orígenes en probabilidad

Los métodos estadístico-matemáticos emergieron desde la teoría de probabilidad, la cual data desde la correspondencia entre Blaise Pascal y Pierre de Fermat (1654). Christian Huygens (1657) da el primer tratamiento científico que se conoce a la materia. El *Ars coniectandi* (póstumo, 1713) de Jakob Bernoulli y la *Doctrina de posibilidades* (1718) de Abraham de Moivre estudiaron la materia como una rama de las matemáticas.[1] En la era moderna, el trabajo de Kolmogórov ha sido un pilar en la formulación del modelo fundamental de la Teoría de Probabilidades, el cual es usado a través de la estadística.

La teoría de errores se puede remontar a la *Ópera miscellánea* (póstuma, 1722) de Roger Cotes y al trabajo preparado por Thomas Simpson en 1755 (impreso en 1756) el cual aplica por primera vez la teoría de la discusión de errores de observación. La reimpresión (1757) de este trabajo incluye el axioma de que errores positivos y negativos son igualmente probables y que hay unos ciertos límites asignables dentro de los cuales se encuentran todos los errores; se describen errores continuos y una curva de probabilidad.

Pierre-Simon Laplace (1774) hace el primer intento de deducir una regla para la combinación de observaciones desde los principios de la teoría de probabilidades. Laplace representó la ley de probabilidades de errores mediante una curva y dedujo una fórmula para la media de tres observaciones. También, en 1871, obtiene la fórmula para la ley de facilidad del error (término introducido por Lagrange, 1744) pero con ecuaciones inmanejables. Daniel Bernoulli (1778) introduce el principio del máximo producto de las probabilidades de un sistema de errores concurrentes.

El método de mínimos cuadrados, el cual fue usado para minimizar los errores en mediciones, fue publicado independientemente por Adrien-Marie Legendre (1805), Robert Adrain (1808), y Carl Friedrich Gauss (1809). Gauss había usado el método en su famosa predicción de la localización del planeta enano Ceres en 1801. Pruebas adicionales fueron escritas por Laplace (1810, 1812), Gauss (1823), James Ivory (1825, 1826), Hagen (1837), Friedrich Bessel (1838), W.F. Donkin (1844, 1856), John Herschel (1850) y Morgan Crofton (1870). Otros contribuidores fueron Ellis (1844), Augustus De Morgan (1864), Glaisher (1872) y Giovanni Schiaparelli (1875). La fórmula de Peters para *r*, el probable error de una observación simple es bien conocido.

El siglo XIX incluye autores como Laplace, Silvestre Lacroix (1816), Littrow (1833), Richard Dedekind (1860), Helmert (1872), Hermann Laurent (1873), Liagre, Didion y Karl Pearson. Augustus De Morgan y George Boole mejoraron la presentación de la teoría. Adolphe Quetelet (1796-1874), fue otro importante fundador de la estadística y quien introdujo la noción del «hombre promedio» *(l'homme moyen)* como un medio de entender los fenómenos sociales complejos tales como tasas de criminalidad, tasas de matrimonio o tasas de suicidios.

1.3 Niveles de medición

Hay cuatro tipos de mediciones o escalas de medición en estadística. Los cuatro tipos de niveles de medición (nominal, ordinal, intervalo y razón) tienen diferentes grados de uso en la investigación estadística. Las medidas de razón, en donde un valor cero y distancias entre diferentes mediciones son definidas, dan la mayor flexibilidad en métodos estadísticos que pueden ser usados para analizar los datos. Las medidas de intervalo tienen distancias interpretables entre mediciones, pero un valor cero sin significado (como las mediciones de coeficiente intelectual o temperatura en grados Celsius). Las medidas ordinales tienen imprecisas diferencias entre valores consecutivos, pero un orden interpretable para sus valores. Las medidas nominales no tienen ningún rango interpretable entre sus valores.
La escala de medida nominal, puede considerarse la escala de nivel más bajo. Se trata de agrupar objetos en clases. La escala ordinal, por su parte, recurre a la propiedad de «orden» de los números. La escala de intervalos iguales está caracterizada por una unidad de medida común y constante. Es importante destacar que el punto cero en las escalas de intervalos iguales es arbitrario, y no refleja en ningún momento ausencia de la magnitud que estamos midiendo. Esta escala, además de poseer las características de la escala ordinal, permite determinar la magnitud de los intervalos (distancia) entre todos los elementos de la escala. La escala de coeficientes o Razones es el nivel de medida más elevado y se diferencia de las escalas de intervalos iguales únicamente por poseer un punto cero propio como origen; es decir que el valor cero de esta escala significa ausencia de la magnitud que estamos midiendo. Si se observa una carencia total de propiedad, se dispone de una unidad de medida para el efecto. A iguales diferencias entre los números asignados corresponden iguales diferencias en el grado de atributo presente en el objeto de estudio.

1.4 Estado actual

Durante el siglo XX, la creación de instrumentos precisos para asuntos de salud pública (epidemiología, bioestadística, etc.) y propósitos económicos y sociales (tasa de desempleo, econometría, etc.) necesitó de avances sustanciales en las prácticas estadísticas.
Hoy el uso de la estadística se ha extendido más allá de sus orígenes como un servicio al Estado o al gobierno. Personas y organizaciones usan la estadística para entender datos y tomar decisiones en ciencias naturales y sociales, medicina, negocios y otras áreas. La estadística es entendida generalmente no como un sub-área de las matemáticas sino como

una ciencia diferente «aliada». Muchas universidades tienen departamentos académicos de matemáticas y estadística separadamente. La estadística se enseña en departamentos tan diversos como psicología, educación y salud pública.

Al aplicar la estadística a un problema científico, industrial o social, se comienza con un proceso o población a ser estudiado. Esta puede ser la población de un país, de granos cristalizados en una roca o de bienes manufacturados por una fábrica en particular durante un periodo dado. También podría ser un proceso observado en varios instantes y los datos recogidos de esta manera constituyen una serie de tiempo.

Por razones prácticas, en lugar de compilar datos de una población entera, usualmente se estudia un subconjunto seleccionado de la población, llamado muestra. Datos acerca de la muestra son recogidos de manera observacional o experimental. Los datos son entonces analizados estadísticamente lo cual sigue dos propósitos: descripción e inferencia.

El concepto de correlación es particularmente valioso. Análisis estadísticos de un conjunto de datos puede revelar que dos variables (esto es, dos propiedades de la población bajo consideración) tienden a variar conjuntamente, como si hubiera una conexión entre ellas. Por ejemplo un estudio del ingreso anual y la edad de muerte entre personas podría resultar en que personas pobres tienden a tener vidas más cortas que personas de mayor ingreso. Las dos variables se dicen a ser correlacionadas. Sin embargo, no se pude inferir inmediatamente la existencia de una relación de causalidad entre las dos variables. El fenómeno correlacionado podría ser la causa de un tercero, previamente no considerado, llamado variable confundida.

Si la muestra es representativa de la población, inferencias y conclusiones hechas en la muestra pueden ser extendidas a la población completa. Un problema mayor es el de determinar que tan representativa es la muestra extraída. La estadística ofrece medidas para estimar y corregir por aleatoriedad en la muestra y en el proceso de recolección de los datos, así como métodos para diseñar experimentos robustos como primera medida, ver diseño experimental.

El concepto matemático fundamental empleado para entender la aleatoriedad es el de probabilidad. La estadística matemática (también llamada teoría estadística) es la rama de las matemáticas aplicadas que usa la teoría de probabilidades y el análisis matemático para examinar las bases teóricas de la estadística.

El uso de cualquier método estadístico es válido solo cuando el sistema o población bajo consideración satisface los supuestos matemáticos del método. El mal uso de la estadística puede producir serios errores en la descripción e interpretación, afectando las políticas sociales, la práctica médica y la calidad de estructuras tales como puentes y plantas de reacción nuclear.

Incluso cuando la estadística es correctamente aplicada, los resultados pueden ser difícilmente interpretados por un no experto. Por ejemplo, el significado estadístico de una tendencia en los datos, que mide el grado al cual la tendencia puede ser causada por una variación aleatoria en la muestra, puede no estar de acuerdo con el sentido intuitivo. El conjunto de habilidades estadísticas básicas (y el escepticismo) que una persona necesita para manejar información en el día a día se refiere como «cultura estadística».

1.5 Métodos estadísticos

1.5.1 Estudios experimentales y observacionales

Un objetivo común para un proyecto de investigación estadística es investigar la causalidad, y en particular extraer una conclusión en el efecto que algunos cambios en los valores de predictores o variables independientes tienen sobre una respuesta o variables dependientes. Hay dos grandes tipos de estudios estadísticos para estudiar causalidad: estudios experimentales y observacionales. En ambos tipos de estudios, el efecto de las diferencias de una variable independiente (o variables) en el comportamiento de una variable dependiente es observado. La diferencia entre los dos tipos es la forma en que el estudio es conducido. Cada uno de ellos puede ser muy efectivo.

Un estudio experimental implica tomar mediciones del sistema bajo estudio, manipular el sistema y luego tomar mediciones adicionales usando el mismo procedimiento para determinar si la manipulación ha modificado los valores de las mediciones. En contraste, un estudio observacional no necesita manipulación experimental. Por el contrario, los datos son recogidos y las correlaciones entre predictores y la respuesta son investigadas.

Un ejemplo de un estudio experimental es el famoso experimento de Hawthorne el cual pretendía probar cambios en el ambiente de trabajo en la planta Hawthorne de la Western Electric Company. Los investigadores estaban interesados en si al incrementar la iluminación en un ambiente de trabajo, la producción de los trabajadores aumentaba. Los investigadores primero midieron la productividad de la planta y luego modificaron la iluminación en un área de la planta para ver si cambios en la iluminación afectarían la productividad. La productividad mejoró bajo todas las condiciones experimentales. Sin embargo, el estudio fue muy criticado por errores en los procedimientos experimentales, específicamente la falta de un grupo control y seguimiento.

Un ejemplo de un estudio observacional es un estudio que explora la correlación entre fumar y el cáncer de pulmón. Este tipo de estudio normalmente usa una encuesta para recoger observaciones acerca del área de interés y luego produce un análisis estadístico. En este caso, los investigadores recogerían observaciones de fumadores y no fumadores y luego mirarían los casos de cáncer de pulmón en ambos grupos.

Los pasos básicos para un experimento son:

Planeamiento estadístico de la investigación, lo cual incluye encontrar fuentes de información, selección de material disponible en el área y consideraciones éticas para la investigación y el método propuesto. Se plantea un problema de estudio,

Diseñar el experimento concentrándose en el modelo y la interacción entre variables independientes y dependientes. Se realiza un muestreo consistente en la recolección de datos referentes al fenómeno o variable que deseamos estudiar. Se propone un modelo de probabilidad, cuyos parámetros se estiman mediante estadísticos a partir de los datos de muestreo. Sin embargo, se mantiene lo que se denominan «hipótesis sostenidas» (que no son sometidas a comprobación). Se valida el modelo comparándolo con lo que sucede en la realidad. Se utiliza métodos estadísticos conocidos como test de hipótesis o prueba de significación.Se producen estadísticas descriptivas.

Inferencia estadística. Se llega a un consenso acerca de qué dicen las observaciones acerca del mundo que observamos. Se utiliza el modelo validado para tomar decisiones o predecir acontecimientos futuros. Se produce un reporte final con los resultados del estudio.

Técnicas estadísticas

Algunas pruebas y procedimientos para investigación de observaciones bien conocidos son:
Prueba t de Student
Prueba de χ^2
Análisis de varianza (ANOVA)
U de Mann-Whitney
Análisis de regresión
Correlación
Iconografía de las correlaciones
Prueba de la diferencia menos significante de Fisher
Coeficiente de correlación producto momento de Pearson
Coeficiente de correlación de rangos de Spearman
Análisis factorial exploratorio
Análisis factorial confirmatorio

1.6 Disciplinas especializadas

Algunos campos de investigación usan la estadística tan extensamente que tienen terminología especializada. Estas disciplinas incluyen:

Física estadística
Estadística industrial
Estadística Espacial
Matemáticas Estadística
Estadística en Medicina
Estadística en Medicina Veterinaria y Zootecnia
Estadística en Nutrición
Estadística en Agronomía
Estadística en Planificación
Estadística en Investigación
Estadística en Restauración de Obras
Estadística en Literatura
Estadística en Astronomía
Estadística en la Antropología (Antropometría)
Estadística en Historia
Estadística militar
Geoestadística
Bioestadística
Estadísticas de Negocios
Estadística Computacional
Estadística en las Ciencias de la Salud
Investigación de Operaciones
Estadísticas de Consultoría
Estadística de la educación, la enseñanza, y la formación
Estadística en la comercialización o mercadotecnia
Cienciometría

Estadística del Medio Ambiente
Estadística en Epidemiología
Minería de datos (aplica estadística y reconocimiento de patrones para el conocimiento de datos)
Econometría (Estadística económica)
Estadística en Ingeniería
Geografía y Sistemas de información geográfica, más específicamente en Análisis espacial
Demografía
Estadística en psicología (Psicometría)
Calidad y productividad
Estadísticas sociales (para todas las ciencias sociales)
Cultura estadística
Encuestas por Muestreo
Análisis de procesos y quimiometría (para análisis de datos en química analítica e ingeniería química)
Confiabilidad estadística
Procesamiento de imágenes
Estadísticas Deportivas

La estadística es una herramienta básica en negocios y producción. Es usada para entender la variabilidad de sistemas de medición, control de procesos (como en control estadístico de procesos o SPC (CEP)), para compilar datos y para tomar decisiones. En estas aplicaciones es una herramienta clave, y probablemente la única herramienta disponible.

1.7 Computación estadística

El rápido y sostenido incremento en el poder de cálculo de la computación desde la segunda mitad del siglo XX ha tenido un sustancial impacto en la práctica de la ciencia estadística. Viejos modelos estadísticos fueron casi siempre de la clase de los modelos lineales. Ahora, complejos computadores junto con apropiados algoritmos numéricos, han causado un renacer del interés en modelos no lineales (especialmente redes neuronales y árboles de decisión) y la creación de nuevos tipos tales como modelos lineales generalizados y modelos multinivel.

El incremento en el poder computacional también ha llevado al crecimiento en popularidad de métodos intensivos computacionalmente basados en remuestreo, tales como tests de permutación y de bootstrap, mientras técnicas como el muestreo de Gibbs han hecho los métodos bayesianos más accesibles. La revolución en computadores tiene implicaciones en el futuro de la estadística, con un nuevo énfasis en estadísticas «experimentales» y «empíricas». Un gran número de paquetes estadísticos está ahora disponible para los investigadores. Los sistemas dinámicos y teoría del caos, desde hace una década, empezaron a interesar en la comunidad hispana, pues en la anglosajona de Estados Unidos estaba ya establecida la «conducta caótica en sistemas dinámicos no lineales» con 350 libros para 1997 y empezaban algunos trabajos en los campos de las ciencias sociales y en aplicaciones de la física. También se estaba contemplando su uso en analítica.

1.8 Críticas a la estadística

Hay una percepción general de que el conocimiento estadístico es intencionada y demasiado frecuentemente mal usado, encontrando maneras de interpretar los datos que sean favorables al presentador. Un dicho famoso, al parecer de Benjamin Disraeli,[2] es: «Hay tres tipos de mentiras: mentiras pequeñas, mentiras grandes y estadísticas». El popular libro *How to lie with*

statistics (‘cómo mentir con las estadísticas’) de Darrell Huff discute muchos casos de mal uso de la estadística, con énfasis en gráficas malintencionadas. Al escoger (o rechazar o modificar) una cierta muestra, los resultados pueden ser manipulados; eliminando outliers por ejemplo. Este puede ser el resultado de fraudes o sesgos intencionales por parte del investigador. Lawrence Lowell (decano de la Universidad de Harvard) escribió en 1909 que las estadísticas, «como algunos pasteles, son buenas si se sabe quién las hizo y se está seguro de los ingredientes».

Algunos estudios contradicen resultados obtenidos previamente, y la población comienza a dudar en la veracidad de tales estudios. Se podría leer que un estudio dice (por ejemplo) que «hacer X reduce la presión sanguínea», seguido por un estudio que dice que «hacer X no afecta la presión sanguínea», seguido por otro que dice que «hacer X incrementa la presión sanguínea». A menudo los estudios se hacen siguiendo diferentes metodologías, o estudios en muestras pequeñas que prometen resultados maravillosos que no son obtenibles en estudios de mayor tamaño. Sin embargo, muchos lectores no notan tales diferencias, y los medios de comunicación simplifican la información alrededor del estudio y la desconfianza del público comienza a crecer.

Sin embargo, las críticas más fuertes vienen del hecho que la aproximación de pruebas de hipótesis, ampliamente usada en muchos casos requeridos por ley o reglamentación, obligan una hipótesis a ser 'favorecida' (la hipótesis nula), y puede también exagerar la importancia de pequeñas diferencias en estudios grandes. Una diferencia que es altamente significativa puede ser de ninguna significancia práctica.

En los campos de la psicología y la medicina, especialmente con respecto a la aprobación de nuevas drogas por la Food and Drug Administration, críticas de la aproximación de prueba de hipótesis se han incrementado en los años recientes. Una respuesta ha sido un gran énfasis en el p-valor en vez de simplemente reportar si la hipótesis fue rechazada al nivel de significancia α dado. De nuevo, sin embargo, esto resume la evidencia para un efecto pero no el tamaño del efecto. Una posibilidad es reportar intervalos de confianza, puesto que estos indican el tamaño del efecto y la incertidumbre. Esto ayuda a interpretar los resultados, como el intervalo de confianza para un α dado indicando simultáneamente la significancia estadística y el efecto de tamaño.

El *p* valor y los intervalos de confianza son basados en los mismos cálculos fundamentales como aquellos para las correspondientes pruebas de hipótesis. Los resultados son presentados en un formato más detallado, en lugar del si-o-no de las pruebas de hipótesis y con la misma metodología estadística.

Una muy diferente aproximación es el uso de métodos bayesianos. Esta aproximación ha sido, sin embargo, también criticada.

El fuerte deseo de ver buenas drogas aprobadas y el de ver drogas peligrosas o de poco uso siendo rechazadas crea tensiones y conflictos (errores tipo I y II en el lenguaje de pruebas de hipótesis).

Se denomina **estadística aplicada** al área de la estadística que se ocupa de inferir resultados sobre una población a partir de una o varias muestras. Es la parte de la estadística que se aplica en cualquier otra rama externa a ella, como psicología, medicina, sociología, historia, biología, marketing, etc.

Los parámetros poblacionales son estimados mediante funciones denominadas "estimadores" o "estadísticos". La estimación de éstos, se hace basándose en la estimación estadística y puede ser puntual, por intervalos o de contraste de hipótesis. En una estimación puntual se obtiene un solo valor con una confianza nula, como cuando se dice que la estatura media de tal población es de 1,72m. En la estimación por intervalos, el nivel de confianza depende de la amplitud del intervalo, es cuando se afirma que el 95% de tal población mide menos de 1,96m. El contraste de hipótesis consiste en verificar estadísticamente si una suposición acerca de una población es cierta o falsa.

La estadística aplicada se apoya totalmente en la utilización de paquetes estadísticos que ayudan a resolver problemas de índole estadística, acortando dramáticamente los tiempos de resolución. Es por esto que en muchas facultades se enseña a utilizar estos programas estadísticos sin que, a veces, el alumno entienda, ni tenga la necesidad de entender cómo funcionan.

Notas

↑ Ver el trabajo de Ian Hacking en *The emergence of probability* para una historia del desarrollo del concepto de probabilidad matemática.
↑ Cf. *Damned lies and statistics: untangling numbers from the media, politicians, and activists*, del profesor Joel Best. Best atribuye este dicho a Disraeli, y no a Mark Twain u otros autores como se cree popularmente.

Bibliografía

Best, Joel (2001). *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists*. University of California Press. ISBN 0-520-21978-3.

Desrosières, Alain (2004). *The Politics of Large Numbers: A History of Statistical Reasoning*, Camille Naish (trad.), Harvard University Press. ISBN 0-674-68932-1.

Hacking, Ian (1990). *The Taming of Chance*. Cambridge University Press. ISBN 0-521-38884-8.
Lindley, D. V. (1985). *Making Decisions*, 2.ª edición edición, John Wiley & Sons. ISBN 0-471-90808-8.

Stigler, Stephen M. (1990). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press/Harvard University Press. ISBN 0-674-40341-X.

Tijms, Henk (2004). *Understanding Probability: Chance Rules in Everyday life*. Cambridge University Press. ISBN 0-521-83329-9.

Volle, Michel (1984). *Le métier de statisticien*, 2.ª ed. edición, Económica. ISBN 2-7178-0824-

2. ESTADÍSTICA DESCRIPTIVA

La **estadística descriptiva** es una parte de la estadística que se dedica a analizar y representar los datos. Este análisis es muy básico, pero estudio. Aunque hay tendencia a generalizar a toda la población las primeras conclusiones obtenidas tras un análisis descriptivo, su poder inferencial es mínimo y debería evitarse tal proceder. Otras ramas de la estadística se centran en el contraste de hipótesis y su generalización a la población.

Algunas de las técnicas empleadas en este primer análisis de los datos se enumeran más abajo en el listado de conceptos básicos. Básicamente, se lleva a cabo un estudio calculando una serie de medidas de tendencia central, para ver en qué medida los datos se agrupan o dispersan en torno a un valor central.

Metodología

Selección y determinación de la muestra.
Obtención de los datos.
Clasificación y organización de los datos.
Análisis descriptivo de los datos.
Representación gráfica de los datos.
Contraste de hipótesis, si procede.
Conclusiones.

Tabla de representación de los datos: Variable característica o suceso en la primera columna y sus frecuencias y porcentajes y acumulativas en las sucesivas columnas.

Representación gráfica: en los ejes de coordenadas: eje vertical para la variable y eje horizontal para frecuencias.

Todos estos elementos son opcionales. Las variables, características o sucesos, con sus correspondientes valores no están siempre presentes, aunque pueden expresarse como intervalos, tiempos, escalas, etc.

Ejemplos
Ejemplos de este tipo de análisis descriptivo pueden encontrarse en la prensa diaria, en la parte de información económico-social: series de tiempo, gráfica de barras, índices de precios, resultados de una encuesta y más elaborado, para más de una variable, en pirámide de edades, comparativas, etc.

Técnicas de análisis gráfico

En esta página web usted encontrará una galería que muestra los principales tipos de gráficos utilizados para describir series estadísticas. Los gráficos se organizan en orden alfabético y no por método de analítico.

http://www.itl.nist.gov/div898/handbook/eda/section3/eda33.htm

Lista de conceptos básicos

La siguiente lista recopila unos conceptos básicos con los que, todo aquel que se pretenda iniciar en las técnicas Estadísticas, debería estar familiarizado.

análisis de series temporales
censo
combinatoria
desviación estándar
diseño experimental
distribución binomial
distribución normal
distribución t
encuesta
error estadístico
estadística inferencial
estadístico
grados de libertad
histograma
media
mediana
moda
muestreo
muestra
parámetro estadístico
población
probabilidad
Prueba de χ^2
regresión estadística
rango
tabla de frecuencias
variable aleatoria
variable estadística
varianza

Bibliografía

Snedecor, G.W. and Cochran, W.G. 1980. Statistical methods. Seventh Ed. Iowa, The Iowa State University Press. 507p.

Sokal, R. R. and Rohlf, 1981. Biometry. W. H. Freeman and Company. USA. 859p.
Steel, R; Torrie, J. 1988. Bioestadística: Principios y procedimientos. Trad. Martínez, R.B. Edit. McGraw-Hill, México. 622 pp.

Zar, Jerrold H. 1999. Biostatistical Analysis. Prentice Hall, New Jersey, USA. 663p.

3. ESTADÍSTICA INFERENCIAL

La inferencia estadística o estadística inferencial es una parte de la Estadística que comprende los métodos y procedimientos para deducir propiedades (hacer inferencias) de una población, a partir de una pequeña parte de la misma (muestra).

La bondad de estas deducciones se mide en términos probabilísticos, es decir, toda inferencia se acompaña de su probabilidad de acierto.

La estadística inferencial comprende los siguientes temas:

Teoría de muestras.
Estimación de parámetros.
Contraste de hipótesis.
Diseño experimental.
Inferencia bayesiana.
Métodos no paramétricos

Método
Un estudio estadístico comprende los siguientes pasos:

Planteamiento del problema: Suele iniciarse con una fijación de objetivos o algunas preguntas como ¿cuál será la media de esta población respecto a tal característica?, ¿se parecen estas dos poblaciones?, ¿hay alguna relación entre... ?. En el planteamiento se definen con precisión la población, la característica a estudiar, las variables, etcétera. Se analizan también en este punto los medios de los que se dispone y el procedimiento a seguir.

Elaboración de un modelo: Se establece un modelo teórico de comportamiento de la variable de estudio. En ocasiones no es posible diseñar el modelo hasta realizar un estudio previo. Los posibles modelos son distribuciones de probabilidad.

Extracción de la muestra: Se usa alguna técnica de muestreo o un diseño experimental para obtener información de una pequeña parte de la población.

Tratamiento de los datos: En esta fase se eliminan posibles errores, se depura la muestra, se tabulan los datos y se calculan los valores que serán necesarios en pasos posteriores, como la media muestral, la varianza muestral Los métodos de esta etapa están definidos por la estadística descriptiva.

Estimación de los parámetros: Con determinadas técnicas se realiza una predicción sobre cuáles podrían ser los parámetros de la población.

Contraste de hipótesis: Los contrastes de hipótesis son técnicas que permiten simplificar el modelo matemático bajo análisis. Frecuentemente el contraste de hipótesis recurre al uso de estadísticos muestrales. Artículo principal: contraste de hipótesis

Conclusiones : Se critica el modelo y se hace un balance. Las conclusiones obtenidas en este punto pueden servir para tomar decisiones o hacer predicciones. El estudio puede comenzar de nuevo a partir de este momento, en un proceso cíclico que permite conocer cada vez mejor la población y características de estudio.

Temas relacionados
Contraste de hipótesis.
Diseño de experimentos.
Estimación.
Inferencia bayesiana.
Intervalo de confianza.
Teoría de muestras

Enlaces externos
Inferencia estadística según el modelo frecuentista, en la web de la Sociedad Andaluza de Enfermedades Infecciosas
Inferencia estadística según el modelo bayesiano, en la web de la Sociedad Andaluza de Enfermedades Infecciosas
http://es.wikipedia.org/wiki/Estad%C3%ADstica_inferencial

Referencis

Snedecor, G.W. and Cochran, W.G. 1980. Statistical methods. Seventh Ed. Iowa, The Iowa State University Press. 507p.

Sokal, R. R. and Rohlf, 1981. Biometry. W. H. Freeman and Company. USA. 859p.
Steel, R; Torrie, J. 1988. Bioestadística: Principios y procedimientos. Trad. Martínez, R.B. Edit. McGraw-Hill, México. 622 pp.

Zar, Jerrold H. 1999. Biostatistical Analysis. Prentice Hall, New Jersey, USA. 663p.

4. HIPÓTESIS (MÉTODO CIENTÍFICO)

Una **hipótesis** puede definirse como proposición cuya veracidad es provisionalmente asumida, como solución provisional (tentativa) para un problema dado o con algún otro propósito investigador. El nivel de verdad que se asume para una hipótesis dependerá de la medida en que los datos empíricos recogidos apoyen lo afirmado en la hipótesis. Esto es lo que se conoce como contrastación empírica de la hipótesis o bien **proceso de validación de la hipótesis**. Este proceso puede realizarse de uno o dos modos: mediante **confirmación** (para las hipótesis universales) o mediante **verificación** (para las hipótesis existenciales).

4.1Hipótesis de investigación

Toda hipótesis constituye, un juicio, una afirmación o una negación de algo. Sin embargo, es un juicio de carácter especial. Las hipótesis son proposiciones provisionales y exploratorias y, por tanto, su valor de veracidad o falsedad depende críticamente de las pruebas empíricas. En este sentido, la replicabilidad de los resultados es fundamental para confirmar una hipótesis como solución de un problema.

La hipótesis de investigación es el elemento que condiciona el diseño de la investigación y responde provisionalmente al problema, verdadero motor de la investigación. Como se ha dicho esta hipótesis es una aseveración que puede validarse estadísticamente. Una hipótesis explícita es la guía de la investigación, puesto que establece los límites, enfoca el problema y ayuda a organizar el pensamiento. Se establece una hipótesis cuando el conocimiento existente en el área permite formular predicciones razonables acerca de la relación de dos o más elementos o variables. Una hipótesis indica el tipo de relación que se espera encontrar; o sea: "existe relación entre a y b"; "el primer elemento es la causa del segundo"; "cuando se presenta esto, entonces sucede aquello", o bien, "cuando esto sí, aquello no". Debe existir una cuantificación determinada o una proporción matemática que permita su verificación estadística.[1]

4.2 Hipótesis en estadística inferencial

En general, en un trabajo de investigación se plantean dos hipótesis mutuamente excluyentes: la hipótesis nula o hipótesis de nulidad (H_0) y la hipótesis de investigación (H_i). Además, es posible plantear hipótesis alternas o hipótesis alternativas. El análisis estadístico de los datos servirá para determinar si se puede o no aceptar H_0. Cuando se rechaza H_0, significa que el factor estudiado ha influido significativamente en los resultados y es información relevante para apoyar la hipótesis de investigación planteada. Es muy importante tener presente que la hipótesis de investigación debe coincidir con la hipótesis alternativa. Plantear hipótesis de investigación que coincidan con Ho supondría una aplicación incorrecta del razonamiento estadístico.

4.3 Identificación de las variables

Toda hipótesis constituye, un juicio, una afirmación o una negación de algo. Sin embargo, es un juicio de carácter especial. Es un juicio científico, técnico o ideológico, en cuanto a su origen o esencia. Siendo así, toda hipótesis lleva implícita un valor, un significado, una solución específica al problema. Esta es la variable, o sea el valor que le damos a la hipótesis. La variable viene a ser el contenido de solución que le damos al problema de investigación

Variable independiente: El valor de verdad que se le da a una hipótesis en relación con la causa, se denomina variable independiente.

Variable dependiente: Denominamos de esta manera a las hipótesis cuando su valor de verdad hace referencia no a la causa, sino al efecto.

Variable interviniente: Será aquella cuyo contenido se refiere a un factor que ya no es causa, tampoco efecto, pero sí modifica las condiciones del problema investigado.

Ejemplos
En esta sección se proponen algunos ejemplos de las diferentes tipologías de hipótesis que pueden hacerse:

Hipótesis de investigación: La computadora con regulador trabaja 100% del tiempo sin fallar. La computadora que se utiliza sin regulador solamente trabaja 80% del tiempo sin fallar.

Hipótesis no direccional: Existe una diferencia entre el nivel de ansiedad de los niños con un coeficiente intelectual alto y aquellos con un coeficiente bajo.

Hipótesis direccional: Los niños con coeficientes intelectuales altos tendrán un nivel de ansiedad mayor que los niños con coeficientes intelectuales bajos.

Hipótesis nula: No existe diferencia en los niveles de ansiedad entre niños con coeficientes intelectuales altos y aquellos que tienen coeficientes intelectuales bajos.

Referencias
↑ Schmelkes, Corina (2007): "Supuestos o hipótesis", en *Manual para la presentación de anteproyectos e informes de investigación*, ed. Oxford, 2ª ed., pp. 37-40.

Snedecor, G.W. and Cochran, W.G. 1980. Statistical methods. Seventh Ed. Iowa, The Iowa State University Press. 507p.

Sokal, R. R. and Rohlf, 1981. Biometry. W. H. Freeman and Company. USA. 859p.
Steel, R; Torrie, J. 1988. Bioestadística: Principios y procedimientos. Trad. Martínez, R.B. Edit. McGraw-Hill, México. 622 pp.

Zar, Jerrold H. 1999. Biostatistical Analysis. Prentice Hall, New Jersey, USA. 663p.

5.CONTRASTE DE HIPÓTESIS

Un **contraste de hipótesis** (también denominado **test de hipótesis** o **prueba de significación**) es una metodología de inferencia estadística para juzgar si una propiedad que se supone cumple una población estadística es compatible con lo observado en una muestra de dicha población. Fue iniciada por Ronald Fisher y fundamentada posteriormente por Jerzy Neyman y Karl Pearson.

Mediante esta teoría, se aborda el problema estadístico considerando una hipótesis determinada H_0y una hipótesis alternativa H_1, y se intenta dirimir cuál de las dos es la hipótesis verdadera, tras aplicar el problema estadístico a un cierto número de experimentos. Está fuertemente asociada a los considerados errores de tipo I y II en estadística, que definen respectivamente, la posibilidad de tomar un suceso verdadero como falso, o uno falso como verdadero.

Existen diversos métodos para desarrollar dicho test, minimizando los errores de tipo I y II, y hallando por tanto con una determinada potencia, la hipótesis con mayor probabilidad de ser correcta. Los tipos más importantes son los test centrados, de hipótesis y alternativa simple, aleatorizados,... Dentro de los test no paramétricos, el más extendido es probablemente el test de Kolmogórov-Smirnov.

5.1Introducción

Si sospechamos que una moneda ha sido trucada para que se produzcan más caras que cruces al lanzarla al aire, podríamos realizar 30 lanzamientos, tomando nota del número de caras obtenidas. Si obtenemos un valor demasiado alto, por ejemplo 25 o más, consideraríamos que el resultado es poco compatible con la hipótesis de que la moneda no está trucada, y concluiríamos que las observaciones contradicen dicha hipótesis.

La aplicación de cálculos probabilísticos permite determinar a partir de qué valor debemos rechazar la hipótesis garantizando que la probabilidad de cometer un error es un valor conocido *a priori*. Las hipótesis pueden clasificarse en dos grupos, según:

Especifiquen un valor concreto o un intervalo para los parámetros del modelo.

Determinen el tipo de distribución de probabilidad que ha generado los datos.

Un ejemplo del primer grupo es la hipótesis de que la media de una variable es 10, y del segundo que la distribución de probabilidad es la distribución normal.

Aunque la metodología para realizar el contraste de hipótesis es análoga en ambos casos, distinguir ambos tipos de hipótesis es importante puesto que muchos problemas de contraste de hipótesis respecto a un parámetro son, en realidad, problemas de estimación, que tienen una respuesta complementaria dando un intervalo de confianza (o conjunto de intervalos de confianza) para dicho parámetro. Sin embargo, las hipótesis respecto a la forma de la distribución se suelen utilizar para validar un modelo estadístico para un fenómeno aleatorio que se está estudiando.

5.2 Planteamiento clásico del contraste de hipótesis

Se denomina hipótesis nula H_0 a la hipótesis que se desea contrastar. El nombre de “nula” indica que H_0 representa la hipótesis que mantendremos a no ser que los datos indiquen su falsedad, y puede entenderse, por tanto, en el sentido de “neutra”. La hipótesis H_0 nunca se considera probada, aunque puede ser rechazada por los datos. Por ejemplo, la hipótesis de que dos poblaciones tienen la misma media puede ser rechazada fácilmente cuando ambas difieren mucho, analizando muestras suficientemente grandes de ambas poblaciones, pero no puede ser *“demostrada”* mediante muestreo, puesto que siempre cabe la posibilidad de que las medias difieran en una cantidad δ lo suficientemente pequeña para que no pueda ser detectada, aunque la muestra sea muy grande.

A partir de una muestra de la población en estudio, se extrae un estadístico (esto es, una valor que es función de la muestra) cuya distribución de probabilidad esté relacionada con la hipótesis en estudio y sea conocida. Se toma entonces el conjunto de valores que es más improbable bajo la hipótesis como región de rechazo, esto es, el conjunto de valores para el que consideraremos que, si el valor del estadístico obtenido entra dentro de él, rechazaremos la hipótesis.

La probabilidad de que se obtenga un valor del estadístico que entre en la región de rechazo aún siendo cierta la hipótesis puede calcularse. De esta manera, se puede escoger dicha región de tal forma que la probabilidad de cometer este error sea suficientemente pequeña.

Siguiendo con el anterior ejemplo de la moneda trucada, la muestra de la población es el conjunto de los treinta lanzamientos a realizar, el estadístico escogido es el número total de caras obtenidas, y la región de rechazo está constituida por los números totales de caras iguales o superiores a 25. La probabilidad de cometer el error de admitir que la moneda está trucada a pesar de que no lo está es igual a la probabilidad binomial de tener 25 "éxitos" o más en una serie de 30 ensayos de Bernoulli con probabilidad de "éxito" 0.5 en cada uno, entonces: 0.0002, pues existe la posibilidad, aunque poco probable, que la muestra nos dé más de 25 caras sin haber sido la moneda trucada.

5.3 Enfoque actual de los contrastes de hipótesis

El enfoque actual considera siempre una hipótesis alternativa a la hipótesis nula. De manera explícita o implícita, la hipótesis nula, a la que se denota habitualmente por H_0, se enfrenta a otra hipótesis que denominaremos hipótesis alternativa y que se denota H_1. En los casos en los que no se especifica H_1 de manera explícita, podemos considerar que ha quedado definida implícitamente como “H_0 es falsa”.

Si por ejemplo deseamos comprobar la hipótesis de que dos distribuciones tienen la misma media, estamos implícitamente considerando como hipótesis alternativa “ambas poblaciones tienen distinta media”. Podemos, sin embargo considerar casos en los que H_1 no es la simple negación de H_0. Supongamos por ejemplo que sospechamos que en un juego de azar con un dado, este está trucado para obtener 6. Nuestra hipótesis nula podría ser “el dado no está trucado” que intentaremos contrastar, a partir de una muestra de lanzamientos realizados, contra la hipótesis alternativa “el dado ha sido trucado a favor del 6”. Cabría realizar otras hipótesis, pero, a los efectos del estudio que se pretende realizar, no se consideran relevantes.

Un test de hipótesis se entiende, en el enfoque moderno, como una función de la muestra, corrientemente basada en un estadístico. Supongamos que se tiene una muestra $X = (X_1, X_2, ...X_n)^t$ de una población en estudio y que se han formulado hipótesis sobre un parámetro θ relacionado con la distribución estadística de la población. Supongamos que se dispone de un estadístico *T*(*X*) cuya distribución con respecto a θ, $F_\theta(t)$ se conoce. Supongamos, también, que las hipótesis nula y alternativa tienen la siguiente formulación:

$$\begin{cases} H_0: \ \theta \in \Theta_0 \\ H_1: \ \theta \in \Theta_1 \end{cases}$$

Un **contraste**, **prueba** o **test** para dichas hipótesis sería una función de la muestra de la siguiente forma:

$$\phi(X) = \begin{cases} 1 \text{ si } T(X) \in \Omega \\ 0 \text{ si } T(X) \notin \Omega \end{cases}$$

Donde $\phi(X) = 1$ significa que debemos rechazar la hipótesis nula, H_0 (aceptar H_1) y $\phi(X) = 0$, que debemos aceptar H_0 (o que no hay evidencia estadística contra H_0). A Ω se la denomina región de rechazo. En esencia, para construir el test deseado, basta con escoger el estadístico del contraste *T*(*X*) y la región de rechazo Ω.
Se escoge Ω de tal manera que la probabilidad de que T(X) caiga en su interior sea baja cuando se da H_0.

5.4 Errores en el contraste

Una vez realizado el contraste de hipótesis, se habrá optado por una de las dos hipótesis, H_0 o H_1, y la decisión escogida coincidirá o no con la que en realidad es cierta. Se pueden dar los cuatro casos que se exponen en el siguiente cuadro:

	H_0 es cierta	H_1 es cierta
Se escogió H_0	No hay error	Error de tipo II
Se escogió H_1	Error de tipo I	No hay error

Si la probabilidad de cometer un error de tipo I está unívocamente determinada, su valor se suele denotar por la letra griega α, y en las mismas condiciones, se denota por β la probabilidad de cometer el error de tipo II, esto es:

$$P(\text{escoger } H_1 | H_0 \text{ es cierta}) = \alpha$$
$$P(\text{escoger } H_0 | H_1 \text{ es cierta}) = \beta$$

En este caso, se denomina **Potencia del contraste** al valor 1-β, esto es, a la probabilidad de escoger H_1 cuando esta es cierta

$$P(\text{escoger } H_1 | H_1 \text{ es cierta}) = 1 - \beta$$.

Cuando es necesario diseñar un contraste de hipótesis, sería deseable hacerlo de tal manera que las probabilidades de ambos tipos de error fueran tan pequeñas como fuera posible. Sin embargo, con una muestra de tamaño prefijado, disminuir la probabilidad del error de tipo I, α, conduce a incrementar la probabilidad del error de tipo II, β.

Usualmente, se diseñan los contrastes de tal manera que la probabilidad α sea el 5% (0,05), aunque a veces se usan el 10% (0,1) o 1% (0,01) para adoptar condiciones más relajadas o más estrictas. El recurso para aumentar la potencia del contraste, esto es, disminuir β, probabilidad de error de tipo II, es aumentar el tamaño muestral, lo que en la práctica conlleva un incremento de los costes del estudio que se quiere realizar.

5.5 Contraste más potente

El concepto de potencia nos permite valorar cual entre dos contrastes con la misma probabilidad de error de tipo I, α, es preferible. Si se trata de contrastar dos hipótesis sencillas sobre un parámetro desconocido, θ, del tipo:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$$

Se trata de escoger entre todos los contrastes posibles con α prefijado aquel que tiene mayor potencia, esto es, menor probabilidad β de incurrir en el error de tipo II.

En este caso el Lema de Neyman-Pearson garantiza la existencia de un contraste de máxima potencia y determina como construirlo.

5.5.1Contraste uniformemente más potente

En el caso de que las hipótesis sean **compuestas**, esto es, que no se limiten a especificar un único posible valor del parámetro, sino que sean del tipo:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases}$$

donde Θ_0y $\Theta_1$son conjuntos de varios posibles valores, las probabilidades α y β ya no están unívocamente determindas, sino que tomarán diferentes valores según los distintos valores posibles de θ. En este caso se dice que un contraste $\phi(X)$tiene tamaño α si
$\alpha = \max_{\theta \in \Theta_0} P_\theta(\phi(X) = 0)$
esto es, si la máxima probabilidad de cometer un error de tipo I cuando la hipótesis nula es cierta es α. En estas circunstancias, se puede considerar β como una función de θ, puesto que para cada posible valor de θ en la hipótesis alternativa se tendría una probabilidad distinta de cometer un error de tipo II. Se define entonces

$$\beta(\theta) = P_\theta(\phi(X) = 1) \quad \forall \theta \in \Theta_1$$

y, la función de potencia del contraste es entonces

$$\mathrm{Pot}(\theta) = 1 - \beta(\theta) \quad \forall\theta \in \Theta_1$$

esto es, la probabilidad de discriminar que la hipótesis alternativa es cierta para cada valor posible de θ dentro de los valores posibles de esta misma hipótesis.

Se dice que un contraste es **uniformemente más potente de tamaño α** cuando, para todo valor $\theta \in \Theta_1 \mathrm{Pot}(\theta)$ es mayor o igual que el de cualquier otro contraste del mismo tamaño. En resumen, se trata de un contraste que garantiza la máxima potencia para todos los valores de θ en la hipótesis alternativa.

Es claro que el caso del contraste uniformemente más potente para hipótesis compuestas exige el cumplimiento de condiciones más exigentes que en el caso del contraste más potente para hipótesis simples. Por ello, no existe un equivalente al Lema de Neyman-Pearson para el caso general.

Sin embargo, sí existen muchas condiciones en las que, cumpliéndose determinadas propiedades de las distribuciones de probabilidad implicadas y para ciertos tipos de hipótesis, se puede extender el Lema para obtener el contraste uniformemente más potente del tamaño que se desee.

5.6 Aplicaciones de los contrastes de hipótesis

Los contrastes de hipótesis, como la inferencia estadística en general, son herramientas de amplio uso en la ciencia en general. En particular, la moderna Filosofía de la ciencia desarrolla el concepto de falsabilidad de las teorías científicas basándose en los conceptos de la inferencia estadística en general y de los contrastes de hipótesis. En este contexto, cuando se desea optar entre dos posibles teorías científicas para un mismo fenómeno (dos hipótesis) se debe realizar un contraste estadístico a partir de los datos disponibles sobre el fenómeno que permitan optar por una u otra.

Las técnicas de contraste de hipótesis son también de amplia aplicación en muchos otros casos, como ensayos clínicos de nuevos medicamentos, control de calidad, encuestas.

5.7 Critica a hipótesis nula

En estadística, una hipótesis nula es una hipótesis construida para anular o refutar, con el objetivo de apoyar una hipótesis alternativa. Cuando se la utiliza, la hipótesis nula se presume verdadera hasta que una evidencia estadística en la forma de una prueba de hipótesis indique lo contrario. El uso de la hipótesis nula es polémico.

5.8 Ejemplos

Hipótesis nula para la distribución ji-cuadrado:
«Si este material genético segrega en proporciones mendelianas, no habrá diferencias entre las frecuencias observadas (Oi) y las frecuencias esperadas (Ei).»
Hipótesis nula para la distribución t de Student:
«Si la humedad no influye sobre el número de huevos por desove, no habrá diferencias entre las medias de esta variable para cada región.»
Temas relacionados

Estadística
Estadístico muestral
Intervalo de confianza
Muestreo
Prueba de Kolmogórov-Smirnov
Prueba de χ^2 (Chi-cuadrado)
Test de la t de Student o t-test
Enlaces externos
Inferencia estadística, apuntes del Departamento de Matemáticas de la Universidad de La Coruña

Referencias

Snedecor, G.W. and Cochran, W.G. 1980. Statistical methods. Seventh Ed. Iowa, The Iowa State University Press. 507p.

Sokal, R. R. and Rohlf, 1981. Biometry. W. H. Freeman and Company. USA. 859p.
Steel, R; Torrie, J. 1988. Bioestadística: Principios y procedimientos. Trad. Martínez, R.B. Edit. McGraw-Hill, México. 622 pp.

Zar, Jerrold H. 1999. Biostatistical Analysis. Prentice Hall, New Jersey, USA. 663p.

Bioestadística: métodos y aplicaciones. Francisca R�us D�az, Francisco Javier Barón Lopez, Elisa Sánchez Font y Luis Parras Guijosa. Universidad de Málaga. Disponible en http://www.bioestadistica.uma.es/baron/bioestadistica.pdf

Apuntes y vídeos de Bioestadística
Material multimedia http://www.bioestadistica.uma.es/baron/apuntes/
http://www.bioestadistica.uma.es/baron/apuntes/
http://www.seh-lelha.org/stat1.htm

6. ESTADÍSTICA MULTIVARIANTE

Los **métodos estadísticos multivariantes** y el análisis multivariante son herramientas estadísticas que estudian el comportamiento de tres o más variables al mismo tiempo. Se usan principalmente para buscar las variables menos representativas para poder eliminarlas, simplificando así modelos estadísticos en los que el número de variables sea un problema y para comprender la relación entre varios grupos de variables. Algunos de los métodos más conocidos y utilizados son la Regresión lineal y el Análisis discriminante.

Se pueden sintetizar dos objetivos claros:

Proporcionar métodos cuya finalidad es el estudio conjunto de datos multivariantes que el análisis estadístico uni y bidimensional es incapaz de conseguir.

Ayudar al analista o investigador a tomar decisiones óptimas en el contexto en el que se encuentre teniendo en cuenta la información disponible por el conjunto de datos analizado.
Existen diferentes modelos y métodos, cada uno con su tipo de análisis:

6.1Métodos de Dependencia

Un Estudio de la regresión nos permite averiguar hasta que punto una variable puede ser prevista conociendo otra. Se utiliza para intentar predecir el comportamiento de ciertas variables a partir de otras, como por ejemplo los beneficios de una película a partir del gasto en márketing y del gasto en producción.

El **análisis de correlación canónica** es un método de análisis multivariante desarrollado por Harold Hotelling. Su objetivo es buscar las relaciones que pueda haber entre dos grupos de variables y la validez de las mismas. Se diferencia del análisis de correlación múltiple en que éste sólo predice una variable dependiente a partir de múltiples independientes, mientras que la correlación canónica predice múltiples variables dependientes a partir de múltiples independientes. La correlación canónica es una correlación lineal y, por tanto, sólo busca relaciones lineales entre las variables.

Al diseñar el experimento hay que considerar el tamaño de la muestra ya que son necesarias un mínimo de observaciones por variable, para que el análisis pueda representar las correlaciones adecuadamente.

Finalmente, hay que interpretar las cargas canónicas para determinar la importancia de cada variable en la función canónica. Las cargas canónicas reflejan la varianza que la variable observada comparte con el valor teórico canónico.

Un análisis discriminante nos puede dar una función discriminante que puede ser utilizada para distinguir entre dos o más grupos, y de este modo tomar decisiones. El **análisis discriminante** es una técnica estadística multivariante cuya finalidad es analizar si existen diferencias significativas entre grupos de objetos respecto a un conjunto de variables medidas sobre los mismos. La naturaleza de las variables debe ser para el caso de la dependiente categórica y para la(s) independiente(s) cuantitativa.

En caso de que estas diferencias existan, intentará explicar en qué sentido se dan y proporcionar procedimientos de clasificación sistemática de nuevas observaciones de origen desconocido en uno de los grupos analizados.

Un análisis multivariante de la varianza (MANOVA), extendiendo el análisis de la varianza (ANOVA), cubre los casos en los que se conozca la existencia de más de una variable dependiente sin poderse simplificar más el modelo.

La regresión logística permite la elaboración de un análisis de regresión para estimar y probar la influencia de una variable sobre otra, cuando la variable dependiente o de respuesta es de tipo dicotómico.

6.2 Métodos de Interdependencia

El análisis de los componentes principales procura determinar un sistema más pequeño de variables que sinteticen el sistema original. En estadística, el análisis de componentes principales (en español ACP, en inglés, PCA) es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos. Intuitivamente la técnica sirve para determinar el número de factores subyacentes explicativos tras un conjunto de datos que expliquen la variabilidad de dichos datos.

Técnicamente, el PCA busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. PCA se emplea sobre todo en análisis exploratorio de datos y para construir modelos predictivos. PCA comporta el cálculo de la descomposición en autovalores de la matriz de covarianza, normalmente tras centrar los datos en la media de cada atributo.

El análisis clúster clasifica una muestra de entidades (individuos o variables) en un número pequeño de grupos de forma que las observaciones pertenecientes a un grupo sean muy similares entre sí y muy disimilares del resto. A diferencia del Análisis discriminante se desconoce el número y la composición de dichos grupos.

La Iconografía de las correlaciones. La iconografía de las correlaciones, uno de los métodos de análisis de datos, consiste en reemplazar una matriz de correlación por un esquema donde las correlaciones «notables» son representadas por uno trazo continuo (correlación positiva), o uno trazo punteado (correlación negativa).

A partir de un cuadro de datos (por ejemplo una hoja de cálculo) que contiene columnas («variables») y líneas («observaciones» de estas variables), la iconografía de las correlaciones elimina las «falsas buenas correlaciones» entre estas variables, esto es, las que son debidas a una variable tercera, y detecta las correlaciones «enmascaradas». El «esquema» final, que presenta solo los vínculos directos entre las variables cualitativas y\o cuantitativas, es un medio de percibir de una ojeada lo esencial, sobre una figura única, quitada las redundancias.

6.3 Métodos Estructurales

Analizan las relaciones existentes entre un grupo de variables representadas por sistemas de ecuaciones simultáneas en las que se suponen que algunas de ellas (denominadas constructos) se miden con error a partir de otras variables observables denominadas indicadores. Los modelos utilizados constan, por lo tanto, de dos partes: un modelo estructural que especifica las relaciones de dependencia existente entre las constructos latentes y un

modelo de medida que especifica como los indicadores se relacionan con sus correspondientes constructos.

Referencias

Esbensen. Kim H. *Multivariate Data Analysis -in practice (5th Edition*).

Sokal, R. R. and Rohlf, 1981. Biometry. W. H. Freeman and Company. USA. 859p.
Steel, R; Torrie, J. 1988. Bioestadística: Principios y procedimientos. Trad. Martínez, R.B. Edit. McGraw-Hill, México. 622 pp.

Zar, Jerrold H. 1999. Biostatistical Analysis. Prentice Hall, New Jersey, USA. 663p.

Bioestadística: métodos y aplicaciones. Francisca R�us D�az, Francisco Javier Barón Lopez, Elisa Sánchez Font y Luis Parras Guijosa. Universidad de Málaga. Disponible en http://www.bioestadistica.uma.es/baron/bioestadistica.pdf

Teknomo, Kardi. Discriminant Analysis Tutorial. http://people.revoledu.com/kardi/ tutorial/LDA/

Teknomo, Kardi. K-Means Clustering Tutorials. http:\\people.revoledu.com\kardi\ tutorial\kMean\

Teknomo, Kardi. Pictorial Introduction to Graph Theory. http:\\people.revoledu.com\kardi\ tutorial\GraphTheory\

7.I ICONOGRAFÍA DE LAS CORRELACIONES

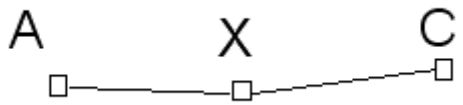
La iconografía de las correlaciones, uno de los métodos de análisis de datos, consiste en reemplazar una matriz de correlación por un *esquema* donde las correlaciones «notables» son representadas por uno trazo continuo (correlación positiva), o uno trazo punteado (correlación negativa).

A partir de un cuadro de datos (por ejemplo una hoja de cálculo) que contiene columnas («variables») y líneas («observaciones» de estas variables), la iconografía de las correlaciones elimina las «falsas buenas correlaciones» entre estas variables, esto es, las que son debidas a una variable tercera, y detecta las correlaciones «enmascaradas». El «esquema» final, que presenta solo los vínculos directos entre las variables cualitativas y\o cuantitativas, es un medio de percibir de una ojeada lo esencial, sobre una figura única, quitada las redundancias.

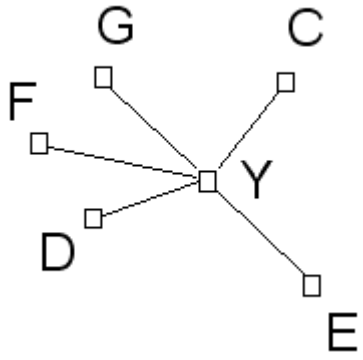
7.1¿Qué es una correlación «notable»?

Una correlación no tiene sentido aisladamente. Recíprocamente una correlación escasa no implica la ausencia de vínculo.

Ejemplo 1 : las variables A y C se correlacionan fuertemente porque sus variaciones son vinculadas las dos a una variable X. En realidad no hay vínculo AC, sino un vínculo XA y un vínculo XC. En otros términos, la correlación entra A y C es redundante, y desaparece, cuando X es mantenido constante (hablamos de correlación parcial escasa con relación a X). Lo deducimos el esquema de las solas correlaciones notables :



Ejemplo 2 : la variable Y depende de varias variables C, D, E, F y G independientes. También la correlación de Y con cada una de ellas, consideradas por separado, es escasa (no "significativa" con sentido probabilista del término). En realidad, existen unos vínculos rigurosos CY, DY, EY, FY y GY. Lo deducimos el esquema de las correlaciones notables :



Selección de los vínculos notables
Ilustrémosla sobre un pequeño ejemplo: en el momento de un control matemático de un nivel de clase de tercer año de bachillerato, ocho alumnos del primer año al último curso, cuyo peso, la edad y la asiduidad conocemos, obtuvieron las notas siguientes:

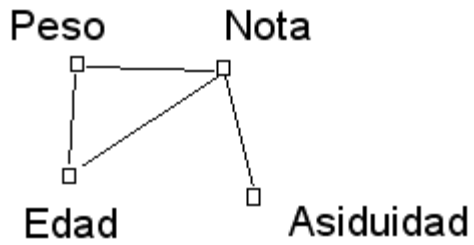
Datos originales

Alumno	Peso	Edad	Asiduidad	Nota
e1	52	12	12	5
e2	59	12,5	9	5
e3	55	13	15	9
e4	58	14,5	5	5
e5	66	15,5	11	13,5
e6	62	16	15	18
e7	63	17	12	18
e8	69	18	9	18

Matriz de correlación

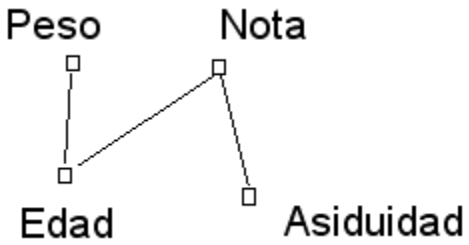
	Peso	Edad	Assiduidad	Nota
Peso	1			
Edad	0,885	1		
Asiduidad	-0,160	-0,059	1	
Nota	0,774	0,893	0,383	1

Coloquemos nuestras cuatro variables al azar sobre el papel, y tracemos uno trazo entre dos de ellas cada vez que su correlación es superior al *umbral* 0,3 en valor absoluto.



¡A la vista de este esquema, la correlación (peso, nota) = 0.774, relativamente fuerte, da a pensar que el peso tiene más influencia sobre la nota que la asiduidad! Pero, por otra parte, tenemos las correlaciones (peso, edad) = 0,885, y (edad, nota) = 0,893.

A partir de estos 3 coeficientes de *correlación total*, la fórmula de la correlación parcial da: *correlación (peso, nota) a edad constante* : = -0,08
¡La correlación entre nota y peso, a edad constante fuertemente bajó (es hasta ligeramente negativa)! De otro término el peso no tiene influencia sobre la nota. Borremos el vínculo entre peso y nota:



En definitiva, un vínculo no es trazado, sea porque su correlación total es inferior al umbral, en valor absoluto, sea porque existe por lo menos una correlación parcial inferior al umbral, en valor absoluto, o de signo contrario a la correlación total.

No es necesario, aquí, de borrar otros vínculos, como se lo verifica a partir de los valores de otras correlaciones parciales:

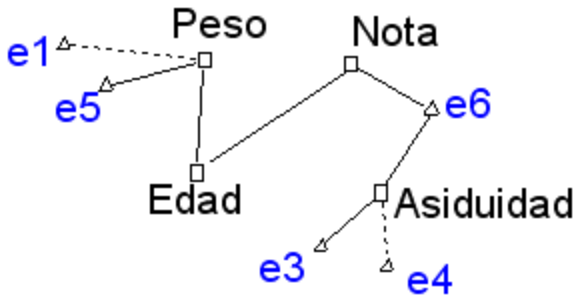
Correlación (peso, nota) a asiduidad constante = 0,92
Correlación (edad, peso) a nota constante = 0,68
Correlación (edad, peso) a asiduidad constante = 0,89
Correlación (edad, nota) a peso constante= 0,71
Correlación (asiduidad, peso) a nota constante = -0,78
Correlación (asiduidad, peso) a edad constante = -0,23
Correlación (asiduidad, nota) a peso constante = 0,81
Correlación (asiduidad, nota) a edad constante = 0,97
Correlación (asiduidad, edad) a peso constante = 0,18
Correlación (asiduidad, edad) a nota constante = -0,97

Instantes notables del análisis
Los datos disponibles permiten llevar más lejos el análisis.
Podemos considerar en efecto cada línea como un «instante» del análisis, caracterizado por una variable indicadora igual a 1 en el instante de la línea considerada, y a 0, en otro caso:

Alumno	Peso	Edad	Asiduidad	Nota	e1	e2	e3	e4	e5	e6	e7	e8
e1	52	12	12	5	1	0	0	0	0	0	0	0
e2	59	12,5	9	5	0	1	0	0	0	0	0	0
e3	55	13	15	9	0	0	1	0	0	0	0	0
e4	58	14,5	5	5	0	0	0	1	0	0	0	0
e5	66	15,5	11	13,5	0	0	0	0	1	0	0	0
e6	62	16	15	18	0	0	0	0	0	1	0	0
e7	63	17	12	18	0	0	0	0	0	0	1	0
e8	69	18	9	18	0	0	0	0	0	0	0	1

Aunque los «instantes» llevan los mismos nombres que los alumnos, hay que recordar que las alumnas son unas líneas (observaciones), mientras que los instantes son unas columnas, que forman parte de las «variables», con el mismo título que las 4 primeras columnas.

Podemos pues adoptar el mismo criterio de trazado de los vínculos para los "instantes" y las variables originales. No obstante, para no agravar el esquema, dibujemos solamente los «instantes» vinculados a una variable por lo menos («instantes notables»).



Los «instantes» son representados por un *triángulo*, por ser mejor distinguidos de las variables originales, que son representadas por un *cuadrado*.

Con relación al esquema precedente, el vínculo entre nota y asiduidad desapareció, reemplazó por los vínculos *(Nota, e6)* y *(Asiduidad, e6)* . Era pues redundante: el alumno e6, muy asiduo y bien anotado, le explica a solas el vínculo *(Nota, Asiduidad)* .

El alumno e3 tiene asiduidad notablemente fuerte, y el alumno e4 asiduidad notablemente escasa (trazo punteado).

Un vínculo es dicho «notable» cuando otros vínculos presentes sobre la figura no bastan con explicarlo.

El alumno e6 tiene en efecto una nota «notable»: 18/20.

Los alumnos e7 y e8 que tienen, también, 18/20, no son notables: no aparecen sobre el esquema, porque, más de edad, sus nota es ya explicada por el vínculo *(edad, anota)*.

Del mismo modo, podemos verificar sobre los datos, que e5 tiene un peso notablemente fuerte para su edad (con relación a los 8 alumnos de la población estudiada); mientras que el alumno e1 tiene un peso notablemente escaso para su edad.

Los vínculos entre cuadrados (variables - variables) subrayan las leyes generales; los vínculos cuadrado-triángulo (variable - instante) subrayan los acontecimientos raros.

Algoritmo de la iconografía de las correlaciones

El principio de la iconografía de las correlaciones es bastante simple para permitir un trazado manual, si el cuadro de datos es pequeño. Si no, hay que recurrir a un programa que contiene, en entrada, la matriz de correlación y el umbral escogido (por ejemplo 0,3). He aquí el algoritmo:

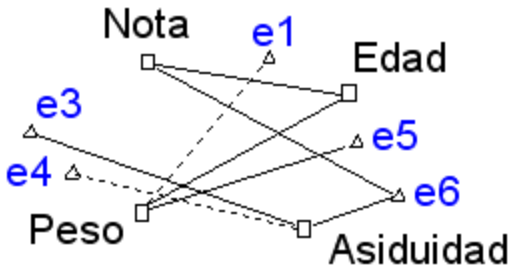
Para evitar las redundancias, el vínculo AB es trazado si y solamente si la correlación total r(A,B) es superior al umbral en valor absoluto, y si las correlaciones parciales r(A,B), con relación a una variable Z, son superiores al umbral, en valor absoluto, y con lo mismo signo que la correlación total, para todo Z entre las variables disponibles, incluido los «instantes».
Este criterio de trazado es estricto, y garantiza la selección de los vínculos *notables*.
Las variables pueden ser cuantitativas y\o cualitativas (siempre y cuando estas últimas utilicen una codificación disyuntiva completa).

7.2 Posición de los puntos sobre el papel

El ejemplo anterior mostró dos tipos de puntos: las variables (*cuadrados*), y los «instantes» (*triángulos*). Una vez encontrados los vínculos entre estos elementos, positivos (*trazos continuos*) o negativos (*trazos punteados*), solo queda dibujarlos sobre el papel.

Toda libertad de posicionamiento es dejada al analista, ya que la interpretación depende de vínculos y no de posiciones.

En lo posible, hay que evitar los cruces inútiles entre vínculos, molestando para la lectura. El esquema siguiente, por ejemplo, es menos legible que el precedente, aunque la interpretación sea la misma (vínculos idénticos):



Varias técnicas pueden ser utilizadas para colocar los puntos de modo automático.

Un primer enfoque consiste en proyectar la nube de puntos de las variables sobre los dos primeros ejes de un análisis de los componentes principales. Pero las proyecciones no son adaptadas siempre a una buena legibilidad cuando hay muchos componentes principales estadísticamente significativos, y particularmente en caso de mezcla de variables cualitativas y cuantitativas.

Otro enfoque consiste en sacar partido de la interpretación geométrica del coeficiente de correlación (coseno), y en dibujar el esquema a la superficie de una esfera a 3 dimensiones.
Al siendo el arco-coseno de la correlación una distancia angular, dos puntos serán tanto más próximos sobre la esfera cuanto serán correlacionados más (positivamente). A la inversa la distancia angular entre dos puntos que se correlacionan negativamente es un ángulo obtuso; si la correlación vale -1, los puntos son opuestos sobre la esfera (ángulo 180°).

Se trata, desde luego, de un mal menor, porque la esfera efectiva no está a 3 dimensiones, sino a *n* dimensiones. Si pues dos puntos que se correlacionan mucho forzosamente son próximos sobre el dibujo, lo inverso no está segura: dos puntos muy próximos sobre el dibujo no se correlacionan forzosamente. No obstante, la ausencia de vínculo trazado levanta la ambigüedad.

Podríamos contemplar muchos otros modos de elección de las posiciones: el más utilizado consiste en escoger como distancia angular el arco-coseno del valor absoluto de la correlación. Así, los puntos que se correlacionan negativamente no son opuestos sobre la esfera, y el vínculo punteado es más corto y atesta menos el esquema.

En práctica, en un enfoque software, una primera variable A es dibujada dondequiera sobre la esfera. Luego la variable B que se correlacionan menos a esta primera es puesta sobre la esfera a la distancia arco-coseno(r(A,B)) de la primera. Colocamos entonces, por triangulación, la variable C la menos correlacionada con ambas primeras. Otros puntos son

puestos poco a poco. Si la cuarta variable tiene una correlación nula con las tres primeras, no es materialmente posible asignarle una posición exacta. Las distancias son vueltas a calcular de modo proporcional a los valores efectivos. Al cabo de un cierto tiempo, la posición de los primeros puntos es vuelta a calcular según los siguientes. Etc. Así, la figura progresivamente es reajustada.

7.3 Elección del umbral

El umbral puede variar entre 0 y 1. Un vínculo es trazado si, no solamente la correlación total pero además todas las correlaciones parciales correspondientes son superiores al umbral en valor absoluto y del mismo signo. Esta condición es severa, y los vínculos que subsisten son ricos, en general, en información.

Aumentar el valor del umbral disminuye el número de vínculos, y clarifica la figura, pero disminuye también la información, sobre todo cuando la variable de interés depende de varias variables independientes.

Es a menudo preferible tomar un umbral bastante bajo. Luego, si la figura completa es demasiado prolija, se puede dibujar sólo los vínculos a la variable de interés.

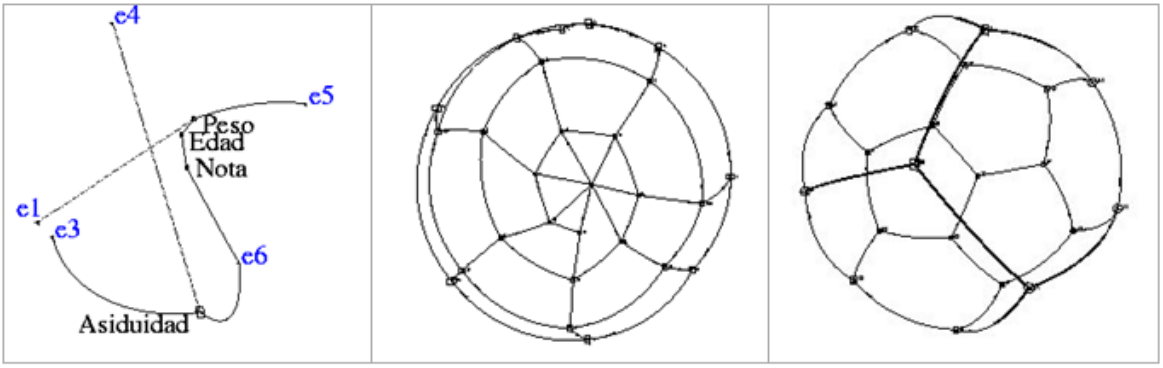
Por ejemplo, cuando se aborda nuevos datos, y cuando no se sabe cual umbral escoger, podremos comenzar por:

un umbral = 0.3 para un análisis de datos;
un umbral = 0.1 para el análisis de los resultados de un diseño de experimentos. En este caso en efecto, todos los factores son controlados, y podemos permitirnos no dibujar los "instantes" (a priori notables por construcción del plano), lo que alivia la figura;
un umbral = 0.01, o menos, podrá hasta ser escogido cuando la tabla de datos comprende varias centenas de observaciones.

En nuestro ejemplo, hasta el umbral nulo, el vínculo (peso, nota) no es trazado, porque la correlación parcial con relación a la edad está con signo contrario a la correlación total. Pero el vínculo (asiduidad, nota) aparece, y hay más instantes notables.

7.4Organización de los vínculos

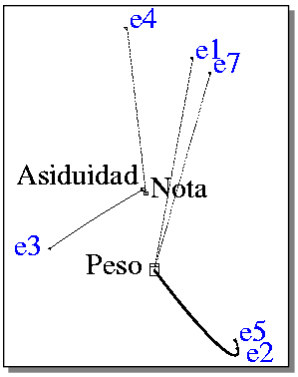
La Iconografía de las Correlaciones pretende poner en evidencia la organización de los vínculos, que puede ser cerrada tanto como jerárquica o continuamente repartida.



La ausencia de eje, cualquiera que sea la dimensión del problema permite reemplazar una multitud de proyecciones bidimensionales por una *imagen única*, o lo esencial aparece de una ojeada.

Una sucesión de tales figuras (eventualmente en forma de dibujo animado) autoriza la representación gráfica de una organización multidimensional evolutiva.

7.5 Retirada de una influencia evidente



Es común, en análisis de datos, disponer de una variable Z cuya influencia, preponderante, y ya bien conocida, enmascara fenómenos más finos que procuramos descubrir.

La solución consiste en trazar el esquema, no de la matriz de correlación total, pero de la matriz de las correlaciones parciales con relación a Z, con el fin de retirar toda influencia lineal de Z si existe allí (creciente o decreciente) sobre otras variables. El esquema revela entonces otra organización, abstracción hecha las variaciones de Z.

Por ejemplo, retiremos el componente de la edad, cuya influencia, preponderante, es bien conocida. El esquema revela entonces la influencia directa de la asiduidad sobre la nota. La edad desapareció de la figura, así como su componente en todas las variables. Y el peso se encuentra aislado.

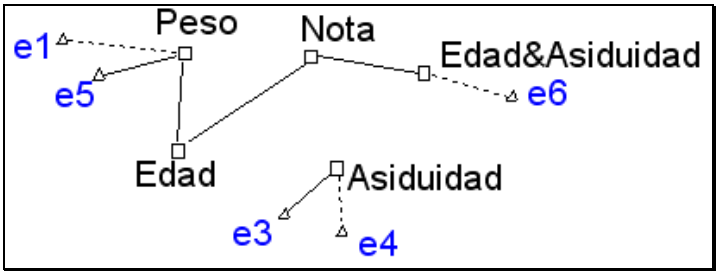
En una tabla de datos que contiene más variables puede ser interesante retirar varias influencias (el resultado no depende del orden en el cual son retirados).

7.6 Interacciones lógicas notables

Lo mismo que los «instantes» son añadidos, más arriba, a la tabla inicial, como de nuevas columnas, lo mismo, podemos añadir otras columnas, por ejemplo funciones de las variables iniciales, en particular las interacciones lógicas, que son unos acoplamientos de variables.

El número de columnas suplementarias importa poco, con tal que se añada sobre el esquema sólo a las que serán vinculadas a uno por lo menos variables iniciales, con el fin de no agravar inútilmente la figura.

Por ejemplo, en respuesta al añadido de nuevas columnas que corresponde a "y" lógica entre dos variables cualquiera, sólo la interacción «Edad&Asiduidad» directamente parece vinculada a la nota.



La interacción lógica aporta algo además a la interpretación (habida cuenta, por supuesto, del pequeño número de variables explicativas disponibles en este ejemplo): para obtener una buena nota no basta con tener mayor edad, hay que también ser asiduo a la clase.

7.7 Base de conocimiento asociada con esquema

Los vínculos del esquema pueden ser descritos de la manera siguiente: a cada vínculo trazado, asociemos una regla del tipo SI … ENTONCES, seguida por el valor del coeficiente de correlación total, precedido por uno «*» si el vínculo es trazado, y de «?» si el vínculo no es trazado, porque «*dudoso*» (el valor de la correlación es superior al umbral a causa de una sola observación).

SI Peso ENTONCES Edad *.885
SI Edad ENTONCES Peso *.885
SI Edad ENTONCES Nota *.893
SI Nota ENTONCES Edad *.893
SI Asiduidad ENTONCES Edad*Asiduidad ?.493
SI Nota ENTONCES Edad*Asiduidad *.960
SI Edad*Asiduidad ENTONCES Nota *.960
SI .e1 ENTONCES Peso *.610
SI .e3 ENTONCES Asiduidad *.484
SI .e4 ENTONCES Asiduidad *.726
SI .e5 ENTONCES Peso *.395
SI .e6 ENTONCES Edad*Asiduidad *.597

Los vínculos entre *variables* son indicados aquí en ambas direcciones, porque la causalidad no es directamente deducible de la correlación.

Los vínculos *«instantes notables» - variables* pueden ser indicadas en una sola dirección, porque la variable emana de su realización en el instante considerado.

Una base de conocimiento puede servir de entrada a un sistema experto; y el utilizador puede enriquecerlo o precisarlo.

Por ejemplo, es contrario al sentido común decir que la edad depende de una buena nota. No obstante lo inverso puede ser posible. Lo mismo, los niños engordan aumentando, pero no es el peso que hace el número de los años. El utilizador puede pues suprimir las reglas «SI Nota ENTONCES Edad .893», «SI Nota ENTONCES Edad *Asiduidad .960» y «SI Peso ENTONCES Edad .885».

La base de conocimiento así modificado da un esquema donde ciertos vínculos son orientados en lo sucesivo. Podemos aplicarle la Teoría de grafos y sacarlo flujos de informaciones.

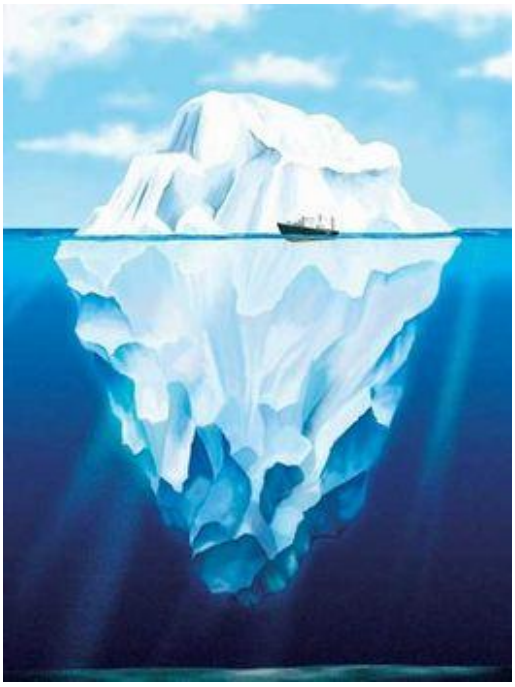
7.8 Campos de aplicación

El método se aplica en campos múltiples. Un medio de no olvidar nada de esencial en un cuadro de datos. Véase un ejemplo de aplicación a un gran cuadro de datos astronómicos difícil de aprehender de una ojeada.

Referencias

Lesty M. (1999) *Une nouvelle approche dans le choix des régresseurs de la régression multiple en présence d'interactions et de colinéarités*. La revue de Modulad, n°22, janvier 1999, pp. 41-77. (en francés).

8.MINERÍA DE DATOS



La minería de datos (DM, Data Mining) consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos.

Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos. Está fuertemente ligado con la supervisión de procesos industriales ya que resulta muy útil para aprovechar los datos almacenados en las bases de datos.

Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico. Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación.

8.1Procesos

Un proceso típico de minería de datos consta de los siguientes pasos generales:

Selección del conjunto de datos, tanto en lo que se refiere a las variables dependientes, como a las variables objetivo, como posiblemente al muestreo de los registros disponibles.

Análisis de las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).

Transformación del conjunto de datos de entrada, se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema.

Seleccionar y aplicar la técnica de minería de datos, se construye el modelo predictivo, de clasificación o segmentación.

Extracción de conocimiento, mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.

Interpretación y evaluación de datos, una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema.

Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Si el modelo final no superara esta evaluación el proceso se podría repetir desde el principio o, si el experto lo considera oportuno, a partir de cualquiera de los pasos anteriores. Esta retroalimentación se podrá repetir cuantas veces se considere necesario hasta obtener un modelo válido.

Una vez validado el modelo, si resulta ser aceptable (proporciona salidas adecuadas y/o con márgenes de error admisibles) éste ya está listo para su explotación. Los modelos obtenidos por técnicas de minería de datos se aplican incorporándolos en los sistemas de análisis de información de las organizaciones, e incluso, en los sistemas transaccionales. En este sentido cabe destacar los esfuerzos del Data Mining Group, que está estandarizando el lenguaje PMML (Predictive Model Markup Language), de manera que los modelos de minería de datos sean interoperables en distintas plataformas, con independencia del sistema con el que han sido construidos. Los principales fabricantes de sistemas de bases de datos y programas de análisis de la información hacen uso de este estándar.

Tradicionalmente, las técnicas de minería de datos se aplicaban sobre información contenida en almacenes de datos. De hecho, muchas grandes empresas e instituciones han creado y alimentan bases de datos especialmente diseñadas para proyectos de minería de datos en las que centralizan información potencialmente útil de todas sus áreas de negocio. No obstante, actualmente está cobrando una importancia cada vez mayor la minería de datos desestructurados como información contenida en ficheros de texto, en Internet, etc.

8.2Protocolo de un proyecto de minería de datos

Un proyecto de minería de datos tiene varias fases necesarias que son, esencialmente:

Comprensión del negocio y del problema que se quiere resolver.
Determinación, obtención y limpieza de los datos necesarios.
Creación de modelos matemáticos.
Validación, comunicación, etc. de los resultados obtenidos.
Integración, si procede, de los resultados en un sistema transaccional o similar.

La relación entre todas estas fases sólo es lineal sobre el papel. En realidad, es mucho más compleja y esconde toda una jerarquía de subfases. A través de la experiencia acumulada en proyectos de minería de datos se han ido desarrollando metodologías que permiten gestionar esta complejidad de una manera más o menos uniforme. Ejemplo de ella es CRISP-DM, se cree que SEMMA es una metodología SAS declara en su página que ésta NO es una metodología

8.3Técnicas de minería de datos

Como ya se ha comentado, las técnicas de la minería de datos provienen de la Inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados.

Las técnicas más representativas son:

1. Redes neuronales.- Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son:

A. El Perceptrón.
B. El Perceptrón multicapa.
C. Los Mapas Autoorganizados, también conocidos como redes de Kohonen.
D. Regresión lineal.- Es la mas utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.

2. Árboles de decisión.- Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema. Ejemplos:

A. Algoritmo ID3.
B. Algoritmo C4.5.

3. Modelos estadísticos.- Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.

4. Agrupamiento o Clustering.- Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Ejemplos:

A. Algoritmo K-means.
B. Algoritmo K-medoids.

Según el objetivo del análisis de los datos, los algoritmos utilizados se clasifican en supervisados y no supervisados (Weiss y Indurkhya, 1998):

Algoritmos supervisados (o predictivos): predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos.

Algoritmos no supervisados (o del descubrimiento del conocimiento): se descubren patrones y tendencias en los datos.

8.4 Ejemplos de uso de la minería de datos

8.4.1Negocios

La minería de datos puede contribuir significativamente en las aplicaciones de administración empresarial basada en la relación con el cliente. En lugar de contactar con el cliente de forma indiscriminada a través de un centro de llamadas o enviando cartas, sólo se contactará con aquellos que se perciba que tienen una mayor probabilidad de responder positivamente a una determinada oferta o promoción.

Por lo general, las empresas que emplean minería de datos ven rápidamente el retorno de la inversión, pero también reconocen que el número de modelos predictivos desarrollados puede crecer muy rápidamente.

En lugar de crear modelos para predecir qué clientes pueden cambiar, la empresa podría construir modelos separados para cada región y/o para cada tipo de cliente. También puede querer determinar qué clientes van a ser rentables durante una ventana de tiempo (una quincena, un mes, ...) y sólo enviar las ofertas a las personas que es probable que sean rentables. Para mantener esta cantidad de modelos, es necesario gestionar las versiones de cada modelo y pasar a una minería de datos lo más automatizada posible.

8.4.1.1Hábitos de compra en supermercados

El ejemplo clásico de aplicación de la minería de datos tiene que ver con la detección de **hábitos de compra en supermercados**. Un estudio muy citado detectó que los viernes había una cantidad inusualmente elevada de clientes que adquirían a la vez pañales y cerveza. Se detectó que se debía a que dicho día solían acudir al supermercado padres jóvenes cuya perspectiva para el fin de semana consistía en quedarse en casa cuidando de su hijo y viendo la televisión con una cerveza en la mano. El supermercado pudo incrementar sus ventas de cerveza colocándolas próximas a los pañales para fomentar las ventas *compulsivas*.

8.4.1.2Patrones de fuga

Un ejemplo más habitual es el de la detección de **patrones de fuga**. En muchas industrias —como la banca, las telecomunicaciones, etc.— existe un comprensible interés en detectar cuanto antes aquellos clientes que puedan estar pensando en rescindir sus contratos para, posiblemente, pasarse a la competencia. A estos clientes —y en función de su valor— se les podrían hacer ofertas personalizadas, ofrecer promociones especiales, etc., con el objetivo último de retenerlos. La minería de datos ayuda a determinar qué clientes son los más proclives a darse de baja estudiando sus patrones de comportamiento y comparándolos con muestras de clientes que, efectivamente, se dieron de baja en el pasado.

8.4.1.3Fraudes

Un caso análogo es el de la detección de transacciones de blanqueo de dinero o de **fraude** en el uso de tarjetas de crédito o de servicios de telefonía móvil e, incluso, en la relación de los contribuyentes con el fisco. Generalmente, estas operaciones fraudulentas o ilegales suelen seguir patrones característicos que permiten, con cierto grado de probabilidad, distinguirlas de las legítimas y desarrollar así mecanismos para tomar medidas rápidas frente a ellas.

8.4.1.4Recursos humanos

La minería de datos también puede ser útil para los departamentos de recursos humanos en la identificación de las características de sus empleados de mayor éxito. La información obtenida puede ayudar a la contratación de personal, centrándose en los esfuerzos de sus empleados y los resultados obtenidos por éstos. Además, la ayuda ofrecida por las aplicaciones para **Dirección estratégica** en una empresa se traducen en la obtención de ventajas a nivel corporativo, tales como mejorar el margen de beneficios o compartir objetivos; y en la mejora de las decisiones operativas, tales como desarrollo de planes de producción o gestión de mano de obra.

8.4.2Comportamiento en Internet

También es un área en boga el del análisis del comportamiento de los visitantes —sobre todo, cuando son clientes potenciales— en una página de **Internet**. O la utilización de la información —obtenida por medios más o menos legítimos— sobre ellos para ofrecerles propaganda adaptada específicamente a su perfil. O para, una vez que adquieren un determinado producto, saber inmediatamente qué otro ofrecerle teniendo en cuenta la información histórica disponible acerca de los clientes que han comprado el primero.

8.4.3Terrorismo

La minería de datos ha sido citada como el método por el cual la unidad Able Danger del Ejército de los EE.UU. había identificado al líder de los atentados del 11 de septiembre de 2001, Mohammed Atta, y a otros tres secuestradores del "**11-S**" como posibles miembros de una célula de Al Qaeda que operan en los EE.UU. más de un año antes del ataque. Se ha sugerido que tanto la Agencia Central de Inteligencia y sus homóloga canadiense, Servicio de Inteligencia y Seguridad Canadiense, también han empleado este método.[1]

8.4.4Juegos

Desde comienzos de la década de 1960, con la disponibilidad de oráculos para determinados juegos combinacionales, también llamados finales de juego de tablero (por ejemplo, para las tres en raya o en finales de ajedrez) con cualquier configuración de inicio, se ha abierto una nueva área en la minería de datos que consiste en la extracción de estrategias utilizadas por personas para estos oráculos. Los planteamientos actuales sobre reconocimiento de patrones, no parecen poder aplicarse con éxito al funcionamiento de estos oráculos. En su lugar, la producción de patrones *perspicaces* se basa en una amplia experimentación con bases de datos sobre esos finales de juego, combinado con un estudio intensivo de los propios **finales de juego** en problemas bien diseñados y con conocimiento de la técnica (datos previos sobre el final del juego). Ejemplos notables de investigadores que trabajan en este campo son Berlekamp en el juego de **puntos-y-cajas** (o Timbiriche) y John Nunn en finales de ajedrez.

8.4.5Ciencia e Ingeniería

En los últimos años la minería de datos se está utilizando ampliamente en diversas áreas relacionadas con la ciencia y la ingeniería. Algunos ejemplos de aplicación en estos campos son:

8.4.5.1Genética

En el estudio de la genética humana, el objetivo principal es entender la relación cartográfica entre las partes y la variación individual en las secuencias del ADN humano y la variabilidad en la susceptibilidad a las enfermedades. En términos más llanos, se trata de saber cómo los cambios en la secuencia de ADN de un individuo afectan al riesgo de desarrollar enfermedades comunes (como por ejemplo el cáncer). Esto es muy importante para ayudar a mejorar el diagnóstico, prevención y tratamiento de las enfermedades. La técnica de minería de datos que se utiliza para realizar esta tarea se conoce como "***reducción de dimensionalidad multifactorial***".[2]

8.4.5.2Ingeniería eléctrica

En el ámbito de la ingeniería eléctrica, las técnicas minería de datos han sido ampliamente utilizadas para monitorizar las condiciones de las instalaciones de alta tensión. La finalidad de esta monitorización es obtener información valiosa sobre el estado del aislamiento de los equipos. Para la vigilancia de las vibraciones o el análisis de los cambios de carga en transformadores se utilizan ciertas técnicas para agrupación de datos (***clustering***) tales como los Mapas Auto-Organizativos (**SOM**, *Self-organizing map*). Estos mapas sirven para detectar condiciones anormales y para estimar la naturaleza de dichas anomalías.[3]

8.4.5.3Análisis de gases

También se han aplicado técnicas de minería de datos para el análisis de gases disueltos (**DGA**, *Dissolved gas analysis*) en transformadores eléctricos. El análisis de gases disueltos se conoce desde hace mucho tiempo como herramienta para diagnosticar transformadores. Los Mapas Auto-Organizativos (**SOM**) se utilizan para analizar datos y determinar tendencias que podrían pasarse por alto utilizando las técnicas clásicas **DGA**.

8.5Minería de datos y otras disciplinas análogas

Suscita cierta polémica el definir las fronteras existentes entre la minería de datos y disciplinas análogas, como pueden serlo la estadística, la inteligencia artificial, etc. Hay quienes sostienen que la minería de datos no es sino **estadística envuelta en una jerga de negocios** que la conviertan en un producto *vendible*. Otros, en cambio, encuentran en ella una serie de **problemas y métodos específicos** que la hacen distinta de otras disciplinas.

El hecho es, que en la práctica la totalidad de los modelos y algoritmos de uso general en minería de datos —redes neuronales, árboles de regresión y clasificación, modelos logísticos, análisis de componentes principales, etc.— gozan de una tradición relativamente larga en otros campos.

8.5.1De la estadística

Ciertamente, la minería de datos bebe de la estadística, de la que toma las siguientes técnicas:

1. Análisis de varianza, mediante el cual se evalúa la existencia de diferencias significativas entre las medias de una o más variables continuas en poblaciones distintos.

2. Regresión: define la relación entre una o más variables y un conjunto de variables predictoras de las primeras.

3. Prueba chi-cuadrado: por medio de la cual se realiza el contraste la hipótesis de dependencia entre variables.

4. Análisis de agrupamiento o *clustering*: permite la clasificación de una población de *individuos* caracterizados por múltiples *atributos* (binarios, cualitativos o cuantitativos) en un número determinado de grupos, con base en las semejanzas o diferencias de los individuos.

5. Análisis discriminante: permite la clasificación de individuos en grupos que previamente se han establecido, permite encontrar la regla de clasificación de los elementos de estos grupos, y por tanto una mejor identificación de cuáles son las variables que definan la pertenencia al grupo.

6. Series de tiempo: permite el estudio de la evolución de una variable a través del tiempo para poder realizar predicciones, a partir de ese conocimiento y bajo el supuesto de que no van a producirse cambios estructurales.

8.5.2De la informática

De la informática toma las siguientes técnicas:

1. Algoritmos genéticos: Son métodos numéricos de optimización, en los que aquella variable o variables que se pretenden optimizar junto con las variables de estudio constituyen un segmento de información. Aquellas configuraciones de las variables de análisis que obtengan mejores valores para la variable de respuesta, corresponderán a segmentos con mayor capacidad reproductiva. A través de la reproducción, los mejores segmentos perduran y su proporción crece de generación en generación. Se puede además introducir elementos aleatorios para la modificación de las variables (mutaciones). Al cabo de cierto número de iteraciones, la población estará constituida por buenas soluciones al problema de optimización, pues las malas soluciones han ido descartándose, iteración tras iteración.

2. Inteligencia Artificial: Mediante un sistema informático que simula un sistema inteligente, se procede al análisis de los datos disponibles. Entre los sistemas de Inteligencia Artificial se encuadrarían los Sistemas Expertos y las Redes Neuronales.
3.
4. Sistemas Expertos: Son sistemas que han sido creados a partir de reglas prácticas extraídas del conocimiento de expertos. Principalmente a base de inferencias o de causa-efecto.

5. Sistemas Inteligentes: Son similares a los sistemas expertos, pero con mayor ventaja ante nuevas situaciones desconocidas para el experto.

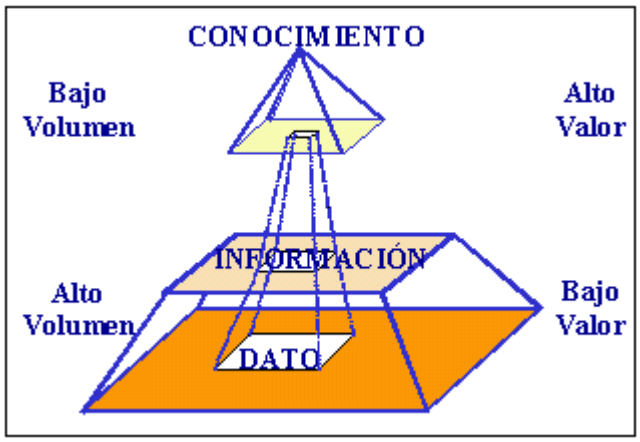
6. Redes neuronales: Genéricamente, son métodos de proceso numérico en paralelo, en el que las variables interactúan mediante transformaciones lineales o no lineales, hasta obtener unas salidas. Estas salidas se contrastan con los que tenían que haber salido, basándose en unos datos de prueba, dando lugar a un proceso de retroalimentación mediante el cual la red se reconfigura, hasta obtener un modelo adecuado.

8.6Minería de datos basada en teoría de la información

Todas las herramientas tradicionales de minería de datos *asumen* que los datos que usarán para construir los modelos contienen la información necesaria para lograr el propósito buscado: obtener suficiente conocimiento que pueda ser aplicado al *negocio* (o problema) para obtener un beneficio (o solución).

El inconveniente es que esto no es necesariamente cierto. Además, existe otro problema mayor aún. Una vez construido el modelo no es posible conocer si el mismo ha capturado toda

la información disponible en los datos. Por esta razón la práctica común es realizar varios modelos con distintos parámetros para ver si alguno logra mejores resultados.



Relación entre dato, información y conocimiento (Molina, 1998).

Un enfoque relativamente nuevo al análisis de datos soluciona estos problemas haciendo que la práctica de la minería de datos se parezca más a una ciencia que a un arte.

En 1948 Claude Shannon publicó un trabajo llamado “Una Teoría Matemática de la Comunicación”. Posteriormente esta pasó a llamarse Teoría de la Información y sentó las bases de la comunicación y la codificación de la información. Shannon propuso una manera de medir la cantidad de información a ser expresada en bits.

En 1999 Dorian Pyle publicó un libro llamado “Data Preparation for Data Mining” en el que propone una manera de usar la Teoría de la Información para analizar datos. En este nuevo enfoque, una base de datos es un canal que transmite información. Por un lado está el mundo real que captura datos generados por el negocio. Por el otro están todas las situaciones y problemas importantes del negocio. Y la información fluye desde el mundo real y a través de los datos, hasta la problemática del negocio.

Con esta perspectiva y usando la Teoría de la Información, es posible medir la cantidad de información disponible en los datos y qué porción de la misma podrá utilizarse para resolver la problemática del negocio. Como un ejemplo práctico, podría encontrarse que los datos contienen un 65% de la información necesaria para predecir qué cliente rescindirán sus contratos. De esta manera, si el modelo final es capaz de hacer predicciones con un 60% de acierto, se puede asegurar que la herramienta que generó el modelo hizo un buen trabajo capturando la información disponible. Ahora, si el modelo hubiese tenido un porcentaje de aciertos de solo el 10%, por ejemplo, entonces intentar otros modelos o incluso con otras herramientas podría valer la pena.

La capacidad de medir información contenida en los datos tiene otras ventajas importantes.
Al analizar los datos desde esta nueva perspectiva se genera un mapa de información que hace innecesario la preparación previa de los datos, una tarea absolutamente imprescindible si se desea buenos resultados, pero que lleva enorme cantidad de tiempo.

Es posible seleccionar un grupo de variables óptimo que contenga la información necesaria para realizar un modelo de predicción.

Una vez que las variables son procesadas con el fin de crear el mapa de información y luego seleccionadas aquellas que aportan la mayor información, la elección de la herramienta que se usará para crear el modelo deja de tener importancia, ya que el mayor trabajo fue realizado en los pasos previos.

8.7Tendencias

La Minería de Datos ha sufrido transformaciones en los últimos años de acuerdo con cambios tecnológicos, de estrategias de marketing, la extensión de los modelos de compra en línea, etc. Los más importantes de ellos son:

1. La importancia que han cobrado los **datos no estructurados** (texto, páginas de Internet, etc.).

2. La **necesidad de integrar** los algoritmos y resultados obtenidos en sistemas operacionales, portales de Internet, etc.

3. La exigencia de que los procesos funcionen prácticamente **en línea** (por ejemplo, que frente a un fraude con una tarjeta de crédito).

4. Los **tiempos de respuesta**. El gran volumen de datos que hay que procesar en muchos casos para obtener un modelo válido es un inconveniente; esto implica grandes cantidades de tiempo de proceso y hay problemas que requieren una respuesta en tiempo real.

8.8Herramientas de software

Existen muchas herramientas de software para el desarrollo de modelos de minería de datos tanto libres como comerciales como, por ejemplo:

- R
- KNIME
- SPSS Clementine (software)
- SAS Enterprise Miner
- RapidMiner
- Weka, http://www.cs.waikato.ac.nz/ml/weka/
- KXEN
- Orange

8.9Referencias

1. ↑ Stephen Haag et al.. *Management Information Systems for the information age*, pp. 28. ISBN 0-07-095569-7.
2. ↑ Xingquan Zhu, Ian Davidson (2007). *Knowledge Discovery and Data Mining: Challenges and Realities*. Hershey, New Your, pp. 18. ISBN 978-1-59904-252-7.
3. ↑ Plantilla:Cite Journal.

Otras fuentes

http://www.lsi.us.es/redmidas/ Red Española de minería de datos

MOLINA, L.C. (1998). *Data mining no processo de extração de conhecimento de bases de dados*. Tesis de máster. São Carlos (Brasil): Instituto de Ciências Matemáticas e Computação. Universidad de São Paulo.

MOLINA, L.C.; RIBEIRO, S. (2001). "Descubrimiento conocimiento para el mejoramiento bovino usando técnicas de data mining". En: *Actas del IV Congreso Catalán de Inteligencia Artificial*. Barcelona, pág. 123-130.

Cómo diseñar grandes variables en bases de datos multidimensionales. Manuel de la Herrán Gascón. ESIDE, Universidad de Deusto y Vicent Castellar. Departament de Matemàtica Aplicada Universitat de València. http://www.uv.es/buso/gv/gv.html

9.PITFALLS OF DATA ANALYSIS (OR HOW TO AVOID LIES AND DAMNED LIES)

Clay Helberg, M.S.
Research Design and Statistics Unit
University of Wisconsin Schools of Nursing and Medicine
600 Highland Ave. K6/380
Madison, WI 53792

ABSTRACT

There seems to be a pervasive notion that "you can prove anything with statistics." This is only true if you use them improperly. In this workshop we'll discuss things that people often overlook in their data analysis, and ways people sometimes "bend the rules" of statistics to support their viewpoint. We will also discuss ways you can make sure your own statistics are clear and accurate. I will include examples from medicine, education, and industry.

This paper presents material covered in a workshop at the Third International Applied Statistics in Industry Conference in Dallas, TX, June 5-7, 1995. The hypertext version of the workshop is available on the World-Wide Web (WWW) at the following location:
http://www.execpc.com/~helberg/pitfalls/

9.1The problem with statistics

We are all familiar with the disparaging quotes about statistics (including "There are three kinds of lies: lies, damned lies, and statistics", attributed to either Mark Twain or Disraeli, depending on whom you ask), and it's no secret that many people harbor a vague distrust of statistics as commonly used. Why should this be the case? It may be assumed that those of us at this conference take our work seriously and value the fruits of our efforts. So, are all those people just paranoid about statistics, or are we as statisticians really kidding ourselves as to our importance in the cosmic scheme of things?

It may be helpful to consider some aspects of statistical thought which might lead many people to be distrustful of it. First of all, statistics requires the ability to consider things from a probabilistic perspective, employing quantitative technical concepts such as "confidence", "reliability", "significance". This is in contrast to the way non-mathematicians often cast problems: logical, concrete, often dichotomous conceptualizations are the norm: right or wrong, large or small, this or that.

Additionally, many non-mathematicians hold quantitative data in a sort of awe. They have been lead to believe that numbers are, or at least should be, unquestionably correct. Consider the sort of math problems people are exposed to in secondary school, and even in introductory college math courses: there is a clearly defined method for finding the answer, and that answer is the only acceptable one. It comes, then, as a shock that different research studies can produce very different, often contradictory results. If the statistical methods used are really supposed to represent reality, how can it be that different studies produce different results? In order to resolve this paradox, many naive observers conclude that statistics must not really provide reliable (in the nontechnical sense) indicators of reality after all. And, the logic goes, if statistics aren't "right", they must be "wrong". It is easy to see how even intelligent, well-

educated people can become cynical if they don't understand the subtleties of statistical reasoning and analysis.

Now, I'm not going to say much about this "public relations crisis" directly, but it does provide a motivation for examining the way we practice our trade. The best thing we can do, in the long run, is make sure we're using our tools properly, and that our conclusions are warranted. I will present some of the most frequent misuses and abuses of statistical methods, and how to avoid or remedy them. Of course, these issues will be familiar to most statisticians; however, they are the sorts of things that can get easily overlooked when the pressure is on to produce results and meet deadlines. If this workshop helps you to apply the basics of statistical reasoning to improve the quality of your product, it will have served its purpose.

We can consider three broad classes of statistical pitfalls. The first involves sources of bias. These are conditions or circumstances which affect the external validity of statistical results. The second category is errors in methodology, which can lead to inaccurate or invalid results. The third class of problems concerns interpretation of results, or how statistical results are applied (or misapplied) to real world issues.

9.2Sources of Bias

The core value of statistical methodology is its ability to assist one in making inferences about a large group (a population) based on observations of a smaller subset of that group (a sample). In order for this to work correctly, a couple of things have to be true: the sample must be similar to the target population in all relevant aspects; and certain aspects of the measured variables must conform to assumptions which underlie the statistical procedures to be applied.

Representative sampling. This is one of the most fundamental tenets of inferential statistics: the observed sample must be representative of the target population in order for inferences to be valid. Of course, the problem comes in applying this principle to real situations. The ideal scenario would be where the sample is chosen by selecting members of the population at random, with each member having an equal probability of being selected for the sample. Barring this, one usually tries to be sure that the sample "parallels" the population with respect to certain key characteristics which are thought to be important to the investigation at hand, as with a stratified sampling procedure.

While this may be feasible for certain manufacturing processes, it is much more problematic for studying people. For instance, consider the construction of a job applicant screening instrument: the population about which you want to know something is the pool of all possible job applicants. You surely won't have access to the entire population--you only have access to a certain number of applicants who apply within a certain period of time. So you must hope that the group you happen to pick isn't somehow different from the target population. An example of a problematic sample would be if the instrument were developed during an economic recession; it is reasonable to assume that people applying for jobs during a recession might be different as a group from those applying during a period of economic growth (even if one can't specify exactly what those differences might be). In this case, you'd want to exercise caution when using the instrument during better economic times.

There are also ways to account for, or "control", differences between groups statistically, as with the inclusion of covariates in a linear model. Unfortunately, as Levin (1985) points out, there are problems with this approach, too. One can never be sure one has accounted for all

the important variables, and inclusion of such controls depends on certain assumptions which may or may not be satisfied in a given situation (see below for more on assumptions).

Statistical assumptions. The validity of a statistical procedure depends on certain assumptions it makes about various aspects of the problem. For instance, well-known linear methods such as analysis of variance (ANOVA) depends on the assumption of normality and independence. The first of these is probably the lesser concern, since there is evidence that the most common ANOVA designs are relatively insensitive to moderate violations of the normality assumption (see Kirk, 1982). Unfortunately, this offers an almost irresistible temptation to ignore *any* non-normality, no matter how bad the situation is. The robustness of statistical techniques only goes so far--"robustness" is not a license to ignore the assumption. If the distributions are non-normal, try to figure out why; if it's due to a measurement artifact (e.g. a floor or ceiling effect), try to develop a better measurement device (if possible). Another possible method for dealing with unusual distributions is to apply a transformation. However, this has dangers as well; an ill-considered transformation can do more harm than good in terms of interpretability of results.

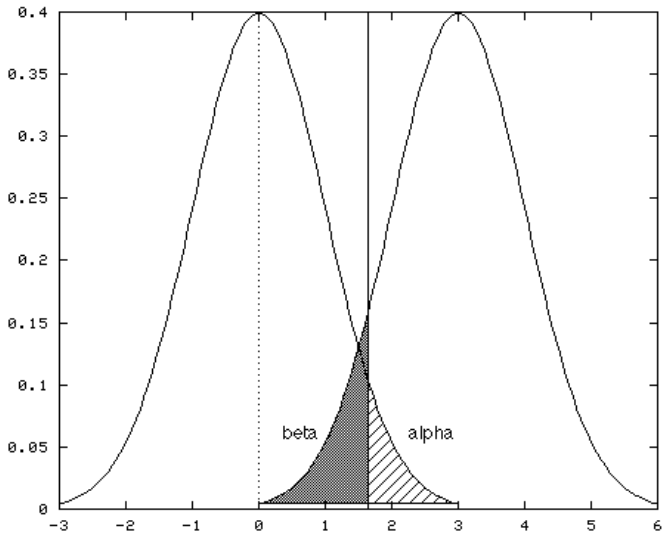
The assumption regarding independence of observations is more troublesome, both because it underlies nearly all of the most commonly used statistical procedures, and because it is so frequently violated in practice. Observations which are linked in some way--parts manufactured on the same machine, students in the same classroom, consumers at the same mall--all may show some dependencies. Therefore, if you apply some statistical test across students in different classrooms, say to assess the relationship between different textbook types and test scores, you're introducing bias into your results. This occurs because, in our example, the kids in the class presumably interact with each other, chat, talk about the new books they're using, and so influence each other's responses to the test. This will cause the results of your statistical test (e.g. correlations or p-values) to be inaccurate.

One way to try to get around this is to aggregate cases to the higher level, e.g. use classrooms as the unit of analysis, rather than students. Unfortunately this requires sacrificing a lot of statistical power, making a Type II error more likely. Happily, methods have been developed recently which allow simultaneous modeling of data which is hierarchically organized (as in our example with students nested within classrooms). One of the papers presented at this conference (Christiansen & Morris) introduces these methods. Additionally, interested readers are referred to Bryk & Raudenbush (1988) and Goldstein (1987) for good overviews of these hierarchical models.

9.3 Errors in methodology

There are a number of ways that statistical techniques can be misapplied to problems in the real world. Three of the most common hazards are designing experiments with insufficient power,ignoring measurement error, and performing multiple comparisons.

Statistical Power. This topic has become quite in vogue lately, at least in the academic community; indeed, some federal funding agencies seem to consider any research proposal incomplete unless it contains a comprehensive power analysis. This graph will help illustrate the concept of power in an experiment. In the figure, the vertical dotted line represents the point-null hypothesis, and the solid vertical line represents a criterion of significance, i.e. the point at which you claim a difference is significant.



Recall that there are two types of errors which can occur when making inferences based on a statistical hypothesis test: a Type I error occurs if you reject the null hypothesis when you shouldn't (the probability of this is what we call "alpha", and is indicated by the cross-hatched region of the graph); a Type II error occurs if you don't reject it when you should (the probability of this is called "beta", and is indicated by the shaded area). Power refers to the probability of avoiding a Type II error, or, more colloquially, the ability of your statistical test to detect true differences of a particular size. The power of your test generally depends on four things: your sample size, the effect size you want to be able to detect, the Type I error rate (alpha) you specify, and the variability of the sample. Based on these parameters, you can calculate the power level of your experiment. Or, as is most commonly done, you can specify the power you desire (e.g. .80), the alpha level, and the minimum effect size which you would consider "interesting", and use the power equation to determine the proper sample size for your experiment. (See Cohen, 1988, for more details on power analysis.)

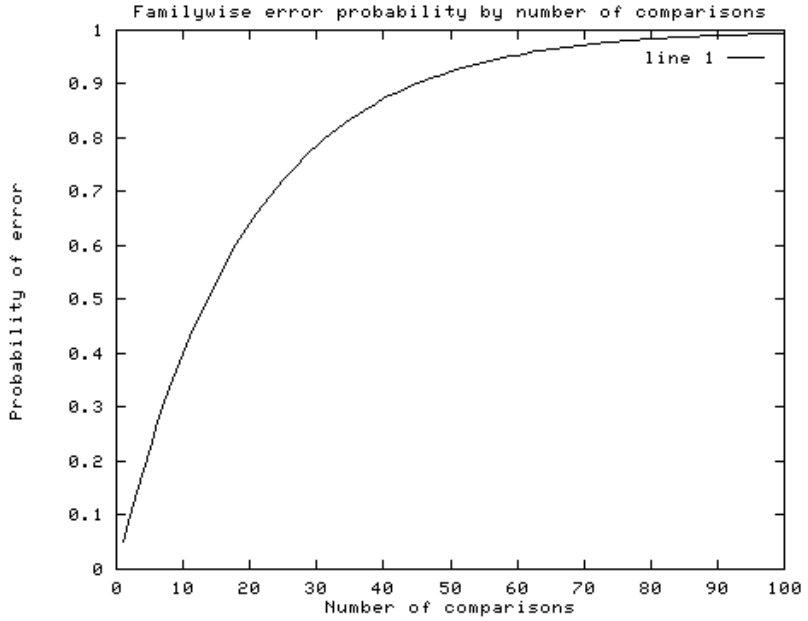
Now, if you have too little power, you run the risk of overlooking the effect you're trying to find. This is especially important if you intend to make inferences based on a finding of no difference. This is what allows advertisers to claim "No brand is better at relieving headaches (or what have you)"--if they use a relatively small sample (say 10 people), of course any differences in pain relief won't be significant. The differences may be there, but the test used to look for them was not sensitive enough to find them.

While the main emphasis in the development of power analysis has been to provide methods for assessing and increasing power (see, e.g. Cohen, 1991), it should also be noted that it is possible to have too much power. If your sample is too large, nearly any difference, no matter how small or meaningless from a practical standpoint, will be "statistically significant". This can be particularly problematic in applied settings, where courses of action are determined by statistical results. (I'll have more to say about this later.).

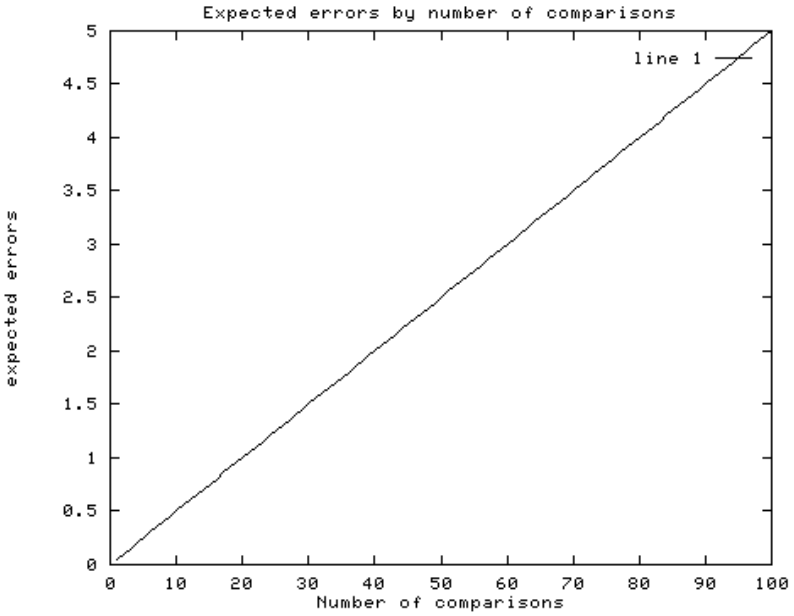
Multiple comparisons. This is a particularly thorny issue, because often what we want to know about is complex in nature, and we really need to check a lot of different combinations of factors to see what's going on. However, doing so in a haphazard manner can be dangerous, if not downright disastrous. Remember that each comparison we make (assuming we're using the standard hypothesis testing model) entails a Type I error risk equal to our predefined alpha.

We might assign the conventional value of .05 to alpha. Each comparison we make has a (1 - .05) = .95 probability of avoiding a Type I error.

Now suppose we have 12 process variables, and we want to see what the relationships are among them. We might be tempted to calculate the 66 possible correlations, and see which ones turn out to be statistically significant. Here's where the fun begins: in the best-case scenario, where the comparisons are independent (not true for this example, but we'll assume it for the sake of argument), the probability of getting all the comparisons right is the product of the probabilities for getting each comparison right. In this case that would be (.95)^66, or about .03. Thus your chance of getting all 66 comparisons right is almost zero. This figure shows the probability of getting one or more errors based on how many comparisons you make, assuming a per-comparison alpha of .05.



In fact, if you were to take 12 completely uncorrelated variables (that is, take a sample from each of 12 uncorrelated variables) and calculate the set of 66 correlations, you should expect to see about 3 spurious correlations in each set. (Note that this is the best-case scenario--if you allow for dependence among the separate tests, the probability of errors is even greater.) This figure shows the expected number of errors based on number of comparisons, assuming a nominal alpha of .05.



So, suppose you calculate your correlations and discover that 10 of them seem to be significant. You'll have a tough time sorting out which ones are real and which are spurious. Several strategies can be used to overcome this problem. The easiest, but probably the least acceptable, is to adjust your alpha criterion (by making it smaller) so that the "familywise" error rate is what you'd like it to be. The problem with this strategy is that it is impractical for large numbers of comparisons: as your alpha for each comparison becomes smaller, your power is reduced to almost nil. The best strategy, but usually an expensive one, is replication--rerun the experiment and see which comparisons show differences in both groups. This is not quite foolproof, but it should give you a pretty good idea which effects are real and which are not. If you can't actually replicate, the next best thing is a technique called cross-validation, which involves setting aside part of your sample as a validation sample. You compute the statistics of interest on the main sample, and then check them against the validation sample to verify that the effects are real. Results that are spurious will usually be revealed by a validation sample.

Measurement error. Most statistical models assume error free measurement, at least of independent (predictor) variables. However, as we all know, measurements are seldom if ever perfect. Particularly when dealing with noisy data such as questionnaire responses or processes which are difficult to measure precisely, we need to pay close attention to the effects of measurement errors. Two characteristics of measurement which are particularly important in psychological measurement are reliability and validity.

Reliability refers to the ability of a measurement instrument to measure the same thing each time it is used. So, for instance, a reliable measure should give you similar results if the units (people, processes, etc.) being measured are similar. Additionally, if the characteristic being measured is stable over time, repeated measurement of the same unit should yield consistent results.

Validity is the extent to which the indicator measures the thing it was designed to measure. Thus, while IQ tests will have high reliability (in that people tend to achieve consistent scores across time), they might have low validity with respect to job performance (depending on the job). Validity is usually measured in relation to some external criterion, e.g. results on a job-

applicant questionnaire might be compared with subsequent employee reviews to provide evidence of validity.

Methods are available for taking measurement error into account in some statistical models. In particular, structural equation modeling allows one to specify relationships between "indicators", or measurement tools, and the underlying latent variables being measured, in the context of a linear path model. For more information on structural equation modeling and its uses, see Bollen (1989).

9.4Problems with interpretation

There are a number of difficulties which can arise in the context of substantive interpretation as well. We go through all these elaborate procedures, chew up time on the mainframe, generate reams of output--eventually, we have to try to make some sense of it all in terms of the question at hand.

Confusion over significance. While this topic has been expounded upon by nearly every introductory statistics textbook, the difference between "significance" in the statistical sense and "significance" in the practical sense continues to elude many statistical dabblers and consumers of statistical results. There is still a strong tendency for people to equate stars in tables with importance of results. "Oh, the p-value was less than .001--that's a really big effect," we hear our clients say. Well, as I pointed out earlier, significance (in the statistical sense) is really as much a function of sample size and experimental design as it is a function of strength of relationship. With low power, you may be overlooking a really useful relationship; with excessive power, you may be finding microscopic effects with no real practical value. A reasonable way to handle this sort of thing is to cast results in terms of effect sizes (see Cohen, 1994)--that way the size of the effect is presented in terms that make quantitative sense. Remember that a p-value merely indicates the probability of a particular set of data being generated by the null model--it has little to say about size of a deviation from that model (especially in the tails of the distribution, where large changes in effect size cause only small changes in p-values).

Precision and Accuracy. These are two concepts which seem to get confused an awful lot, particularly by those who aren't mathematically inclined. It's a subtle but important distinction: precision refers to how finely an estimate is specified (akin to number of decimal places given, e.g. 4.0356 is more precise than 4.0), whereas accuracy refers to how close an estimate is to the true value. Estimates can be precise without being accurate, a fact often glossed over when interpreting computer output containing results specified to the fourth or sixth or eighth decimal place. My advice: don't report any more decimal places than you're fairly confident are reflecting something meaningful. Thus, if your standard error of a mean is 1.2, there's no sense in reporting the third or fourth decimal place of the estimated mean--it will only lull the unwary into a false sense of security.

Causality. I don't think anything has caused as much mischief in research and applied statistics as unclear thinking about causality. Assessing causality is the *raison d'êtr*e of most statistical analysis, yet its subtleties escape many statistical consumers.

The bottom line on causal inference is this: you must have random assignment. That is, the experimenter must be the one assigning values of predictor variables to cases. If the values are not assigned or manipulated, the most you can hope for is to show evidence of a relationship of some kind. Observational studies are very limited in their ability to illuminate

causal relationships. Take, for example, an hypothesized relationship between number of health-care visits and socioeconomic status (SES), i.e. the higher your SES, the more you visit the clinic. There are three possible explanations for this:

one is that people with high SES have the means to pay for frequent clinic visits (SES -> visits); another is that people who visit their doctor frequently are in better health and so are able to be more productive at work, get better jobs, etc. (visits -> SES);

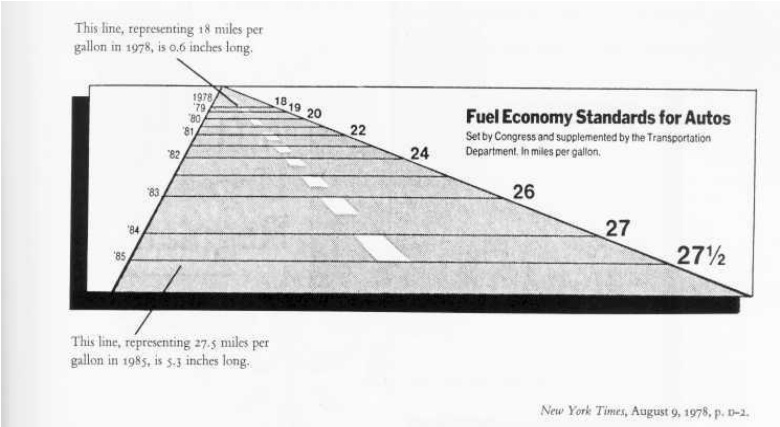
the third is that something else (e.g. size of city) affects both clinic visitation and SES independently (larger cities have more clinics and offer better paying jobs), making them go up and down together (visits <- X -> SES).

I want to point out here that this factor of causal inference (i.e. random assignment) is the key *regardless of the statistical methodology used.* We've all had it drummed into our heads that "correlation is not causation". Unfortunately, some people seem to interpret that as implying that correlation (and regression) can't be used for causal analysis; or worse, that experimentally oriented statistical designs (e.g. ANOVA) are necessary and sufficient conditions for causal inference. Neither of these interpretations is correct; if you assign values to a predictor variable (e.g. by manipulating drug dosages), it is perfectly legitimate to use a correlation coefficient or a regression equation to generate inferences about the effectiveness of the drug. Conversely, if you're measuring relationships between political affiliation and self-esteem, it doesn't matter what sort of elaborate ANOVA design you put together--you still won't have a warrant for making causal statements about what causes what, since you aren't assigning people to political parties.

Now, of course, many of the things we might wish to study are not subject to experimental manipulation (e.g. health problems/risk factors). If we want to understand them in a causal framework, we must be very cautious. It will require a multifaceted approach to the research (you might think of it as "conceptual triangulation"), use of chronologically structured designs (placing variables in the roles of antecedents and consequents), and plenty of replication, to come to any strong conclusions regarding causality.

Graphical Represenations. There are many ways to present quantitative results numerically, and it is easy to go astray (or to lead your audience astray, if you are less than scrupulous) by misapplying graphical techniques. Tufte's book, *The Visual Display of Quantitative Information,* gives a wealth of information and examples on how to construct good graphs, and how to recognize bad ones. I'll present a few of these examples here.

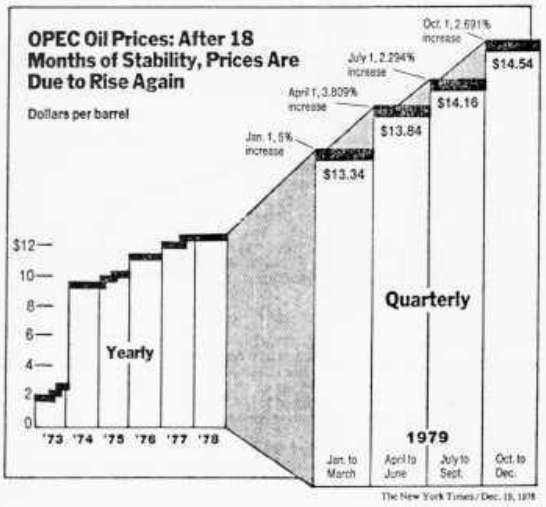
One of the principles Tufte introduces to indicate the relationship between the data and the graphic is a number he calls "the Lie Factor". This is simply the ratio of the difference in the proportion of the graphic elements versus the difference in the quantities they represent. The most informative graphics are those with a Lie Factor of 1. Here is an example of a badly scaled graphic, with a lie factor of 14.8:



(from Tufte, 1983, p. 57)

Another key element in making informative graphs is to avoid confounding design variation with data variation. This means that changes in the scale of the graphic should always correspond to changes in the data being represented. Here is an example which violates this principle:

Design variation corrupts this display:

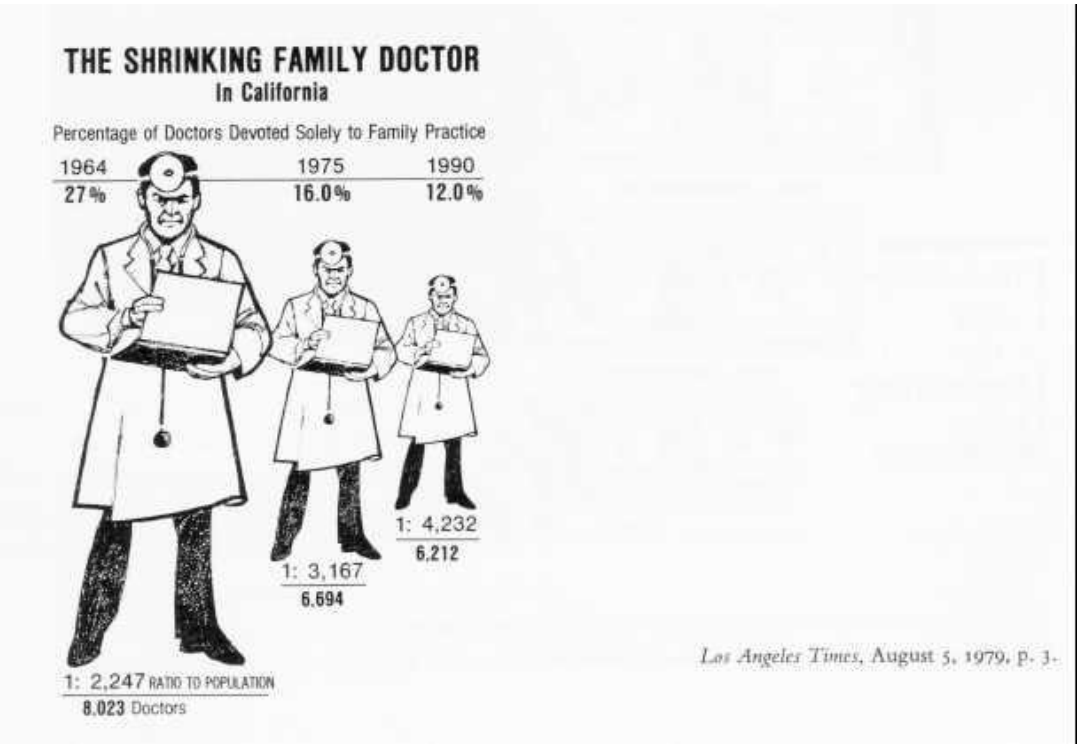


New York Times, December 19, 1978, p. D-7.

(from Tufte, 1983, p. 61)

Notice that from 1973-1978, each bar in the graph represents one year, whereas for 1979 each bar represents only one quarter. Also notice that the vertical scale changes between '73-'78 and the four quarters of '79. It is very difficult to determine from this graph what the real trend is due to the confusion between the design variation and data variation.

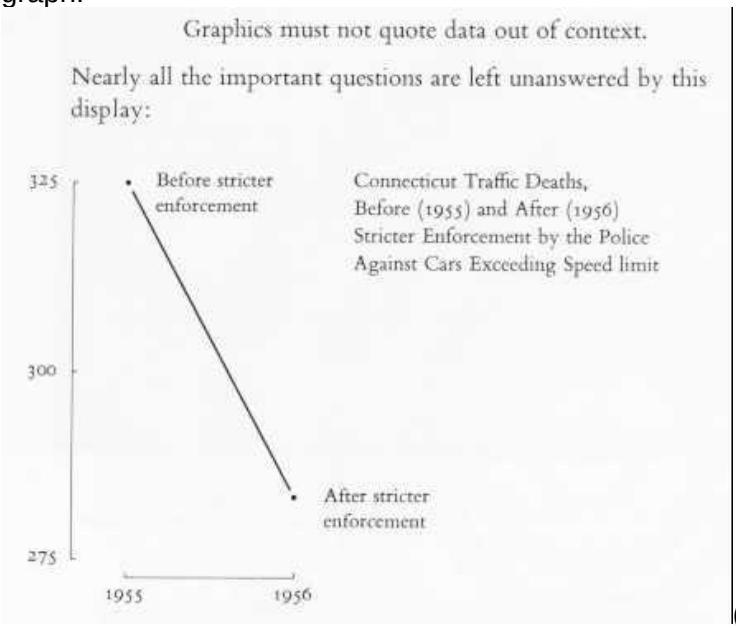
Another trouble spot with graphs is multidimensional variation. This occurs where two-dimensional figures are used to represent one-dimensional values. What often happens is that the size of the graphic is scaled both horizontally and vertically according to the value being graphed. However, this results in the area of the graphic varying with the *square* of the underlying data, causing the eye to read an exaggerated effect in the graph. An example:



(from Tufte, 1983, p. 69)

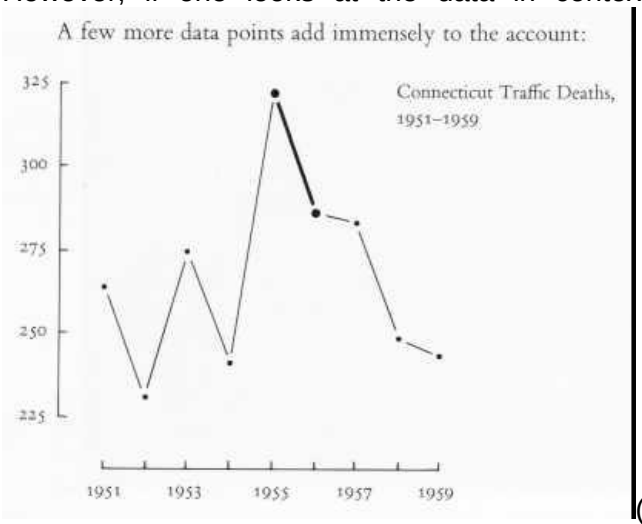
This graph has a lie factor of about 2.8, based on the variation between the area of each doctor graphic and the number it represents.

And, one more point about graphs: be sure to include enough context to make the graph meaningful. For instance, one may be tempted to draw unwarranted conclusions based on this graph:



(from Tufte, 1983, p. 74)

However, if one looks at the data in context, the pattern becomes less cut-and-dried:



(from Tufte, 1983, p. 74)

Clearly, there are other forces at work on the traffic situation besides the stricter enforcement. This information would be completely missed if all you had to look at was the former graph.

9.5 Summary

In this paper I've discussed some of the trickier aspects of applied data analysis. Here are the important points in condensed form:
Be sure your sample is representative of the population in which you're interested.

Be sure you understand the assumptions of your statistical procedures, and be sure they are satisfied. In particular, beware of hierarchically organized (non-independent) data; use techniques designed to deal with them.

Be sure you have the right amount of power--not too little, not too much.
Be sure to use the best measurement tools available. If your measures have error, take that fact into account.

Beware of multiple comparisons. If you must do a lot of tests, try to replicate or use cross-validation to verify your results.

Keep clear in your mind what you're trying to discover--don't be seduced by stars in your tables; look at magnitudes rather than p-values.

Use numerical notation in a rational way--don't confuse precision with accuracy (and don't let the consumers of your work do so, either).

Be sure you understand the conditions for causal inference. If you need to make causal inference, try to use random assignment. If that's not possible, you'll have to devote a lot of effort to uncovering causal relationships with a variety of approaches to the question.
Be sure your graphs are accurate and reflect the data variation clearly.

References

Bollen, K. (1989). *Structural Equations with Latent Variables.* New York: John Wiley & Sons.
Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models.* Newbury Park: Sage Publications.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.
Goldstein, H.I. (1987) *Multilevel Models in Educational and Social Research.* London: Oxford University Press.

Kirk, R. (1982). Experimental Design: Procedures for the Behavioral Sciences. Monterrey, CA: Brooks/Cole.

Levin, J.R. (1985). Some methodological and statistical "bugs" in research on children's learning. In M. Pressley & C.J. Brainerd (Eds.), *Cognitive Learning and Memory in Children.* New York: Springer-Verlag.

Tufte, E.R. (1983). *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics Press.

Related reading

Banks, D. (1993). Is Industrial Statistics out of control? *Statistical Science, 8,* 356-377 (plus commentary).

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997-1003.

Huff, Darrrell. (1954). *How to Lie with Statistics.* New York: W.W. Norton & Company.
King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science, 30* 666-687.

Levin, J.R. & Levin, M.E. (1993). Methodological problems in research on academic retention programs for at-risk minority college students. *Journal of College Student Development, 34,* 118-124.

Nester, M.R. (1996). An Applied Statistician's Creed. *Applied Statistics, 45,* 401-410.
Paulos, J.A. (1988). Innumeracy: mathematical illiteracy and its consequences. New York: Hill & Wang.

Acknowledgements
The author would like to thank Roger L. Brown, Ph.D., Linda J. Baumann, Ph.D., RN, CS, and Joel Levin, Ph.D., for their helpful comments on this workshop, as well as the stimulating input of the many subscribers to the Usenet newsgroups sci.stat.consult and sci.stat.edu (A.K.A. STAT-L and EDSTAT-L, respectively). This work was supported by the American Cancer Society through grants PBR-51a and PBR-51b, Richard R. Love, PI.

10.POSITIVISMO

El **positivismo** es un **sistema filosófico** que se basa en el **método experimental** y que rechaza los conceptos universales y las nociones a priori. Para los positivistas, el único **conocimiento** válido es el conocimiento científico que surge de la afirmación positiva de las teorías tras la aplicación del **método científico**.

El desarrollo del positivismo está vinculado a las consecuencias de la **Revolución Francesa**, que convirtió al ser humano y a la **sociedad** en objeto de estudio científico. Esta novedad requería de una nueva epistemología para legitimar los conocimientos obtenidos.

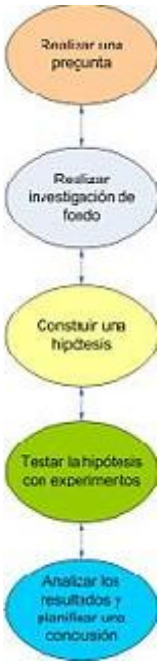
El francés **Augusto Comte** y el británico **John Stuart Mill** suelen ser señalados como los padres de esta epistemología y del positivismo en general. Ambos sostuvieron que cualquier actividad filosófica o científica debe llevarse a cabo mediante el análisis de los hechos reales que fueron verificados por la experiencia.

La epistemología positivista recibió diversas críticas por parte de quienes creían que sus objetos de estudio (como el **hombre** y la **cultura**) no podían ser analizados con el mismo método que se utiliza en las ciencias naturales. La creación de significado y la intencionalidad, por ejemplo, son exclusivas de los seres humanos.

La **hermenéutica** fue una de las corrientes que se opuso al positivismo, buscando la comprensión de los fenómenos y no la explicación. **Bertrand Russell** y **Ludwig Wittgenstein** estuvieron entre los pensadores que intentaron separar la **ciencia** de la metafísica.

Positivismo también es, por último, la actitud práctica, la afición excesiva a los goces materiales y la tendencia a valor los aspectos materiales de la realidad por sobre todas las cosas.

10.1Método científico

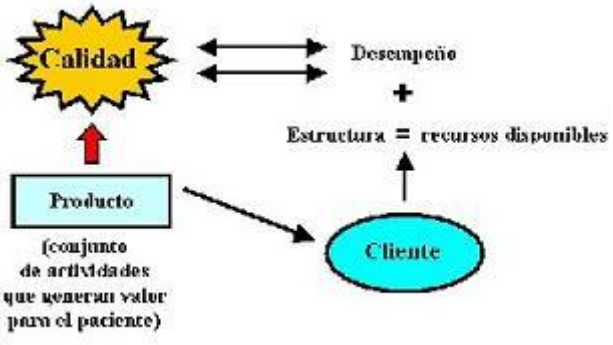


El concepto de **método** proviene del griego *methodos* (**"camino"** o **"vía"**) y hace referencia al **medio utilizado para alcanzar un fin**.

El **método científico**, por lo tanto, se refiere al conjunto de pasos necesarios para obtener **conocimientos válidos** (científicos) mediante **instrumentos confiables**. Este método intenta proteger al investigador de la subjetividad.

El método científico se basa en la **reproducibilidad** (la capacidad de repetir un determinado experimento en cualquier lugar y por cualquier **persona**) y la **falsabilidad** (toda proposición científica tiene que ser susceptible de ser falsada).

Entre los pasos necesarios que conforman el método científico, se encuentran la **observación** (consiste en aplicar los sentidos a un objeto o a un fenómeno, para estudiarlo tal como se presenta en realidad), la **inducción** (acción y efecto de extraer, a partir de determinadas observaciones, el principio particular de cada una de ellas), el **planteamiento de la hipótesis** (mediante la observación), la **demostración o refutación de la hipótesis**, y la presentación de la **tesis** o **teoría científica**.



Entre los distintos tipos de métodos científicos, aparecen el **método empírico-analítico** (se basa en la lógica empírica, distinguiendo los elementos de un fenómeno y revisando cada uno de ellos por separado), el **método experimental** (que comprende el **método hipotético deductivo**, el **método de la observación científica** y el **método de la medición**), el **método hermenéutico** (estudia la coherencia interna de los textos), el **método dialéctico** (considera los fenómenos históricos y sociales en continuo movimiento), el **método fenomenológico** (conocimiento acumulativo) y el **método histórico** (relacionado al conocimiento de las distintas etapas de los objetos en su sucesión cronológica).

10.2Método inductivo o inductivismo

El **método inductivo** o **inductivismo** es un **método científico** que **obtiene conclusiones generales a partir de premisas particulares**. Se trata del método científico más usual, que se caracteriza por cuatro etapas básicas: la observación y el registro de todos los hechos: el análisis y la clasificación de los hechos; la derivación inductiva de una generalización a partir de los hechos; y la contrastación.

Esto supone que, tras una primera etapa de observación, análisis y clasificación de los hechos, se deriva una hipótesis que soluciona el problema planteado. Una forma de llevar a cabo el método inductivo es proponer, apartir de la **observación repetida de objetos o acontecimientos de la misma naturaleza**, una conclusión para todos los objetos o eventos de dicha naturaleza.

El razonamiento inductivo puede ser **completo** (se acerca a un **razonamiento deductivo** ya que la conclusión no aporta más información que la dada por las premisas) o **incompleto** (la conclusión va más allá de los datos que aportan las premisas; a mayor cantidad de datos, mayor probabilidad. Sin embargo, la verdad de las premisas no garantiza la verdad de la conclusión).

Ejemplo de razonamiento inductivo completo:
Pedro y Marta tienen tres perros: Pancho, Pepe y Toto.
Pancho es de color negro.
Pepe es de color negro.
Toto es de color negro.
Por lo tanto, todos los perros de Pedro y Marta son de color negro.

Ejemplo de razonamiento inductivo incompleto.

Pancho es un perro de color negro.
Pepe es un perro de color negro.
Toto es un perro de color negro.
Por lo tanto, todos los perros son de color negro.

Como puede verse, en el segundo ejemplo todas las premisas son verdaderas, pero la conclusión es falsa.

El **inductivismo** o **método lógico inductivo** es un método científico que saca conclusiones generales de algo particular. Este ha sido el método científico más común, pero también han surgido otras escuelas epistemológicas que han desarrollado otros como el falsacionismo y los paradigmas de Kuhn.

El inductivismo se caracteriza por tener 4 etapas básicas:

Observación y registro de todos los hechos
Análisis y clasificación de los hechos
Derivación inductiva de una generalización a partir de los hechos
Contrastación

En una primera etapa se deberían observar y registrar todos los hechos y luego analizarlos y clasificarlos ordenadamente.

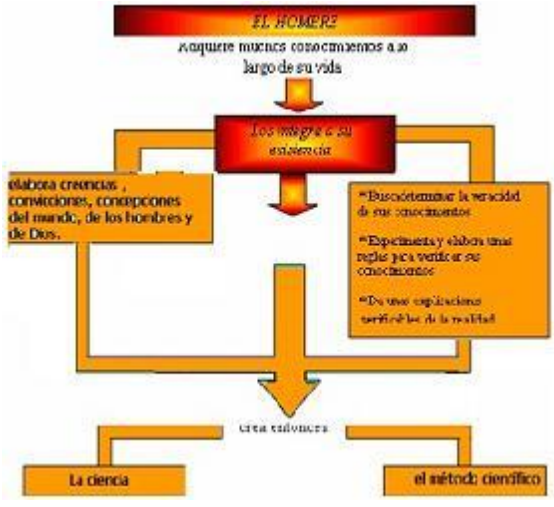
A partir de los datos procesados se deriva una hipótesis que solucione el problema basada en el análisis lógico de los datos procesados. Esta derivación de hipótesis se hace siguiendo un razonamiento inductivo.

En la última etapa se deduce una implicación contrastadora de hipótesis. Esta implicación debería ocurrir en el caso de que la hipótesis sea verdadera, así si se confirma la implicación contrastadora de hipótesis quedará validada la hipótesis principal.
Opuestamente al razonamiento inductivo en el cual se formulan leyes a partir de hechos observados, el razonamiento deductivo infiere esos mismos hechos basándose en la ley general. Según Bacon la inducción es mejor que la deducción porque mientras que de la inducción se pasa de una particularidad a una generalidad, la deducción es de la generalidad.

Se divide en:

Método deductivo directo de conclusión inmediata: Se obtiene el juicio de una sola premisa, es decir que se llega a una conclusión directa sin intermediarios.
Método deductivo indirecto o de conclusión mediata: La premisa mayor contiene la proposición universal, la premisa menor contiene la proposición particular, de su comparación resulta la conclusión. Utiliza silogismos.

10.3Método deductivo



El **método deductivo** es un **método científico** que considera que **la conclusión está implícita en las premisas**. Por lo tanto, supone que las conclusiones sigue necesariamente a las premisas: si el razonamiento deductivo es válido y las premisas son verdaderas, la conclusión sólo puede ser verdadera.

El razonamiento deductivo fue descrito por los filósofos de la **Antigua Grecia**, entre ellos **Aristóteles**. Cabe destacar que la palabra

deducción proviene del verbo **deducir** (del latín *deducĕre*), que significa sacar consecuencias de un principio, proposición o supuesto.

El método deductivo infiere los hechos observados basándose en la ley general (a diferencia del **inductivo**, en el cual se formulan leyes a partir de hechos observados). Hay quienes creen, como el filósofo **Francis Bacon**, que la inducción es mejor que la deducción, ya que se pasa de una particularidad a una generalidad.

El método deductivo puede dividirse en **método deductivo directo de conclusión inmediata** (cuando se obtiene el juicio de una sola premisa, sin intermediarios) y **método deductivo indirecto o de conclusión mediata** (cuando la premisa mayor contiene la proposición universal y la premisa menor contiene la proposición particular, la conclusión resulta de su comparación).

En todos los casos, los investigadores que siguen el método deductivo comienzan con el planteamiento del conjunto axiomático de partida (donde los supuestos deben incorporar sólo las características más importantes de los fenómenos, con coherencia entre los postulados) y continúan con el proceso de deducción lógica (partiendo siempre de los postulados iniciales). Así, pueden enunciar **leyes de carácter general**, a las que se llegan partiendo del conjunto axiomático y a través del proceso de deducción.

El **pensamiento deductivo** parte de categorías generales para hacer afirmaciones sobre casos particulares.

Hablamos de razonamiento deductivo cuando observando una cosas muchas veces se declama lo visto en toda las cosas de la misma especie válido la conclusión debe poder derivarse necesariamente de las premisas aplicando a éstas algunas de las reglas de inferencia según las reglas de transformación de un sistema deductivo o cálculo lógico. Al ser estas reglas la aplicación de una ley lógica o tautología y, por tanto una verdad necesaria y universal, al ser aplicada a las premisas como caso concreto permite considerar la inferencia de la conclusión como un caso de razonamiento deductivo.

Dicho de otro modo, la conjunción o producto de todas las premisas cuando es verdadero, es decir, todas y cada una de las premisas son verdaderas, entonces se implica la verdad de la conclusión.

Por medio de un razonamiento de estas características se concede la máxima solidez a la conclusión, las premisas implican lógicamente la conclusión. Y la conclusión es una consecuencia lógica de las premisas.
Silogismo

El **silogismo** es una forma de razonamiento deductivo que consta de dos proposiciones como premisas y otra como conclusión, siendo la última una inferencia necesariamente deductiva de las otras dos. Fue formulado por primera vez por Aristóteles, en su obra lógica recopilada como *El Organon*, de sus libros conocidos como Primeros Analíticos (en griego, *Proto Analytika*, en latín –idioma en el que se reconoció la obra en Europa Occidental-, *Analytica Priora*).

Aristóteles consideraba la lógica como lógica de relación de términos. Los términos se unen o separan en los juicios. Los juicios aristotélicos son considerados desde el punto de vista de

unión o separación de dos términos, un sujeto y un predicado. Hoy se hablaría de proposiciones.

La diferencia entre juicio y proposición es importante. La proposición afirma un hecho como un todo, que es o no es, como contenido lógico del conocimiento. El juicio, en cambio, atribuye un predicado a un sujeto lógico del conocimiento. Esto tiene su importancia en el concepto mismo del contenido de uno y otra, especialmente en los casos de negación, como se ve en la problemática de la lógica silogística.

Mantenemos aquí la denominación de juicio por ser lo más acorde con lo tradicional, teniendo en cuenta que este tipo de lógica, como tal, está en claro desuso, sustituida por la lógica simbólica en la que esta lógica es interpretada como lógica de clases. Ver cálculo lógico.
La relación entre los términos de un juicio, al ser comparado con un tercero que hace de "término medio", hace posible la aparición de las posibles conclusiones. Así pues, el silogismo consta de dos juicios, premisa mayor y premisa menor, en los que se comparan tres términos, de cuya comparación se obtiene un nuevo juicio como conclusión.

La lógica trata de establecer las leyes que garantizan que, de la verdad de los juicios comparados (premisas), se pueda obtener con garantía de verdad un nuevo juicio verdadero (conclusión).

10.4Razonamiento abductivo

La **abducción** (del latín *abductio* y esta palabra de *ab* –desde lejos– *ducere* llevar) es un tipo de razonamiento inicialmente puesto en evidencia por Aristóteles en su Analytica priora (II, 25); tal razonamiento opera con una especie de silogismo en donde la premisa mayor es considerada cierta mientras que la premisa menor es solo probable, por este motivo la Conclusión a la que se puede llegar tiene el mismo grado de probabilidad que la premisa menor.

Según el filósofo Charles Sanders Peirce, la abducción es algo más que una suerte de silogismo; es una de las tres formas de razonamiento junto a la deducción y la inducción.
Lógica

En la abducción a fin de entender un fenómeno se introduce una Regla que opera en forma de hipótesis para considerar dentro de tal regla al posible resultado como un Caso particular. En otros términos: en el caso de una *deducción* se obtiene una Conclusión « **q** » de una Premisa « **p** », mientras que el razonar abductivo consiste en explicar « **q** » mediante « **p** » considerando a p como hipótesis explicativa.

De este modo la abducción es la operación lógica por la que surgen hipótesis *novedosas*.
En muchos casos las abducciones no son sino las conjeturas espontáneas de la razón. Para que esas hipótesis surjan se requiere el concurso de la imaginación y del instinto. La abducción es como un destello de comprensión, un saltar por encima de lo sabido; para la abducción es preciso dejar libre a la mente. Peirce habla en ese sentido del *musement*, un momento más instintivo que racional en el que hay un flujo de ideas, hasta que de pronto se ilumina la sugerencia, según el mismo Peirce la "abducción es el primer paso del razonamiento científico" (*Collected papers* 7.218) ya que desde el inicio se efectúa una restricción de hipótesis aplicables a un fenómeno.

Según ese filósofo estadounidense el pensar humano tiene tres posibilidades de crear inferencias o tres diversos modos de razonar: el deductivo, el inductivo y el abductivo.

Ejemplos

Un ejemplo de deducción:
Regla: "Todas las bolillas de la bolsa x son blancas".
Caso: "Estas bolillas provienen de la bolsa x".
Deducción: "Estas bolillas son blancas".
Un ejemplo de inducción:
Caso: "Estas bolillas proceden de la bolsa x"
Caso: "Estas bolillas son blancas".Inducción: "En la bolsa x todas las bolillas son blancas"
Un ejemplo de abducción:
Regla: "Todos las bolillas de la bolsa x son blancas".
Caso: "Estas bolillas son blancas"
Abducción: "Estas bolillas proceden de la bolsa x".

10.5Crítica

En la *deducción* la Conclusión se obtiene de la Premisa: dada la Regla y el Caso, el resultado hace explícito algo ya implícito en las premisas (se dice aquí que "se va de lo universal a lo singular"). La *inducción* en cambio permite crear una Regla (hipotética) a partir de un Caso y otro Caso (se va de los singular a lo "universal"). A diferencia de la deducción y como la misma abducción, la inducción **no** es lógicamente válida sin confirmaciones externas (en los ejemplos dados, bastaría *una* excepción a la regla para que la regla quedase falsada, por ejemplo, bastaría una bolilla negra...por más que la excepción *puede* reforzar en cierto modo a la regla precisamente por su carácter de excepcionalidad), la inducción y la abducción no son válidas sin una ratificación empírica y pese a todas las posibles ratificaciones empíricas siempre parece existir el riesgo de una excepción.

Siguiendo con los ejemplos dados y observando que, tenemos bolillas blancas y teniendo a disposición una Regla como para dar una explicación (sabemos que todas las bolillas de la bolsa x son blancas) entonces podemos hipotetizar válidamente que *quizás*, *probablemente*, estas bolillas blancas procedan de la bolsa x. De este modo (pese a la incertidumbre) hemos incrementado nuestro conocimiento en cuanto sabemos ya algo más: al principio sabíamos que (por ejemplo) "las bolillas eran blancas", ahora sabemos que pueden corresponder al conjunto de la bolsa x.

Por estar fundamentada en el juego de hipótesis probables, es que Peirce ha considerado a la abducción "como la única forma de razonar que es realmente susceptible de incrementar nuestro saber, o, mejor dicho, al hipotetizar, crear **nuevas ideas** y prever. En lo real las tres formas de inferencia lógica (abducción, deducción, inducción) permiten incrementar la consciencia, aunque en orden y medida diferentes; al respecto opina Peirce que sólo la abducción está totalmente dedicada al enriquecimiento cognitivo... aunque al precio de un cierto riesgo de error, si bien se observa la abducción ésta aparece como el modo inferencial más inductivo.

La abducción, como la inducción, no contiene en sí una validez lógica y debe ser confirmada, la confirmación sin embargo jamás podrá ser absoluta sino sólo probable, existirá una abducción correcta si la Regla elegida para explicar la Conclusión se confirma tantas veces de

modo que la probabilidad prácticamente equivale a una razonable certeza y si no existen otras Reglas que expliquen igualmente bien o mejor los fenómenos en cuestión.

En cierto modo la abducción, precisamente por su imprecisión original implica un modo de pensar *no lineal* (existe aquí alguna analogía con el pensamiento lateral).

Para el semiótico Umberto Eco el razonar abductivo es el «razonar del detective» en cuanto en ella se pueden relacionar diversos indicios dentro de una hipótesis explicativa válida.

Definir: Explicar con claridad y precisión la significación de una palabra o la naturaleza de una cosa. Lo que se definen son conceptos, ideas formadas con las características esenciales.

Explicar: Desarrollar un concepto o tema dado a partir de un marco teórico, exponer cualquier materia, texto o problema con palabras claras para hacerlo entendible dando a conocer su causa o motivo.

Caracterizar: Determinar las peculiaridades de cada concepto o tema.
Comparar: Establecer semejanzas, diferencias y relaciones entre dos o más series de datos, hechos o conceptos, sacando conclusiones pertinentes.

Relacionar o vincular: Realizar correspondencia entre conceptos o temas, asociar. Al vincular conceptos se exponen diferentes realidades o elementos buscando los puntos que tienen en común

Clasificar: Agrupar en clases y categorías, los elementos para organizarlos, conforme a sus principios generales.

Diferenciar: Encontrar distinciones entre conceptos o temas.

Analizar: Descomposición de un todo complejo en elementos simples, es decir, la explicación de las diferentes partes que componen un tema o concepto hasta llegar a sus principios elementales.

Fundamentar: Elaborar una argumentación o justificación, desde un marco teórico determinado para afirmar o refutar hipótesis.

11. ESTADÍSTICA: SOFTWARE GRATUITO

Mstat *Windows* *Mac OSX* *Linux* Análisis estadístico para Windows, Mac y Linux. Estadística descriptiva e inferencial paramétrica y no paramétrica.

Instat Análisis estadístico para Windows. Estadística descriptiva e inferencial paramétrica y no paramétrica. Modulo para aplicaciones climáticas.

XLStatistics Current version: 08.05.12. Análisis estadístico con Excel. Estadística descriptiva e inferencial paramétrica y no paramétrica.

PAST Análisis estadístico univariado, multivariado. Índices de diversidad. Estadística descriptiva e inferencial paramétrica y no paramétrica.

MacAnova Análisis estadístico para Macs y Windows, análisis de poder

OpenEpi Plataforma independiente. Enfatiza análisis epidemiológico.

The R Project for Statistical Computing. Gran variedad de análisis, muy poderoso pero requiere de usuarios experimentados. Opera en base a comandos.

Remuestreo Sofware para análisis estimación y pruebas de hipótesis utilizando remuestreo.

PSPP is a program for statistical analysis of sampled data. It is a Free replacement for the proprietary program SPSS, and appears very similar to it with a few exceptions. http://www.jdmp.org/misc/related-software/

Minería de datos

Java Data Mining Package (e.g. clustering, regression, classification, graphical models, optimization) http://www.jdmp.org/

Gretl: Gnu Regression, Econometrics and Time-series Library
http://gretl.sourceforge.net/win32/

RapidMiner is the world-wide leading open-source data mining solution due to the combination of its leading-edge technologies and its functional range. Applications of RapidMiner cover a wide range of real-world data mining tasks. http://rapid-i.com/content/blogcategory/38/69/

Lecturas de apoyo:

Parametric and Resampling Statistics: Two Different Philosophies of Hypothesis Testing -- Or is it Three? http://www.uvm.edu/~dhowell/StatPages/Resampling/philosophy.html
Randomization Tests

http://www.uvm.edu/~dhowell/StatPages/Resampling/RandomizationTests.html
Bootstrapping Approaches to Inference

http://www.uvm.edu/~dhowell/StatPages/Resampling/Bootstrapping.html
Lunneborg, C. E. (2000) Random assignment of available cases: Let the inference fit the design. http://faculty.washington.edu/lunnebor/Australia/randomiz.p df

http://faculty.washington.edu/lunnebor/Australia/randomiz.pdf

Análisis de Poder

G*Power 3 Análisis estadístico de poder para las siguientes pruebas estadísticas: *F* , *t*, χ^2, familia de z y algunas pruebas exactas.

12.BIBLIOGRAFÍA GENERAL

Hossein Arsham. Razonamiento Estadístico para Decisiones Gerenciales. Disponible en: http://home.ubalt.edu/ntsbarsh/Business-stat/opre504S.htm

Herramientas para el Análisis de Decisión: Análisis de Decisiones Riesgosas. Hossein
Disponible en: Arsham. http://home.ubalt.edu/ntsbarsh/Business-stat/opre/SpanishP.htm
Statistical analysis of survey data ch19fin3.pdf.
Disponible en: http://unstats.un.org/unsd/hhsurveys/FinalPublication/ch19fin3.pdf

Manual de Estadística. David Ruiz Muñoz. Universidad Pablo de Olavide. ISBN: 84-688-6153-7. 91 págs. PDF 709 Kb. Disponible en: http://www.eumed.net/cursecon/libreria/drm/ped-drm-est.htm

National Institute of Standards and Technology . NIST/SEMATECH e-Handbook of Statistical Methods. 2003. Disponible en: http://www.itl.nist.gov/div898/handbook/index.htm
Hossein Arsham.Ciencia de la Administración Aplicada para Gerentes y Lideres Gerenciales: Toma de decisiones estratégicas acertadas. 2006. Disponible en: http://home.ubalt.edu/ntsbarsh/Business-stat/opre/Spanish.htm

Víctor Manuel Quesada Ibarguen y Juan Carlos Vergara Schmalbach. Estadistíca básica con aplicaciones en Ms Excel. Disponible en: http://www.eumed.net/libros/2007a/239/indice.htm
Richard Lowry. Concepts and Applications of Inferential Statistics. Vassar College
Disponible en: http://faculty.vassar.edu/lowry/webtext.html.

Aplicaciones

Kurtz, Janis C., Jacksonb, Laura E. and Fisher William S. Strategies for evaluating indicators based on guidelines from the Environmental Protection Agency's Office of Research and Development. Ecological Indicators. Volume 1, Issue 1, August 2001, Pages 49-60.