Estadística

Nancy Lacourly

11 de noviembre de 2002

‘2

A mes parents pour leur affection

A Juan por sus valiosos consejos

A Poupée, Rodrigo y Fran por su ayuda y cariño

A ma filleule Carole

Prefacio

Muchas personas prefieren situaciones con riesgo nulo a enfrentar eventos aleatorios o arriesgados.Tomar decisión con incertidumbre no es parte de la cultura de cualquier persona. Incluso, aunque los juegos de azar son muy populares, su teoría es poco conocida.

En la actualidad la estadística es una herramienta necesaria para muchas otras disciplinas donde fenómenos aleatorios son estudiados para obtener y entender informaciones en vista de tomar decisiones relativas a poblaciones de gran tamaño.

Enseñar la estadística se volvió una necesidad pero su dificultad constituye un desafío.

El curso de estadística es parte del plan común de ingeniería y para algunas carreras es el único curso de estadística que tendrá el alumno. Se espera, introducir al alumno al razonamiento y al modelamiento estadístico

El libro comprende en particular una introducción al muestreo, a la metodología básica de la Inferencia Estadística y a los métodos multidimensionales con el modelo lineal.

Se busca preparar al futuro profesional en la aplicación de modelos estadísticos para tratar fenómenos aleatorios en física, mecánica o economía entre otros, así como trabajar con grandes volúmenes de datos que en la actualidad pueden ser estudiados fácilmente.

Existe una versión interactiva de este libro, que hemos llamado libro orgánico (disponible en la pagina $http://www.dim.uchile.cl/\sim estadistica$), en la cual hay actividades que esperamos ayuden a profundizar los temas del curso.

La puesta a punto de estas actividades interactivas fue realizada por Laurence Jacquet.

Un muy especial agradecimiento a Lorena Cerda que con mucha paciencia me permitió evitar estropear el magnífico idioma de Miguel de Cervantes.

Finalmente este libro no había sido posible sin el financiamiento del proyecto IDEA+ Fondef D99I1049 y del Departamento de Ingeniería Matemática de la Universidad de Chile.

Nancy Lacourly

Octubre 2002

Índice General

Capítulo 1

LA ESTADÍSTICA, ¿QUÉ ES?

La **estadística** es una rama del método científico que trata datos empíricos, es decir datos obtenidos contando o midiendo propiedades sobre poblaciones de fenómenos naturales, cuyo resultado es "incierto". Ofrece métodos utilizados en la recolección, la agregación y el análisis de los datos.

En teoría de las probabilidades, los estudiantes, estudiaron el experimento relativo a tirar un dado y hicieron el supuesto que el dado no estaba cargado (los seis sucesos elementales son equiprobables), lo que permite deducir que la probabilidad de sacar "un número par" es igual a 1/3. A partir de un modelo probabilístico adecuado, se deduce nuevos modelos o propiedades. En estadística tratamos responder, por ejemplo, a la pregunta *¿el dado está cargado?*, comprobando si el modelo probabilístico de equiprobabilidad subyacente esta en acuerdo con datos experimentales obtenidos tirando el dado un cierto número de veces. Se propone entonces un modelo probabilístico que ajuste bien los datos del experimento. En resumen, en estadística se tiene un problema a resolver o una *hipótesis de trabajo*, por ejemplo el dado es equilibrado. Se hace un *experimento*, aquí es lanzar el dado, que proporciona datos de los cuales se busca concluir sobre la *hipótesis de trabajo*.

No hay que confundir el uso de la palabra **estadísticas** (plural), que designa un conjunto de datos observados y la palabra **estadística** (singular), que designa la rama del método científico que trata estos datos observados.

Esta introducción se inicia con una breve presentación histórica de la estadística, para seguir con algunos ejemplos de problemas estadísticos. Siguen las etapas del razonamiento que permite resolver tales problemas. Terminamos con introducción a la teoría de muestreo, que es la base de la solución de todo problema estadístisco.

> *Hay tres tipos de mentira: las piadosas, las crueles y las estadísticas.*
> *Atribuido a Mark Twain por el primer ministro inglés Benjamin Disraeli (1804-1881).*

1.1 HISTORIA DEL AZAR Y DE LA ESTADÍSTICA

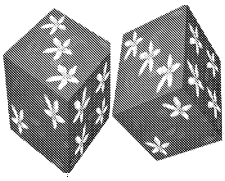
El desarrollo de la computación trastornó los progresos de la estadística y su enseñanza. Vamos a ver aquí cómo y por quién se desarrollo la estadística, desde la prehistoria hasta la

actualidad. Es difícil separar la evolución de la estadística sin considerar la historia de las probabilidades. El progreso de ambas disciplinas puede verse como la historia de una única ciencia: la ciencia del azar.

La prehistoria

La estadística descriptiva tiene su origen mil o dos miles años antes de Cristo, en Egipto, China y Mesopotamia, donde se hacían censos[1] para la administración de los imperios. Los egipcios tuvieron el barómetro económico más antiguo: un instrumento llamado "nilometro", que medía el caudal del Nilo y servia para definir un índice de fertilidad, a partir del cual se fijaba el monto de los impuestos. Con la variabilidad del clima ya conocían el concepto de incertidumbre.

Paralelamente, el concepto de azar es tan antiguo como los juegos (los dados y los juegos con huesos que en Chile llamamos "payayas" son antiquísimos) y motivó desde antaño las reflexiones de los filósofos. En las ideas de Aristóteles (384-322 AC) se encuentran tres tipos de nociones de probabilidad, que definen más bien actitudes frente al azar y la fortuna, que siguen vigentes hoy en día: (1) el azar no existe y refleja nuestra ignorancia; (2) el azar proviene de causas múltiples y (3) el azar es divino y sobrenatural. Sin embargo, pasó mucho tiempo antes de que alguien intentara cuantificar el azar y sus efectos.



La edad Media

Durante la edad media hubo una gran actividad científica y artística en Oriente y el nombre de *azar* parece haber venido desde Siria a Europa. La flor de zahar, que aparecía en los dados de la época podría ser el origen de la palabra. Las compañías aseguradoras iniciaron investigaciones matemáticas desde tiempos muy antiguos, y en siglo XVII aparecieron los primeros famosos problemas de juegos de azar. En la sociedad francesa, el juego era uno de los entretenimientos más frecuentes. Los juegos cada vez más complicados y las apuestas muy elevadas hicieron sentir la necesidad de calcular las probabilidades de los juegos de manera racional. El caballero de Méré, un jugador apasionado, escribiendo sobre ciertos juegos de azar a Blaise Pascal (1623-1662), un austero cristiano jansenista que vivía en un distinto mundo al de nuestro caballero, y dejaría más tarde la matemática por la teología..., dio origen a una correspondencia entre algunos matemáticos de la época. Las preguntas de De Méré permitieron, en particular, iniciar una discusión entre Blaise Pascal y Pierre Fermat

[1] La palabra censo viene de la palabra latina censere que significa fijar impuestos.

(1601-1665) y así el desarrollo de la teoría de las probabilidades. En el siglo anterior, los italianos Tartaglia (1499-1557), Cardano (1501-1576), e incluso el gran Galileo (1564-1642) abordaron algunos problemas numéricos de combinaciones de dados.

En cada juego de azar, dados, cartas o ruleta, por ejemplos, cada una de las jugadas debe dar un resultado tomado de un conjunto finito de posibilidades (números de 1 a 6 para el dado, 52 posibilidades para las cartas o 38 para la ruleta). Si el juego de azar es "correcto" (sin trampas) no se puede predecir de antemano el resultado que se obtendrá en una jugada. Es lo que define el azar del juego. Se observa una cierta simetría en los posibles resultados: son todos igualmente posibles, es decir que el riesgo para un jugador es el mismo cualquiera sea la opción que juega. De aquí surgió la primera definición de una medida de probabilidad para un determinado suceso:

$$p = \frac{a}{b}$$

donde a es el número de casos *favorables* (el número de casos que producen el suceso) y b el número de casos posibles. Por ejemplo, la probabilidad de sacar un "6" en el lanzamiento de un dado es $p = \frac{1}{6}$, de sacar un corazón de un paquete de 52 cartas es $p = \frac{1}{4}$ o un número par en la ruleta (considerando que "0" y "00" son ni pares y ni impares) es $p = \frac{18}{38}$. El caballero De Méré, que jugaba con frecuencia, había acumulado muchas observaciones en diversos juegos y constató una cierta regularidad en los resultados. Esta regularidad, a pesar de tener como base un hecho empírico, permitió relacionar la frecuencia relativa de la ocurrencia de un suceso y su probabilidad. Si f es la frecuencia absoluta de un suceso (el número de veces que ocurrió) en n jugadas, como el número de casos favorables debería ser aproximadamente igual a na, $f \approx \frac{na}{b}$ y entonces la probabilidad de que ocurra el suceso será:

$$p = \frac{a}{b} \approx \frac{f}{n}$$

En un juego, De Méré encontraba una contradicción en su interpretación de la probabilidad a partir de la frecuencia relativa que obtuvo empíricamente. Pascal y Fermat pudieron mostrarle que sus cálculos eran erróneos y que la interpretación propuesta era correcta. De Méré siguió planteando problemas que no pudieron resolver los matemáticos de su época. Sin embargo, Jacques de Bernoulli (1654-1705), el primero de una famosa familia de matemáticos suizos, dio una demostración de la ley de los Grandes Números y Abraham de Moivre enunció el teorema de la regla de multiplicación de la teoría de la probabilidad.

Según Richard Epstein, la ruleta es el juego de casino más antiguo que está todavía en operación. No se sabe a quien atribuirlo: puede ser Pascal, el matemático italiano Don Pasquale u otros. La primera ruleta fue introducida en París en 1765.

La demografía

Las reglas de cálculo desarrolladas hasta entonces para los juegos de azar vieron sus aplicaciones en otras disciplinas. Los censos demográficos, que se hacían desde la antigüedad, requieren recolectar

El problema de los puntos: supongamos que dos jugadores, Abel y Bertrán, interrumpen un juego secuencial en el cual a Abel le falta A y a Bertrán le falta B para ganar. ¿Como tienen que repartirse las apuestas? Es uno de los famosos problemas propuestos por De Méré y que fue resuelto por Fermat y Pascal (1984)

Después de una larga correspondencia, Fermat y Pascal llegaron a la misma solución del problema, por caminos distintos, Fermat usando la combinatoria y Pascal el razonamiento por inducción, lo que tranquilizó a ambos respecto a la justeza de sus razonamientos. De paso, construyeron entre los dos los fundamentos del cálculo de probabilidades a partir de los juegos de azar.

muchos datos. En Inglaterra, a pesar que John Grant tenia la noción de las tablas de mortalidad, es Edmund Halley (1656-1742) que construye por primera vez una tabla de mortalidad utilizando observaciones.

La demografía y los seguros de vida se aprovecharon de este desarrollo de la teoría de las probabilidades. Consideremos, por ejemplo, el sexo de una sucesión de niños recién nacidos. Se puede ver como la repetición del lanzamiento de una moneda, con niño y niña en vez de cara y sello. De la misma manera, podemos considerar un conjunto de hombres mayores de 50 años. Al final del año, una cierta proporción sigue viva. Durante el siglo XVIII, Pierre Simon y Marqués de Laplace (1749-1827), paso, por primera vez, de la observación estadística a la creación de un concepto probabilístico, reconociendo estos problemas como similares a los de un juego, encontrando las correspondientes frecuencias relativas, lo que permitió determinar la probabilidad que nazca una niña, o que un hombre mayor que 50 años muera en el año.

Si bien la extensión de los juegos de azar a la demografía o a la matemática actuarial fue extremadamente importante, su planteamiento tiene grandes limitaciones debido a que considera todos los resultados posibles simétricos. ¿Qué pasa cuando una situación real no puede expresarse como un juego de azar? Por ejemplo, Daniel Bernoulli, careciendo de datos sobre la mortalidad producida por la viruela a distintas edades, supuso que el riesgo de morir de la enfermedad era el mismo en todas las edades. Lo que evidentemente es muy discutible.

Christiaan Huygens (1629–1695), matemático holandés, astronomo y físico, descubrio la teoría ondulatoria de la luz, y contribuyo a la ciencia en general y en particular a la dinamica.

La noción de esperanza matemática se encuentra en sus trabajos. Escribía: si espero A ó B, y que puedo obtener uno ó el otro, puedo decir que mi esperanza vale (A+B)/2.

La teoría de los errores y la distribución normal

Durante los siglos XVIII y XIX la estadística se expandió sin interrupción mientras la teoría de las probabilidades no mostró progreso. Una de las aplicaciones importante fue desarrollada al mismo tiempo por Gauss (1777-1855), Legendre (1752-1833) y Laplace: el análisis numérico de los errores de mediciones en física y astronomía. ¿Cómo determinar el mejor valor leído por un instrumento que entrega diferentes mediciones del mismo fenómeno? Si tenemos n mediciones de un mismo fenómeno $x_1, x_2, ..., x_n$, deberíamos tener $x_1 = x_2 = ... = x_n$ si no tuvieramos errores. En su anexo sobre el método de los mínimos cuadrados, "Nuevos métodos para la determinación de las órbitas

de los cometas", Legendre propone determinar el valor único z de la medición de manera que una función de los errores sea mínima:

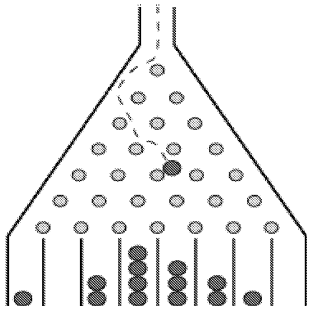
$$\min_z \sum_{i=1}^{n} (x_i - z)^2$$

La solución es el promedio de las mediciones.

Esta función cuadrática encuentra su justificación en la distribución normal con Gauss y Laplace, aunque la distribución de los errores fue estudiada mucho antes por Thomas Simpson (1710-1761), que hizo los supuestos que esta distribución tenía que ser simétrica y que la probabilidad de errores pequeños debería ser más grande que la de los errores grandes. Adolfe Quetelet (1796-1874), un astrónomo belga, hace los primeros intento de aplicar la estadística a las Ciencias Sociales. Una de sus contribuciones fue el concepto de *persona promedio*, persona cuya acción e ideas corresponde al resultado promedio obtenido sobre la sociedad entera.

En 1840, Sir Francis Galton (1822-1911), primo de Charles Darwin, partió de una distribución discreta y la fue refinando hasta llegar en 1857 a una distribución continua muy parecida a la distribución normal. Galton inventó incluso una máquina llamada quincunx o máquina de Galton, que permite ilustrar la distribución normal.

En 1840, Sir Francis Galton (1822-1911), primo de Charles Darwin, partió de una distribución discreta y la fue refinando hasta llegar en 1857 a una distribución continua muy parecida a la distribución normal. Galton inventó incluso una máquina llamada quincunx o máquina de Galton, que permite ilustrar la distribución normal.



La distribución normal es la ley en la cual todo el mundo cree:
Los experimentadores creen que es un teorema de la Matemática,
y los matemáticos que es un hecho experimental.
El astrónomo Lippman.

Nacimiento de la estadística Moderna

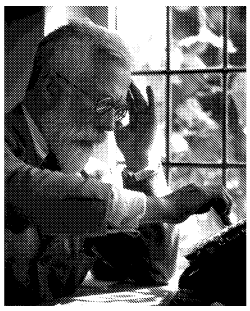
Es con la introducción de nuevas aplicaciones que la teoría de las probabilidades del siglo XVIII funda la estadística matemática. El término de *estadística* se debe posiblemente a G. Achenwall (1719-1772), profesor de la Universidad de Göttingen, tomando del latín la palabra *status*.[2] Achenwall creía, y con razón, que los datos de la nueva ciencia (la estadística) serían el aliado más eficaz de los gobernantes.

Aparte de la demografía y la matemática actuarial, otras disciplinas introdujeron la teoría de las probabilidades. Fue el inicio de la mecánica estadística, debido a Maxwell (1831-1879) y Bolzmann, quienes dieron también una justificación de la distribución normal en la teoría cinética de los gases.

[2]El término latino status significa estado o situación.

La estadística se empezó a usar de una manera u otra en todas las disciplinas, a pesar de un estancamiento de la teoría de las probabilidades. En particular, muchos vieron la dificultad de aplicar el concepto de simetría, o de casos igualmente posibles, en todas las aplicaciones. Hubo que esperar a que Andrey Nickolaevich Kolmogorov (1903-1987) separara la determinación de los valores de las probabilidades de sus reglas de cálculo.

Los primeros resultados importantes de la estadística Matemática se deben al inglés Karl Pearson (1857-1936) y a otros investigadores de la escuela biométrica inglesa tal como Sir Ronald Fisher (1890-1962), que tuvo mucha influencia en el campo de la genética y la agricultura.



Sir Ronald Fisher es considerado como uno de los fundadores de la estadística moderna por todas sus contribuciones.

Estudia en Rothamsted el diseño de experimentos introduciendo el concepto de randomización y del análisis de la varianza. En 1921 crea el concepto de verosimiltud, propone el método de máxima verosimiltud y estudia los tests de hipótesis.

La segunda mitad del siglo XX: la revolución computacional

Los científicos, especialmente los ingleses, desarrollaron métodos matemáticos para la estadística, pero en la práctica manipularon cifras durante medio siglo sin disponer de verdaderas herramientas de cálculo. La llegada de los computadores revolucionó el desarrollo de la estadística. El francs J. P. Benzécri y el norte americano J. W. Tuckey fueron los pioneros en repensar la estadística en función de los computadores. Mejoraron, adaptaron y crearon nuevos instrumentos para estudiar grandes volúmenes de datos: nuevas técnicas y herramientas gráficas.

> *El modelo tiene que adaptarse a los datos y no al revés.*
> *Jean-Paul Benzécri, 1965*

Cálculo de probabilidades y estadística

Algunas palabras para concluir. Si bien la historia de la estadística no se puede separar de la historia del cálculo de las probabilidades, la estadística no puede considerarse como una simple aplicación del cálculo de las probabilidades. Podemos comparar esta situación a la de la geometría y la mecánica. La mecánica usa conceptos de la geometría, y sin embargo es una ciencia a parte.

El cálculo de las probabilidades es una teoría matemática y la estadística es una ciencia aplicada donde hay que dar un contenido concreto a la noción de probabilidad. Como ilustración citemos el experimento de Weldon (1894), que lanzó 315.672 veces un dado (bajo la supervisión de un juez) y anotó que 106.602 veces salió un 5 o un 6. La frecuencia teórica debería ser 0.3333... si el dado hubiera sido perfectamente equilibrado. La frecuencia observada aquí fue 0.3377. ¿Deberíamos concluir que el dado estaba cargado? Es una pregunta concreta que es razonable considerar. El cálculo de las probabilidades no responde a esta pregunta y es la estadística la que permite hacerlo.

El geómetra no se interesa por saber si existen en la práctica objetos que puedan considerarse como líneas rectas. Hay que tener cuidado cuando se razona por analogía con otras ramas de las matemáticas aplicadas, porque a este nivel no nos preocupamos solamente de las relaciones entre cálculo y razonamiento. Admitamos el derecho del matemático de desinteresarse del problema, como matemático, pero tenemos que asumir la responsabilidad de resolver la dificultad, como psicólogo, lógico o estadístico, a menos que estemos dispuestos a poner la probabilidad en el campo de la matemática pura y sus aplicaciones en el frontis de nuestras academias.
Kendall, 1949.

1.2 ¿DONDE SE USA LA ESTADÍSTICA?

Actualmente el gobierno de cada país recolecta sistemáticamente datos relativos a su población, su economía, sus recursos naturales y su condición política y social para tomar decisiones. En las actividades industriales o comerciales las estadísticas son parte de la organización así como en los sectores agrícolas y forestales, donde se requieren predicciones de la producción. En la investigación científica (medicina, física, biología, ciencias sociales, etc.) el rol de la estadística es primordial.

Estadísticas y el Estado

Un estado necesita conocer su población: En Chile los censos permiten obtener estadísticas demográficas y de vivienda y los métodos estadísticos hacer predicciones dentro el periodo de 10 años que transcurre entre dos censos. Para poder elaborar una planificación de la salud, el gobierno tiene que tener informaciones sobre las necesidades de la población (datos demográficos, enfermedades según las estaciones, etc.) y un inventario de las infraestructuras de salud. En función de estas informaciones, se crean nuevos hospitales, se amplían antiguos consultorios, etc.. Para erradicar la pobreza o definir una política de empleo, hay que estudiar el origen del problema. En el campo de la agricultura, se requiere hacer buenas predicciones de la producción (de trigo, por ejemplo) y decidir si estas permitirán satisfacer la demanda. En la explotación de los bosques es importante estimar los volúmenes y la calidad de la madera esperada en una zona dada para la planificación de las cosechas y los requerimientos de la demanda.

Estadísticas y empresas

Una fábrica o una empresa de servicios requiere saber de sus recursos, producción, demanda y la competencia de sus productos. Estos problemas involucran el control de calidad de los productos en los procesos de fabricación y los estudios de mercado, entre otros. Una compañía de Seguros de Vida requiere estimar la probabilidad de que una persona de una cierta edad y cierto sexo fallezca antes de alcanzar una determinada edad, de manera a fijar el monto de su póliza. Un productor de fertilizante tiene que evaluar la eficacia de su producto. Hará, por ejemplo, un experimento para medir el efecto de su fertilizante sobre la cosecha de choclo.

Estadísticas y ciencias

En la investigación de ciencias como la física, la química, la biología o ciencias sociales, se busca verificar las leyes formuladas a partir de experimentos que se analizan mediante métodos estadísticos.

Un físico busca el valor de una constante numérica, que aparece en una relación exacta. Sin embargo, el experimento que le permitirá obtener la constante en el laboratorio conlleva perturbaciones en las mediciones. Tomar el promedio de varias mediciones será la mejor forma de resolver su problema. En la clasificación de planta o animales se usan procedimientos de muestreo aleatorio para contarlos. Las famosas leyes de Mendel, a pesar de referirse a caractéres genéticos cualitativos, pueden considerarse como leyes estadísticas.

Estadísticas y educación

Un psicólogo mide las aptitudes mentales de algunos estudiantes y les da un método de estudio. El rendimiento permitirá evaluar el método de estudio en función de las aptitudes mentales. La psicometría es la rama de la psicología que trata mediciones relativas a habilidades mentales de individuos. En educación, la psicometría permite, mediante tests llevados a escalas numéricas, medir características psicológicas relativas al comportamiento, el aprendizaje y el rendimiento de los estudiantes.

1.3 EL PENSAMIENTO ESTADÍSTICO

Si bien el cálculo de las probabilidades es una teoría matemática abstracta, que deduce consecuencias de un conjunto de axiomas, la estadística trata encontrar un modelo que refleja mejor los datos obtenidos a partir de experimentos y necesita, entonces, dar una interpretación con-creta a la noción de probabilidad. Varias interpretaciones fueron propuestas por los estadísticos, que se pueden resumir en dos puntos de vista diferentes: la noción frecuentista y la noción intuicionista.

El punto de vista *frecuentista* asocia la noción de probabilidad a la noción empírica de frecuencia, basada en observaciones aleatorias repetidas, mientras que el punto de vista *intuicionista* liga la noción de probabilidad al grado de creencia subjetiva que uno tiene sobre la ocurrencia de un suceso.

Todos los días se habla en las noticias de población para referirse a un grupo de personas que tienen algo en común, como la población de los chilenos o la población de los niños de Santiago. Para el estadístico, este concepto se refiere a un conjunto de elementos (personas, objetos, plantas, animales, etc.) sobre los cuales se obtienen informaciones para sacar conclusiones sobre el grupo. Cuando obtener mediciones sobre cada elemento de la población (un censo) resulta ser muy largo y caro, se puede observar una parte de ella (una muestra), es decir solamente un grupo de elementos elegidos de la población.

Un sociólogo quiere, por ejemplo, determinar el ingreso anual promedio de las familias que viven en Santiago. Recolectar esta información en todas las familias en Santiago sería un largo y costoso proceso. El sociólogo podrá entonces usar una muestra. Eso es posible porque no se interesa en el ingreso anual de cada familia en particular, pero sí en el ingreso anual promedio de la totalidad de las familias que viven en Santiago y eventualmente en la repartición de estos ingresos en la población.

Para saber cual es el número total N de peces viviendo en un lago, sería difícil pescarlos todos. Se pueden pescar aleatoriamente algunos, sea $A = 200$ por ejemplo, marcarlos y devolver al lago. Se vuelve a pescar al azar, sea $n = 100$ por ejemplo, y observar el número k de marcados encontrados en la segunda muestra. Se puede estimar al número total N de peces en el lago, suponiendo que

la proporción de peces marcados en el lago y la proporción de peces marcados en la muestra son iguales:

$$\frac{A}{N} = \frac{k}{n} \Rightarrow N = \frac{n}{k}A$$

Por ejemplo, si se encontró $k = 16$ peces marcados en la segunda muestra de $n = 100$ peces, se estimaría que hay $N = \frac{100}{16} \times 200 = 1250$.

Un candidato a una elección presidencial encarga a un centro de estudio de opiniones un análisis sobre el porcentaje de votos que podría obtener en la elección que tendrá lugar en un mes más. El centro de estudio hace un sondeo de opiniones sobre 1500 personas elegidas al azar en la población que votan y le informa al candidato que si la elección tuviera lugar este mismo día tendría 45% de votos contra 55% de su adversario y agrega con un error porcentual de $2,52\%$ con un nivel de confianza de 95%. Con este pronostico el candidato concluye que tiene muy poca posibilidad de ser elegido, salvo si cambia su campaña electoral.

El problema es entonces cómo elegir una muestra para poder sacar conclusiones que sean válidas para la población entera. En este caso cada individuo o elemento de la muestra no tiene un interés por separado, sino, solamente por que es parte de la población. La teoría de muestreo nos ofrece métodos para obtener muestras. Distinguiremos entonces la **estadística descriptiva**, la actividad que consiste en resumir y representar informaciones, de la **inferencia estadística**, un conjunto de métodos que consisten en sacar resultados sobre una muestra para inferir conclusiones sobre la población de donde proviene esta muestra.

Todos los problemas citados anteriormente son distintos; algunos se podrán basar en datos censales y otros en datos muestrales. Pero hay elementos y una línea general del razonamiento que son los mismos para todos los problemas.

Población y muestras

Los datos experimentales son obtenidos sobre conjuntos de individuos u objetos, sobre los cuales se quiere conocer algunas características. Llamaremos **unidad de observación** a estos individuos y la totalidad de estas unidades de observación se llama **población**. La población puede ser finita: la población de un país en una encuesta de opinión; el conjunto de ampolletas fabricadas por una máquina; los árboles de un bosque.

La población puede ser considerada también como infinita y hipótetica: la población de todos los posibles lanzamientos que se puede hacer con una moneda; la población definida por el caudal de un rio; la población definida por el tiempo de vida de una ampolleta; el tiempo de espera en un paradero de buses. En estos casos la población es definida por el conjunto de los reales $\mathbb{R}$ o un intervalo de $\mathbb{R}$ y generalmente tal población esta definida por una variable aleatoria y su distribución de probabilidad.

Frecuentemente la población a estudiar, aún si es finita, es demasiado grande. Se extrae entonces solamente un subconjunto de la población, llamada **subpoblación o muestra** sobre la cual se observan mediciones llamadas **variables**. Los elementos de la muestra podrán ser repetidos o no y el orden de extracción podrá ser relevante o no.

Por ejemplo se toma un subconjunto de la población de un país; se lanza 100 veces una moneda; se considera los tiempos de vida de 150 ampolletas.

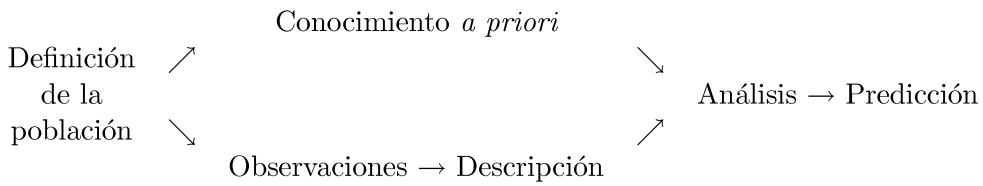
El estadístico trata entonces de inferir informaciones sobre la población a partir de los valores observados en la muestra. La muestra podrá no ser **representativa** de la población en el sentido que algunas características de interés podrán ser sobreestimadas o subestimadas.

Definición 1.3.1 *Se dice que una muestra es representativa de una población si toda unidad de observación podrá aparecer en la muestra y esto con una probabilidad conocida.*

Etapas de un estudio estadístico

Un estudio estadístico se descompone generalmente en varias etapas:

- Definición del problema: objetivos y definición de la población
- Determinación del muestreo.
- Recolección de los datos.
- Análisis descriptivo de los datos.
- Análisis inferencial o matemático de los datos. Se usa toda información útil al estudio
- Conclusión del estudio: Decisión o predicción.



Recolección de los datos

Se distinguen los censos de los muestreos. En un censo los datos se recolectan sobre la totalidad de las unidades de observación de la población considerada y en una muestra se recoge información sólo sobre una parte de la población. *¿Cómo entonces sacar una muestra de una población finita o de una distribución de probabilidad desconocida para obtener informaciones fidedignas sobre la población de la cual provienen?* La forma de elegir la muestra depende del problema (teorías del diseño de muestreo y del diseño de experimentos). Puede ser muy compleja, pero generalmente la muestra está obtenida aleatoriamente y lleva a aplicar la teoría de las probabilidades.

Descripción estadística de los datos

La descripción estadística permite resumir, reducir y presentar gráficamente el contenido de los datos con el objeto de facilitar su interpretación, sin preocuparse si estos datos provienen de una muestra o no. Las técnicas utilizadas dependerán del volumen de las unidades de observación, de la cantidad de las variables, de la naturaleza de los datos y de los objetivos del problema. Esta etapa del estudio es una ayuda para el análisis inferencial.

Análisis inferencial o matemático de los datos

El análisis, la etapa más importante del razonamiento estadístico, se basa en un modelo matemático o probabilístico.

La inferencia estadística consiste en métodos para extrapolar características obtenidas sobre una muestra hacia la población. Se basa en modelos que dependen de los objetivos del estudio, de los datos y eventualmente del conocimiento *a priori* que se puede tener sobre el fenómeno estudiado. El modelo no está en general totalmente determinado (es decir, se plantea una familia de modelos de un cierto tipo); por ejemplo, la familia de las distribuciones normales, la familia de las distribuciones de Poisson o Beta o un modelo lineal. Estos modelos tendrán algunos elementos indeterminados llamados **parámetros**. Se trata entonces de precisar lo mejor posible tales parámetros desconocidos a partir de datos empíricos obtenidos sobre una muestra: **es el problema de estimación estadística**. Por otro lado, antes o durante el análisis, se tienen generalmente consideraciones teóricas respecto del problema estudiado y se trata entonces de comprobarlas o rechazarlas a partir de los datos empíricos: **es el problema de test estadístico**.

Por ejemplo, se quiere estudiar la duración de las ampolletas de $100W$ de la marca ILUMINA. No podemos esperar que se quemen todas las ampolletas producidas durante un período dado para sacar ciertas conclusiones. Se observa entonces el tiempo de duración de una muestra de 500 ampolletas, por ejemplo. Nos preguntamos entonces:

- ¿Cómo seleccionar las 500 ampolletas?
- ¿Cómo extrapolar o inferir las conclusiones obtenidas sobre la muestra de las 500 ampolletas a la totalidad de las ampolletas ILUMINA de $100W$?

Se responde a la primera pregunta con la teoría de muestreo y a la segunda con la inferencia estadística.

Decisión o predicción

El análisis esta condicionado por la finalidad del estudio, que consiste generalmente en tomar una decisión o proceder a alguna predicción. Por ejemplo, decidir si las ampolletas ILUMINA están conforme a las normas de calidad (duración 2500 horas), si un tratamiento es eficaz para combatir la hipertensión. Predecir el IPC del próximo mes, las temperaturas mínima y máxima de mañana en Santiago, el porcentaje de votos de un candidato en una elección, a partir de algunas muestras.

1.4 MUESTREO: VER PARA CREER

Un problema importante de la estadística es la selección de una muestra. Esta dependerá de la población, de las mediciones que se recolectarán sobre las unidades de observación y del problema a estudiar. La teoría de muestreo consiste en una collección de métodos particulares para diferentes situaciones.

En los problemas citados anteriormente, el problema sería cómo seleccionar las 500 ampolletas ILUMINA o cómo extrapolar las conclusiones obtenidas de la muestra a la totalidad de las ampolletas, o predecir el resultado a una elección. Por lo tanto, nos preguntamos

¿Qué esperamos de una muestra para responder correctamente a los estudios planteados?

Para obtener un valor aceptable de la duración media de las ampolletas, hay que seleccionar correctamente la muestra con un tamaño de muestra suficientemente grande. Una muestra no está correctamente seleccionada sino se obtiene a partir de toda la población. En este caso puede resultar sesgada, es decir, algunas características medidas en la muestra podrían sobreestimar o subestimar las mismas características de la población. Otro problema es el tamaño de la muestra, que puede ser demasiado pequeño para la variabilidad de la variable estudiada la población y sus características.

El sesgo puede provenir de diferentes fuentes de errores de procedimiento, en particular de la forma de extraer la muestra y de la forma de medir o del problema que se quiere resolver.

La forma de evitar el problema de la extracción consiste en sacar la muestra de manera aleatoria a partir de la población entera. Este método se basa en el principio de que la muestra debe obtenerse de la manera más objetiva posible.

La determinación del tamaño de la muestra es lo más delicado. Veremos que el error o la precisión del resultado, en definitiva, depende no solamente del tamaño de la muestra sino que también de la variabilidad en la población. Sin embargo, en la práctica no se conoce en general la variabilidad en la población, más aún, es una de la característica de la población que se quiere conocer. Por otra parte, no siempre se puede tomar el tamaño de muestra que uno quisiera debido a los costos de obtención de los datos. Se debe buscar entoncess un compromiso entre la precisión deseada y los costos.

En resumen, una muestra está correctamente seleccionada cuando es sacada de manera aleatoria a partir de toda la población y es suficientemente grande para tener una precisión aceptable. Las condiciones que debe tener una muestra son:

- Que no tenga *sesgo*, es decir que las características de la muestra no sobreestimen o no subestimen las características de la población que se pretende evaluar.
- Que todo elemento de la población tenga la posibilidad de ser elegido en la muestra. Además la selección debería ser objetiva, es decir sin que ningún factor personal intervenga. De aquí que se da un carácter aleatorio al muestreo, y se asigna a cada elemento de la población una probabilidad de selección no nula.
- Para poder inferir hacia la población debemos poder dar una formalización matemática que permita estudiar las propiedades de la muestra, especialmente los errores asociados al muestreo. Debemos entonces conocer las probabilidades asignadas a cada elemento de la población.

> *Un muestreo se dice aleatorio o probabilístico si todo elemento de la población tiene una probabilidad* **no nula** *y* **conocida** *de ser seleccionado en la muestra.*

El muestreo aleatorio se basa entonces en el principio de una muestra *objetiva* donde todo elemento tiene cierta probabilidad conocida de estar seleccionado.

Los valores de las variables obtenidos sobre los elementos de la muestra se llaman **valores muestrales**. Si la muestra se obtiene de un muestreo aleatorio, los valores muestrales son variables

aleatorias cuya distribución depende de la población. Las características calculadas a partir de los valores muestrales son aleatorias también.

Ahora bien, cuando se emiten conclusiones sobre una población sólo a partir de valores obtenidos sobre una muestra aleatoria, están afectadas de **errores debidos al muestreo** y el muestreo no es la única fuente de error. Se tienen generalmente a los **errores de medición**. Los errores de medición pueden influir sobre la precisión de las conclusiones. Si tienen un carácter aleatorio, pueden compensarse o bien ser sistemáticos.

Veremos que los errores de muestreo decrecen cuando el tamaño de la muestra crece, pero los errores de medición crecen generalmente con este tamaño. Lo ideal es entonces tener un buen equilibrio entre estos dos tipos de errores. Pero es difícil en la práctica evaluar los errores de medición.

La variabilidad real en la población es otro factor importante que interviene en la variabilidad de los resultados obtenidos de una muestra (Esquema en la figura 1.1).

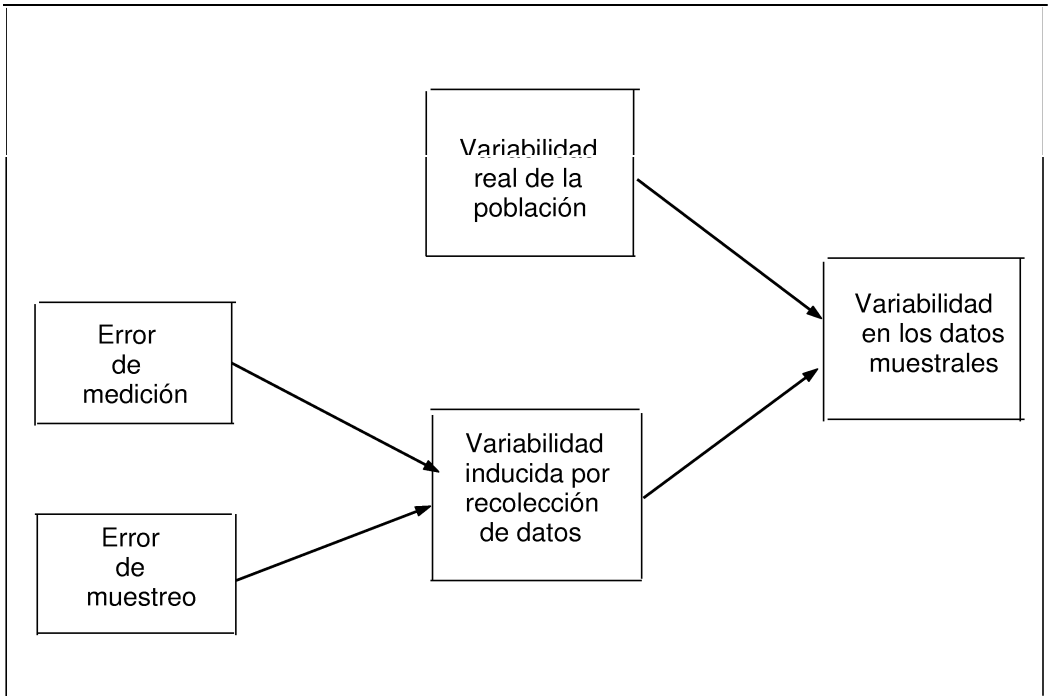


Figura 1.1: Esquema de las variabilidades

Consideremos el caso de una población finita de tamaño N. Se llama **fracción de muestreo** a la proporción entre el tamaño n de la muestra y el tamaño N de la población:

$$\frac{n}{N}$$

La teoría de muestreo permite determinar la fracción de muestreo para un error de muestreo dado y definir un procedimiento para seleccionar las unidades de observación de la muestra de manera de producir una muestra **representativa** de la población de donde están extraídas, es decir para que la muestra dé un imagen reducida pero fiel de la población. Hay varias formas de obtener

la representatividad dependiendo de la complejidad de la población tratada. Se distinguen los muestreos aleatorios de los muestreos sistemáticos.

Cualquier sea el tipo de muestreo elegido, la población debe estar perfectamente definida y todos sus elementos identificables sin ambigüedad.

El muestreo aleatorio simple (m.a.s.) permite sacar muestras de tamaño dado, cada una equiprobable, de una población finita o infinita. Se debe distinguir el m.a.s. con reemplazo del m.a.s. sin reemplazo.

En lenguaje probabilista:

- Dado un experimento aleatorio $\mathcal{E}$ y una población (o espacio muestral) $\mathcal{P}$ de sucesos elementales equiprobables, el conjunto de n repeticiones independientes del experimento $\mathcal{E}$ es **una muestra aleatoria simple con reemplazo de tamaño** n. La muestra obtenida es entonces una n-tupla de $\mathcal{P}$.
- **Una muestra aleatoria simple sin reemplazo** (o sin repetición) se obtiene de la población $\mathcal{P}$ de sucesos elementales equiprobables realizando el experimento $\mathcal{E}$:
  - sobre $\mathcal{P}$. Se obtiene un suceso a_1 con equiprobabilidad;
  - sobre $\mathcal{P} \setminus \{a_1\}$. Se obtiene un suceso a_2 con equiprobabilidad;
  - sobre $\mathcal{P} \setminus \{a_1, a_2\}$. Se obtiene un suceso a_3 con equiprobabilidad, etc... hasta completar la muestra de tamaño n.

  La muestra obtenida es entonces un subconjunto de $\mathcal{P}$. Se observará que los sucesos a_i no son independientes en este caso.

En una población finita de tamaño N con todos sus elementos equiprobables, el número total de muestras posibles sin reemplazo de tamaño n es igual a $\left(\begin{array}{c} N \\ n \end{array} \right)$. Luego cada muestra tiene una probabilidad igual a:

$$\frac{1}{\left(\begin{array}{c} N \\ n \end{array} \right)}$$

En el caso del muestreo aleatorio con reemplazo el número total de muestras posibles es igual a $\left(\begin{array}{c} N+n-1 \\ n \end{array} \right)$ o sea $\left(\begin{array}{c} N+n-1 \\ N-1 \end{array} \right)$.

En efecto el número total de muestras posibles con reemplazo es el número de soluciones del problema *(Pb)*:

$$(Pb): \quad x_1 + x_2 + ... + x_N = n \quad con \quad x_i \in I\!N$$

Sea $f(N,n)$ el número de soluciones del problema *(Pb)* que buscaremos por inducción sobre N.

Para $N = 1$, tenemos una sola solución: $f(1,n) = 1$.
Para $N = 2$, observamos que $x_2 = n - x_1$ con $x_1 = 0, 1, ..., n$. Tenemos entonces $n+1$ soluciones: $f(2,n) = n+1$ es decir $f(2,n) = \left(\begin{array}{c} 1+n \\ 1 \end{array} \right)$.

Supongamos que es cierto para $N-1$: $f(N-1,n) = \begin{pmatrix} N+n-2 \\ N-2 \end{pmatrix}$.

Para N, la ecuación del problema *(Pb)* se puede escribir:

$$(Pb): \quad x_1 + x_2 + ... + x_{N-1} = n - x_N \quad con \quad x_N = 0, 1, ..., n$$

Lo que equivale a escribir las $n+1$ ecuaciones:

$$\begin{cases} x_1 + x_2 + ... + x_{N-1} = n \\ x_1 + x_2 + ... + x_{N-1} = n-1 \\ ... \\ x_1 + x_2 + ... + x_{N-1} = 0 \end{cases}$$

Observando que la primera ecuación tiene $f(N-1,n)$ soluciones, la segunda $f(N-1,n-1)$,..., y la última tiene $f(N-1,0)$ ecuaciones, el número de soluciones del problema *(Pb)* es

$$M = \sum_{j=0}^{n} f(N-1,j)$$

$$M = \begin{pmatrix} N+n-2 \\ N-2 \end{pmatrix} + \begin{pmatrix} N+n-3 \\ N-2 \end{pmatrix} + ... + \begin{pmatrix} N-2 \\ N-2 \end{pmatrix}$$

Como

$$\begin{pmatrix} m \\ p \end{pmatrix} = \begin{pmatrix} m-1 \\ p-1 \end{pmatrix} + \begin{pmatrix} m-1 \\ p \end{pmatrix} = \sum_{j=1}^{m-p+1} \begin{pmatrix} m-j \\ p-1 \end{pmatrix}$$

$$M = \begin{pmatrix} N+n-1 \\ N-1 \end{pmatrix} = f(N,n)$$

El muestreo aleatorio simple es un método para obtener muestras de tamaño fijo de tal forma que todas las muestras de mismo tamaño tengan la misma probabilidad de selección. Pero no es la única forma de proceder.

El muestreo sistemático se basa en una regla de selección no aleatoria efectuando saltos en una lista de los elementos de la población. Por ejemplo en una población formada de mil pozos listados, se determina una muestra de 100 pozos seleccionando un pozo de cada 10 de la lista. Para que este procedimiento produzca un muestreo aleatorio simple basta que la lista de los elementos sea construida al azar.

Este procedimiento tiene entonces una ventaja práctica, pero obliga a controlar que estos pozos no tengan justamente algunas particularidades.

Sin embargo, se puede buscar asegurar una mejor representatividad relativa a un aspecto particular. Si las unidades de observación son clasificadas según un criterio, por ejemplo los pozos sean ordenados en la lista en función de su profundidad (de menor a mayor profundidad), y si además este criterio esta correlacionado con las variables de interés, entonces se tendra en la muestra pozos de todas las profundidades. Pero, lo anterior, requiere conocer la profundidad para todos los pozos de la población.

El muestreo a probabilidades desiguales permite atribuir a ciertas unidades de observación una probabilidad mayor que a otras. Se usa cuando las unidades de observación de la población tienen tamaño distintos, y se estima que mientras más grande, más información aporta. Se toma entonces probabilidades proporcionales al tamaño de la observación. Por ejemplo, para la población de las empresas en Chile, se pueden seleccionar proporcionalmente a su número de empleados; para la población de los campos agrícolas, se elige proporcionalmente a la superficie.

El muestreo estratificado se basa en una partición de la población en clases homogéneas (con respecto a algunas características definidas a priori) llamadas **estratos**. Se hace un muestreo aleatorio al interior de cada estrato y los muestreos son independientes entre los estratos. Este tipo de muestreo permite aplicar métodos de muestreo diferentes en los estratos. Su objetivo es disminuir el error de muestreo para un tamaño muestral total fijo. La repartición de la muestra entre los estratos depende si se busca disminuir el error muestral a nivel global o a nivel de cada estrato.

El inconveniente de este tipo de muestreo es que la estratificación puede resultar eficaz para algunas variables, en particular las variables de estratificación, pero muy poco eficaz para otra.

Sea por ejemplo la población de todos los hogares de la Región Metropolitan. Un muestreo estratificado según dos criterios - comuna y tipo de alojamiento- y un muestreo aleatorio simple con una fracción de muestreo igual al interior de los estratos permite alcanzar una mejor representatividad. Conociendo, por ejemplo, el tamaño de los hogares de toda la población se podría hacer un muestreo sistématico en vez de un muestreo aleatorio simple.

El muestreo por etapas se usa en caso de encuestas complejas. Si consideramos la población de todos los hogares chilenos, un muestreo estratificado según la comuna llevaría a demasiado estratos. Se podría estratificar según la región, o bien proceder en dos etapas: seleccionar al azar algunas comunas (unidades de observación primarias) y en cada comuna seleccionada elegir una muestra de hogar. En cada etapa se puede usar probabilidades iguales o desiguales.

El muestreo por conglomerados es un caso particular de muestreo por etapas, donde en la última etapa se selecciona todas las unidades de observación. Por ejemplo, en la primera etapa se elige algunas comunas, en la segunda etapa se elige manzanas y en la tercera y última etapa se toma todos los hogares de las manzanas elegidas.

Capítulo 2

DISTRIBUCIONES EN EL MUESTREO

2.1 INTRODUCCIÓN

Los métodos estadísticos permiten confrontar modelos matemáticos o probabilísticos con los datos empíricos obtenidos sobre una muestra aleatoria:

> ***Considerando mediciones obtenidas sobre una muestra de tamaño*** n***,***
> ***se busca deducir propiedades de la población de la cual provienen.***

Ejemplo 2.1.1 Se saca una muestra al azar de 500 ampolletas ILUMINA del mismo tipo en un proceso de producción y se considera sus tiempos de vida. Si el proceso de fabricación no ha cambiado, las fluctuaciones entre las ampolletas observadas pueden considerarse como aleatorias y además que todas las observaciones provienen de una misma variable aleatoria X de distribución desconocida abstracta llamada **distribución de población** $F(x) = I\!P(X \leq x)$ del tiempo de vida de este tipo de ampolleta.

Ejemplo 2.1.2 El ministerio de la salud quiere conocer la talla promedio μ de las mujeres chilenas mayores de 15 años. En este caso la población no es abstracta ya que se podría medir la talla de todas las chilenas mayores de 15 años y entonces determinar la distribución de población y, por lo tanto, calcular la talla media de la población. Sin embargo es muy difícil de realizar en la práctica aún si la población es finita, dado que es muy grande. La función de distribución de población se considera entonces como continua y incluso abstracta con una expresión teórica (una distribución normal en general) y se usa una muestra al azar, por ejemplo de 1000 chilenas mayores de 15 años y se mide sus tallas.

Ejemplo 2.1.3 La compañía Dulce compró una máquina para llenar sus bolsas de azúcar de 1 kg. La máquina no puede ser perfecta y el peso varía de una bolsa a otra. Si se acepta una variación en el peso de las bolsas, esta debería ser pequeña y la media del peso debería ser igual a 1 kg. Un buen modelo estadístico para el peso es una distribución Normal de media nula y varianza pequeña (el modelo Normal se obtiene de la teoría de los errores párrafo 1.1) .

Ejemplo 2.1.4 Un candidato a una elección presidencial quiere planear su campaña electoral a partir de un sondeo de opiniones sobre una muestra de votantes. ¿Los resultados del sondeo le permitirían inferir el resultado de la elección? Se puede construir un modelo de Bernoulli cuyo parámetro es la probabilidad que un elector vote por el candidato. El candidato saber si esta probabilidad será mayor que 50%.

Ejemplo 2.1.5 Una máquina produce diariamente un lote de piezas. Un criterio basado sobre normas de calidad vigentes permite clasificar cada pieza fabricada como defectuosa o no defectuosa. El cliente aceptará el lote si la proporción de piezas θ defectuosas contenidas en el lote no sobrepasa el 2%. El fabricante tiene que controlar entonces la proporción θ de piezas defectuosas contenidas en cada lote que fábrica. Pero si la cantidad de piezas N de cada lote es muy grande, no podrá examinar cada una para determinar el valor de θ. Como el ejemplo anterior se puede construir un modelo de Bernoulli cuyo parámetro aquí es la probabilidad que una pieza este defectuosa. El cliente querá saber entonces si esta probabilidad sera mayor que el 2%.

Si se tiene una sola variable aleatoria X cuya función de distribución F de población es generalmente desconocida, obteniendo observaciones de esta variable X sobre una muestra, buscaremos conocer la función de distribución F. Los valores $x_1, x_2, ..., x_n$ de la v.a. X obtenidos sobre una muestra de tamaño n son **los valores muestrales**.

Se quiere saber entonces de que manera estos valores muestrales procuren información sobre algunas características de la población. Esta pregunta no es posible de responder directamente, hay que transformarla en otra pregunta: **si suponemos que la población tiene una distribución F_o ¿cual seriá la probabilidad de obtener la muestra que obtuvimos?**

Si la probabilidad es pequeña, se concluye que la distribución de la población no es F_o. Si la probabilidad es alta, aceptamos F_o. Se busca, entonces, **estimar** características de la distribución F_o a partir de los valores muestrales, por ejemplo, la media y la varianza.

2.2 TIPOS DE VARIABLES

La cantidad y la naturaleza de las cacterísticas que se puede medir sobre los elementos de una población $\mathcal{P}$ son muy diversos. Supondremos aquí una sola variable en estudio que es una función

$$X : \mathcal{P} \longrightarrow Q$$

Se distingue la naturaleza de la variable X según el conjunto Q:

- variable cuantitativa (también llamada intervalar) si Q es un intervalo de $\mathbb{R}$ ó todo $\mathbb{R}$. Por ejemplo: la edad, el peso ó la talla de una persona. Estas variables se consideran como reales continuas aún si se miden de manera discontinua (en año, en kg ó cm).
- variable discreta si Q es un subconjunto de $\mathbb{N}$. Por ejemplo, el número de hijos de una familia. Se habla de variable discreta.
- variable cualitativa (o nominal) si Q es un conjunto finito de atributos (ó modalidades ó categorías) no numéricos. Por ejemplo: el estado civil, el sexo, la ocupación de una persona ó los nombres de los candidatos a una elección.

- variable ordinal si Q es un conjunto de atributos no numéricos que se pueden ordenar. Por ejemplo, el ranking de la crítica cinematográfica.

Los métodos estadísticos dependen del tipo de variables consideradas. Es entonces interesante poder transformar una variable de un tipo a otro. Por ejemplo, la edad se puede transformar en una variable nominal o ordinal considerando como conjunto Q un conjunto de clases de edad. Según la precisión requerida de la variable edad y los métodos utilizados se usará la edad como variable cuantitativa o ordinal.

2.3 DISTRIBUCIÓN EMPÍRICA

En este párrafo vemos la distribución de los valores muestrales obtenidos a partir de un muestreo aleatorio simple. Distinguimos el estudio según el tipo de variable.

2.3.1 Caso de variables numéricas (reales)

Consideramos una muestra aleatoria simple $x_1, ...x_n$ independientes e idénticamente distribuidas (i.i.d.) del ejemplo 2.1.1 del tiempo de vida de las ampolletas ILUMINA. La proporción de ampolletas con tiempo de vida menor que x define una función de distribución, que depende de la muestra (Figura 2.1).

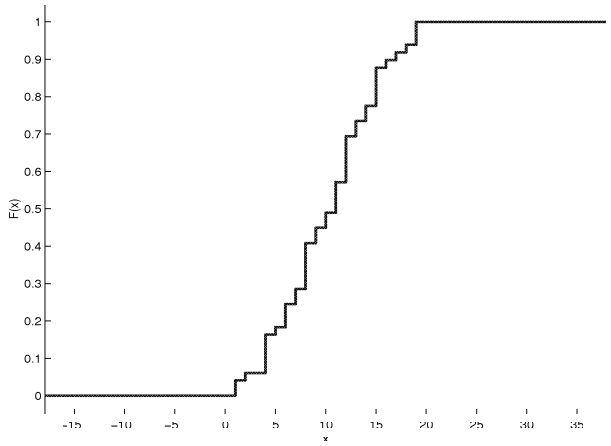


Figura 2.1: Una función de distribución empírica

Definición 2.3.1 *Sean $x_1, x_2, ..., x_n$, los valores muestrales obtenidos de un m.a.s. de X. Se llama la función de distribución empírica a la proporción de observaciones de la muestra inferiores o iguales a x;*

$$F_n(x) = \frac{Card\{x_i | x_i \leq x\}}{n}$$

La función de distribución empíca $F_n(x)$ tiene las propiedades de una función de distribución:

- $F_n : I\!R \longrightarrow |0, 1|$.

- El muestreo es equiprobable: si n es el tamaño de la muestra, $p_i = \frac{1}{n}$ para todo elemento de la muestra. Luego $F_n(x)$ es la probabilidad de encontrar una observación x_i menor que x en la muestra.
- $F_n(x)$ es monótona no decreciente; tiene límites a la derecha y a la izquierda; es continua a la derecha; $F(-\infty) = 0$; $F(+\infty) = 1$. Además los puntos de discontinuidad son con salto y en número finito.

Además para x fijo, $F_n(x)$ es una variable aleatoria y $nF_n(x)$ es una v.a. igual a la suma de variables de Bernoulli independientes de mismo parámetro $F(x)$. En efecto, si definamos

$$Y_i = \begin{cases} 1 & si \quad X_i \leq x \\ 0 & si \quad X_i > x \end{cases}$$

Las variables Y_i son variables aleatorias de Bernoulli de parámetro igual a la probabilidad que $X_i \leq x$ es decir $F(x)$. Luego $nF_n(x) = \sum_{i=1}^{n} Y_i$ sigue una distribución binomial: $nF_n(x) \sim \mathcal{B}(n, F(x))$.

Teorema 2.3.2 *Para todo x, $F_n(x)$ converge casi-seguramente hacia el valor teórico $F(x)$ (se denota $F_n(x) \xrightarrow{c.s.} F(x)$).*

Demostración Como $nF_n(x) \sim \mathcal{B}(n, F(x))$, se concluye de la ley fuerte de los grandes números que:

$$I\!P(\lim_n F_n(x) = F(x)) = 1$$

■

Se espera entonces que la función de distribución empíca $F_n(x)$ no sea tan diferente de la función de distribución de la población cuando n es suficientemente grande. Se tiene dos otros resultados que lo confirman (no se demuestran estos teoremas).

Teorema 2.3.3 *(Glivenko-Cantelli)*

$$D_n = \sup_x \mid F_n(x) - F(x) \mid \longrightarrow 0$$

Teorema 2.3.4 *(Kolmogorov) La distribución asintótica de D_n es conocida y no depende de la distribución de X:*

$$\lim_{n\to\infty} I\!P(\sqrt{n}D_n < y) = \sum_{k=-\infty}^{+\infty} (-1)^k exp(-2k^2y^2)$$

2.3.2 Caso de variables nominales u ordinales

En el ejemplo 2.1.4 de la elección presidencial, la población $\mathcal{P}$ esta constituida por la totalidad de los N votantes. Si hay r candidatos, la variable X de interés es el voto que va emitir el votante:

$$X : \mathcal{P} \longrightarrow Q$$

donde $Q = \{q_1, q_2, ..., q_r\}$ es el conjunto de los r candidatos. Si el votante i ha elegido el candidato q_j, $X_i = q_j$ $(i = 1, 2, ..., N)$. Es una variable nominal y los candidatos son los atributos $q_1, q_2, ..., q_r$.

Si m_j es el número de votos que recibe el candidatos q_j, su proporción de votos en la población es $p_j = \frac{m_j}{N} = \frac{card\{X(i)=q_j | i=1,2,...,N\}}{N}$.

Se interpreta p_j como la probabilidad que un votante vote por el candidato q_j. El conjunto $p_1, p_2, ..., p_r$ constituye la función de probabilidad definida sobre el conjunto Q de los candidatos relativa a la población total de los votantes: $I\!P(X = q_j) = p_j \ (\forall j = 1, ..., r)$.

Una encuesta de opiniones previa a la elección tratará de acercarse a los valores $p_1, p_2, ..., p_r$ de la función de probabilidad de la población.

Sea una muestra aleatoria de $n = 1500$ personas en la cual los candidatos recibieron las proporciones de votos $f_n(q_1)$, $f_n(q_2)$,...,$f_n(q_r)$, con $f_n(q_j) = \frac{Card\{X_i=q_j\}}{n}$, $(\forall j = 1, ..., r)$. Estas proporciones pueden escribirse como la media de variables de Bernoulli.

Sean las r variables de Bernoulli Y_j $(\forall j)$ asociadas a la variable X:

$$Y_j(i) = \begin{cases} 1 & si \quad X_i = q_j \\ 0 & si \quad X_i \neq q_j \end{cases}$$

Si $Y_j(1), Y_j(2), ..., Y_j(n)$, $(\forall j)$ son los valores muestrales,

$$f_n(q_j) = \frac{\sum_{i=1}^{n} Y_j(i)}{n} \qquad \forall j$$

Como la distribución $nf_n(q_j) \sim \mathcal{B}(n, p_j)$, $f_n(q_j) \xrightarrow{c.s.} p_j (\forall q_j \in Q)$.

Se observará que las r v.a. binomiales $nf_n(q_j)$ no son independientes entre si: $\sum_j nf_n(q_j) = n$. Veremos más adelante que estas r variables binomiales forman un vector aleatorio llamado *vector multinomial*.

2.4 DISTRIBUCIONES EN EL MUESTREO Y EN LA POBLACIÓN

Sea una v.a. X de distribución F. Sean $x_1, x_2, ..., x_n$ valores muestrales independientes obtenidos sobre una muestra aleatoria de tamaño n de esta distribución. Si nos interesa estudiar la media μ de la población (esperanza de la distribución F), la muestra nos permitirá **estimarla**. Pero si se saca otra muestra del mismo tamaño obtendremos posiblemente otro valor de la estimación de μ. **El resultado de la estimación es aleatorio**. El carácter aleatorio del resultado proviene de la aleatoriedad de la muestra y además su distribución depende del tamaño y del tipo de muestreo que se aplique. Es decir, los valores muestrales y las funciones de estos que permiten estimar son variables aleatorias.

Vimos la relación entre la distribución empírica y la distribución de población, luego, como la distribución empírica permite acercarse a la distribución de población. De la misma manera el estudio de la relación entre de las distribuciones de las estimaciones y la distribución de la población permitirá hacer inferencia de los valores muestrales hacia características de la población tales como μ.

Definición 2.4.1 *Las funciones de los valores muestrales son variables aleatorias llamadas* **estadísticos**[1] *y las distribuciones de los estadísticos se llaman* **distribuciones en el muestreo**.

[1]No confundir con el estadístico, profesional o investigador que trabaja en estadística

Generalmente no ignoramos todo de la distribución de la población y por eso hacemos supuestos sobre está. Es decir, suponemos que la distribución de población pertenece a una familia de distribuciones teóricas. Por ejemplos, si X es la talla de los hombres adultos chilenos, podremos suponer que X sigue una distribución normal, o si X es la proporción del tiempo ocupado diariamente mirando TV, podremos suponer una distribución beta ó si X es el número de clientes en la cola de una caja de una banco podremos suponer una distribución de Poisson. En este caso, solamente algunas características quedarán desconocidas, como por ejemplo la media y la varianza para la distribución normal ó el parámetro λ para la distribución de Poisson. Estas características desconocidas de la distribución de la población son llamados **los parámetros** que buscamos a estudiar. Los estadísticos y sus distribuciones en el muestreo (ó sus distribuciones asintóticas cuando se hace tender n el tamaño de la muestra a $+\infty$) permiten **estimar** los parámetros desconocidos de la distribución de la población.

Se llama **estimador** de θ al estadístico que permite estimar un parámetro θ de una distribución de población. Como el estimador es una variable aleatoria, sus fluctuaciones tienen que estudiarse. Una medición de las fluctuaciones de un estimador T en el muestreo con respecto al parámetro θ de la distribución de población es el **error cuadrático medio** $E[(S-\theta)^2]$ ó su raíz llamada **el error estándar**, que permite medir la precisión del estimador T con respecto al parámetro θ. El problema es que no se conoce a θ.

Veamos a continuación las propiedades de algunos estadísticos conocidos, tales como la proporción, la media o la varianza en la muestra.

2.4.1 Proporción muestral

Supongamos que $x_1, x_2, ..., x_n$ son los valores muestrales i.i.d obtenidos de una población de Bernoulli de parámetro p.

Consideremos, en primer lugar, el caso de una población infinita o una población finita con reemplazo. Por ejemplo, $x_i = 1$ si se saca "cara" y $x_i = 0$ si se saca "sello" en el lanzamiento i de n lanzamientos independientes de una moneda. El parámetro p es la probabilidad de sacar "cara", que vale $\frac{1}{2}$ en el caso de una moneda equilibrada. O bien en un proceso de control de calidad, $x_i = 1$ si la pieza fabricada i es defectuosa y $x_i = 0$ en el caso contrario. La probabilidad p es la probabilidad de que una pieza sea defectuosa en este proceso y $1-p$ es la probabilidad que no sea defectuosa.

Se define la proporción muestral o empírica como $f_n = \sum_{i=1}^{n} \frac{x_i}{n}$ la proporción de caras (ó piezas defectuosas) encontradas entre las n observadas. Veamos que nf_n sigue una distribución $\mathcal{B}(n,p)$:

$$P(f_n = \frac{k}{n}) = P(nf_n = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \quad (k = 0, 1, ..., n)$$

Tenemos $E(f_n) = p$ y $Var(f_n) = p(1-p)/n$. Es decir que la distribución de la proporción empírica f_n esta centrada en el parámetro p y su dispersión depende del tamaño n de la muestra:

$$E((f_n - p)^2) = Var(f_n) = \frac{p(1-p)}{n}$$

El error estándar es entonces: $\varepsilon(f_n - p) = \sqrt{\frac{p(1-p)}{n}}$

Observamos que se tiene la convergencia en media cuadrática:

$$f_n \xrightarrow{m.c.} p$$

En efecto $[\varepsilon(f_n - p)]^2 = E((f_n - p)^2) \longrightarrow 0$

Además se tienen las otras convergencias de f_n hacia p (en probabilidad y casi segura): La convergencia en media cuadrática implica la convergencia en probabilidad ó por la ley débil de los grandes números: la diferencia $|f_n - p|$ es tal que para $\epsilon > 0$ dado:

$$\lim_{n\to(\infty)} I\!P(|f_n - p| < \epsilon) = 1$$

La convergencia casi segura: $f_n \xrightarrow{c.s.} p$, es decir

$$I\!P(\lim_{n\to(\infty)} f_n = p) = 1$$

Además se tiene la convergencia en ley hacia una normal: $f_n \xrightarrow{ley} \mathcal{N}(p, p(1-p)/n)$.

En el caso de una población finita de tamaño N con un muestreo sin reemplazo se obtiene una distribución hipergeométrica:

$$I\!P(nf_n = k) = \frac{\begin{pmatrix} Np \\ k \end{pmatrix}\begin{pmatrix} N(1-p) \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

Se obtiene en este caso un error estándar: $\varepsilon(f_n - p) = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$

Si el tamaño N de la población es grande con respecto al tamaño de la muestra, se tienen los mismos resultados que los del muestreo con reemplazo. Si N es pequeño, conviene usar los resultados del muestreo sin reemplazo. La última formula muestra que el tamaño de la muestra necesario para alcanzar un error ε dado es casi independiente del tamaño N de la población:

$$n = \frac{Np(1-p)}{p(1-p) + \varepsilon^2(N-1)}$$

Se presenta a continuación los tamaños muestrales necesarios para obtener un error $\varepsilon = 0.05$ y $\varepsilon = 0.025$ cuando $p = 0.5$ (Tabla 2.1). Se observa que el tamaño de la muestra aumenta poco cuando aumenta el tamaño de la población, pero que aumenta mucho cuando se quiere disminuir el error estándar. Para N muy grande se requiere observar cuatro veces más unidades para disminuir el error a la mitad.

N	500	1000	5000	10000	50000	∞
n para $\varepsilon = 0.05$	83	91	98	99	100	100
n para $\varepsilon = 0.025$	222	286	370	385	397	400

Tabla 2.1: Tamaño de la muestra, tamaño de la población y error estándar

2.4.2 Media muestral

Sean $x_1, x_2, ..., x_n$, los valores muestrales i.i.d. de una v.a. X. Se define la **media muestral o media empírica** como

$$\bar{x}_n = \sum_{i=1}^{n} \frac{x_i}{n}$$

Si la distribución de población tiene como esperanza μ, $E(x_i) = \mu$ y $Var(x_i) = \sigma^2$ para todo i, entonces $E(\bar{x}_n) = \mu$. Lo que significa que el **promedio** de los valores $\bar{x}_n$ dados por las distintas muestras de tamaño n coincide con la media μ de la población. Pero para una muestra dada, el valor $\bar{x}_n$ se encontrará en general un poco por debajo ó encima de μ debido a las fluctuaciones del muestreo. La pregunta entonces es ¿Cómo evaluar esta fluctación? La respuesta esta dada por la varianza de $\bar{x}_n$, es decir la dispersión promedio de $\bar{x}_n$ alrededor de μ, que depende de la varianza σ^2 de la población:

$$Var(\bar{x}_n) = \frac{\sigma^2}{n}$$

Observamos que la dispersión de los valores de $\bar{x}_n$ alrededor de μ disminuye cuando el tamaño n de la muestra crece. Además utilizando la desigualdad de Chebychev encontramos que para un ϵ dado:

$$I\!P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

Nota 2.4.2 *Si el muestreo es aleatorio sin reemplazo en una población finita de tamaño N entonces $Var(\bar{x}_n) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$. Cuando la población es infinita ($N \to \infty$) se obtiene la expresión de la varianza del caso de valores muestrales independientes $Var(\bar{x}_n) = \frac{\sigma^2}{n}$.*

Si además la distribución de población es normal entonces la distribución en el muestreo de $\bar{x}_n$ también lo es. Si los valores muestrales x_i no provienen necesariamente de una distribución normal pero si son i.i.d., entonces la distribución asintótica de $\frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}$ es $\mathcal{N}(0,1)$ (TEOREMA CENTRAL DEL LIMITE).

Teorema 2.4.3 *(Liapounoff): Si $x_1, x_2, ..., x_n...$ es una succesión de v.a. independientes tales que*

- *sus varianzas $v_1, v_2, ..., v_n, ...$ son finitas*
- *la suma $S_n = \sum_1^n v_j$ crece con n pero los cocientes $\frac{v_j}{S_n}$ tienden hacia cero cuando n crece (condición de Lindeberg)*

Entonces si $Z_n = \sum_1^n X_j$, la distribución de la v.a. $\varrho_n = \frac{Z_n - E(Z_n)}{\sigma_{Z_n}}$, cuando n aumenta, tiende hacia una forma independiente de las distribuciones de las X_j que es la distribución $\mathcal{N}(0,1)$.

De aquí el rol privilegiado de la distribución normal en estadística. Se observará que la propiedad no es cierta si no se cumple la condición de Lindberg. Muchas distribuciones empíricas son representables por una distribución normal, pero no es siempre el caso. En particular en hidrología , el caudal de los ríos, que es la suma de varios ríos más pequeños, no se tiene la independencia entre las componentes que intervienen y se obtiene distribuciones claramente asimétricas.

2.4.3 Varianza muestral

Sea una m.a.s. $x_1, ...x_n$ i.i.d., con $\mathrm{E}(x_i) = \mu$ y $Var(x_i) = \sigma^2$ $(\forall i)$. Se define la **varianza muestral** o la **varianza empírica** como la dispersión promedio de los valores muestrales con respecto de la media muestral:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

Se puede escribir también:

$$S_n^2 = \frac{1}{n}\sum x_i^2 - \bar{x}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - (\bar{x}_n - \mu)^2$$

Propiedades:

- $S_n^2 \xrightarrow{c.s.} \sigma^2$ $\quad (\frac{1}{n}\sum_{i=1}^{n} x_i^2 \xrightarrow{c.s.} E(X^2)$ y $\bar{x}_n^2 \xrightarrow{c.s.} [E(X)]^2)$.

- Cálculo de $E(S_n^2)$
  $E(S_n^2) = E(\frac{1}{n}\sum(x_i^2 - \bar{x}_n)^2) = E(\frac{1}{n}\sum(x_i^2 - \mu)^2 - (\bar{x}_n - \mu)^2)$

  $E(S_n^2) = \frac{1}{n}\sum Var(x_i) - Var(\bar{x}_n) = \frac{1}{n}\sum\sigma^2 - \frac{\sigma^2}{n}$

  $E(S_n^2) = \frac{n-1}{n}\sigma^2 \longrightarrow \sigma^2$.

- Cálculo de $Var(S_n^2)$
  $Var(S_n^2) = \frac{n-1}{n^3}((n-1)\mu_4 - (n-3)\sigma^4)$

  en que $\mu_4 = E((X-\mu)^4)$ es el momento teórico de orden 4 de la v.a. X.

  Se deja este cálculo como ejercicio.

  $Var(S_n^2) \approx \frac{\mu_4 - \sigma^4}{n} \longrightarrow 0$.

- $S_n^2 \xrightarrow{m.c.} \sigma^2$ $\quad (E((S_n^2 - \sigma^2)^2) \longrightarrow 0)$.

- Cálculo de $Cov(\bar{x}_n, S_n^2)$
  $Cov(\bar{x}_n, S_n^2) = E((\bar{x}_n - \mu)(S_n^2 - \frac{n-1}{n}\sigma^2))$

  $Cov(\bar{x}_n, S_n^2) = E[\frac{1}{n}\sum(x_i - \mu)(\frac{1}{n}\sum(x_j - \mu)^2 - (\bar{x}_n - \mu)^2 - \frac{n-1}{n}\sigma^2)]$

  Como $E(x_i - \mu) = 0 \ \forall i$ y $E(x_i - \mu)(x_j - \mu) = 0 \ \forall (i,j)$

  $Cov(\bar{x}_n, S_n^2) = \frac{1}{n^2}E(\sum(x_i - \mu)^3) - E((\bar{x}_n - \mu)^3)$

  $Cov(\bar{x}_n, S_n^2) = \frac{1}{n^2}E(\sum(x_i - \mu)^3) - \frac{1}{n^3}E(\sum x_i^3)$

  $Cov(\bar{x}_n, S_n^2) = \frac{\mu_3}{n} - \frac{\mu_3}{n^2} = \frac{n-1}{n^2}\mu_3$, donde $\mu_3 = E((X-\mu)^3)$.

  Si $n \to +\infty$, $Cov(\bar{x}_n, S_n^2) \to 0$, lo que no significa que hay independencia. Además si la distribución es simétrica ($\mu_3 = 0$), entonces $Cov(\bar{x}_n, S_n^2) = 0$.

2.4.4 Caso de una distribución normal

Si una m.a.s. $x_1, ...x_n$ i.i.d con $x_i \sim \mathcal{N}(\mu, \sigma^2)$ $(\forall i)$, entonces $\bar{x}_n \sim \mathcal{N}(\mu, \sigma^2/n)$. Además S_n^2 sigue una distribución conocida llamada *ji-cuadrado a n-a grados de libertad* y denotada χ^2_{n-1}.

En efecto $S_n^2 = \frac{1}{n}\sum(x_i - \mu)^2 - (\bar{x}_n - \mu)^2$. Luego $\frac{nS_n^2}{\sigma^2} = \sum(\frac{x_i-\mu}{\sigma})^2 - (\frac{\bar{x}_n-\mu}{\sigma/\sqrt{n}})^2$.

Como las v.a. $(\frac{x_i-\mu}{\sigma})$ son i.i.d. de una $\mathcal{N}(0,1)$, entonces $U = \sum(\frac{x_i-\mu}{\sigma})^2$ es una suma de los cuadrados de n v.a. independientes de $\mathcal{N}(0,1)$ cuya distribución es fácil de calcular y se llama **Ji-cuadrado con n grados de libertad y se denota χ^2_n**. Por otro lado, $(\frac{\bar{x}_n-\mu}{\sigma/\sqrt{n}})^2$, que es el cuadrado de una distribución $\mathcal{N}(0,1)$ sigue una distribución χ^2 con 1 grado de libertad.

Estudiamos entonces **la distribución χ^2_r**

Recordemos en primer lugar la distribución de $Y = Z^2$ cuando $Z \sim \mathcal{N}(0,1)$.

Sea Φ la función de distribución de $Z \sim \mathcal{N}(0,1)$ y F la distribución de $Y = Z^2$:

$$F(y) = I\!P(Y \leq y) = I\!P(Z^2 \leq y) = I\!P(-\sqrt{y} \leq Z \leq \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$$

Se deduce la función de densidad f de Y:

$$f(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} exp(-y/2) \quad \forall y > 0$$

Se dice que Y sigue una distribución Ji-cuadrado con 1 grado de libertad ($Y \sim \chi^2_1$).

Observando que la χ^2_1 tiene una distribución Gamma particular $\Gamma(1/2, 1/2)$, la función generatriz de momentos (f.g.m.) se escribe:

$$\Psi_Y(t) = E(e^{tY}) = (\frac{1}{1-2t})^{1/2} \quad \forall t < \frac{1}{2}$$

Sea entonces $U = \sum_1^r Y_i = \sum_1^r Z_i^2$ en que las Z_i^2 son χ^2_1 independientes, entonces

$$\Psi_U(t) = (\frac{1}{1-2t})^{r/2}$$

que es la f.g.m. de una distribución $Gamma(\frac{r}{2}, \frac{1}{2})$.

De esta manera se deduce la función de densidad de $U \sim \chi^2_r$, una Ji-cuadrado con r g.l.:

$$f(u) = \frac{1}{2^{r/2}} \frac{u^{r/2-1}}{\Gamma(r/2)} exp(-u/2) \quad \forall u > 0$$

Se observa que $E(U) = r$ y $Var(U) = 2r$ y se tiene el siguiente resultado:

Corolario 2.4.4 *La suma de k variables aleatorias independientes y de distribución χ^2 a r_1, r_2,...,r_k g.l. respectivamente sigue una distribución χ^2 a $r_1 + r_2 + ... + r_k$ g.l.*

Aplicamos estos resultados al cálculo de la distribución de S_n^2 cuando $X \sim \mathcal{N}(\mu, \sigma^2)$

Teorema 2.4.5 *Si los valores muestrales $x_1, ...x_n$ son i.i.d. de la $\mathcal{N}(\mu, \sigma^2)$, entonces la v.a. nS_n^2/σ^2 sigue una distribución χ^2_{n-1}*

Demostración Sea $\underline{X}$ el vector de las n v.a. x_i y una transformación ortogonal $\underline{Y} = B\underline{X}$ tal que la primera fila de B es igual a $(1/\sqrt{n}, ..., 1/\sqrt{n})$. Se tiene entonces que:

- $y_1 = \sqrt{n}\bar{x}_n$
- $\sum y_i^2 = \sum x_i^2 = \sum (x_i - \bar{x}_n)^2 + n\bar{x}_n^2$ ($y_2^2 + ... + y_n^2 = nS_n^2$)
- $(y_1 - \sqrt{n}\mu)^2 + y_2^2 + ... + y_n^2 = (x_1 - \mu)^2 + ... + (x_n - \mu)^2$

La densidad conjunta de $y_1, ..., y_n$ es entonces proporcional a:

$$exp\{-(y_1 - \mu\sqrt{n})^2 + y_2^2 + ... + y_n^2\}/2\sigma^2$$

Luego $y_1^2, ..., y_n^2$ son independientes y

$$\sqrt{n}\bar{x}_n = y_1 \sim \mathcal{N}(\sqrt{n}\mu, \sigma^2)$$

$$nS_n^2/\sigma^2 = y_2^2 + ... + y_n^2\}/\sigma^2 \sim \chi^2_{n-1}$$

■

Además $\bar{x}_n$ y S_n^2 son independientes.

Teorema 2.4.6 *Sean $x_1, x_2, ..., x_n$ v.a. i.i.d., entonces $\bar{x}_n$ y S_n^2 son independientes si y sólo si los valores x_i provienen de una distribución normal.*

La demostración que no es fácil se deduce del teorema 2.4.5 y del corolario 2.4.4.

Definamos a continuación la distribución **t de Student**,[2] que tiene muchas aplicaciones en inferencia estadística como la distribución χ^2.

Definición 2.4.7 *Si X e Y son dos v.a. independientes, $X \sim \mathcal{N}(0,1)$ e $U \sim \chi_r^2$, entonces la v.a. $T = \frac{X}{\sqrt{U/r}}$ tiene una distribución t de Student a r grados de libertad (denotada t_r).*

Buscamos la función de densidad de la variable T. Si $f(x,y)$ es la densidad conjunta de (X,Y) y $f_1(x)$ y $f_2(y)$ las densidades marginales de X e Y respectivamente, entonces $f(x,y) = f_1(x)f_2(y)$.

$$f_1(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2}) \quad \forall x \in I\!R$$

$$f_2(y) = \frac{1}{2^{r/2}} \frac{y^{r/2-1}}{\Gamma(r/2)} exp(-y/2) \quad \forall y > 0$$

El jacobiano del cambio de variables $X = T\sqrt{W/r}$ e $Y = W$ es $J = \sqrt{W/r}$. Deducimos la densidad conjunta de (T, W):

$$g(t,w) = \sqrt{\frac{w}{r}} \frac{e^{-\frac{t^2 w}{2r}}}{\sqrt{2\pi}} \frac{w^{\frac{r}{2}-1} e^{-\frac{w}{2}}}{2^{\frac{r}{2}}\Gamma(\frac{r}{2})} \quad \forall w > 0, \ -\infty < t < \infty$$

[2] Student es un seudónimo utilizado por el estadístico inglés W. S. Gosset (1876-1937) para publicar.

$$g(t,w) = \frac{w^{\frac{r-1}{2}} e^{-\frac{1}{2}(1+\frac{t^2}{r})w}}{\sqrt{2^{r+1}\pi r}\Gamma(\frac{r}{2})} \quad \forall w > 0, \ -\infty < t < \infty$$

$$h(t) = \frac{\Gamma(\frac{r+1}{2})(1+\frac{x^2}{r})^{-(\frac{r+1}{2})}}{\sqrt{r\pi}\Gamma(\frac{r}{2})} \quad t \in I\!R$$

Se observa que la función de densidad de T es simétrica, $E(T) = 0$ para $r > 1$ y $var(T) = \frac{r}{r-2}$ para $r > 2$. Además para $r = 1$ se tiene la distribución de Cauchy y para r grande se puede aproximar la distribución de T a una $\mathcal{N}(0,1)$.

Aplicando estos resultados, deducimos que la distribución de la v.a.

$$V = \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/(n-1)}}$$

sigue una distribución t de Student con $n-1$ grados de libertad.

2.4.5 Estadísticos de orden

Hay otros aspectos importantes de una distribución a estudiar, en particular su forma. Por ejemplos, si es simétrica o entre que rango de valores podrían estar los valores muestrales. Para este estudio se consideran otros estadísticos, que son los estadísticos de orden y los cantiles.

Se define los **estadísticos de orden** $X_{(1)}, ..., X_{(n)}$, como los valores muestrales ordenados de menor a mayor: $(X_{(1)} \leq X_{(2)} ... \leq X_{(n)})$. Los estadísticos de orden cambian de una muestra a la otra. Son variables aleatorios. Por ejemplo, sean 3 muestras de tamaño 5 provenientes de la misma población $\mathcal{P} = \{1,2,3,4,5,6,7,8,9\}$:

muestra 1:	{2,	5,	3,	1}
muestra 2:	{8,	5,	,2	7}
muestra 3:	{1,	5,	9,	4}

Entonces $X_{(1)}$ toma los valores 1, 2 y 1 y $X_{(2)}$ toma los valores 2, 5 y 4, etc.

Nos interesamos frecuentemente a $X_{(1)} = min\{X_1, ..., X_n\}$ y $X_{(n)} = max\{X_1, ..., X_n\}$. Estos valores cambian con la muestra.

En el curso de probabilidades y procesos estocásticos se estudiaron las distribuciones de estos estadísticos de orden en función de la distribución de población $F(x)$ de X. En particular, recordamos estos resultados.

- La distribución de $X_{(1)}$ es: $1 - (1 - F(x))^n$
- La distribución de $X_{(n)}$ es: $(F(x))^n$

El rango $W = X_{(n)} - X_{(1)}$ o $(X_{(1)}, X_{(2)})$ son otros estadísticos interesantes a estudiar. Para más detalles pueden consultar H. David|4|.

2.4.6 Cuantiles muestrales

Definición 2.4.8 *Dada una función de distribución $F(x)$ de X, se llama cuantil de orden α al valor x_α tal que $F(x_\alpha) = \alpha$.*

Cuando la distribución F es invertible, $x_\alpha = F^{-1}(\alpha)$.

En el caso empírico, se usa la distribución empírica.

Si tomamos $\alpha = 1/2$, entonces $x_{1/2}$ es tal que hay tantos valores muestrales por debajo que por arriba de $x_{1/2}$. Este valor $x_{1/2}$ se llama **mediana muestral o mediana empírica**. Se llaman **cuartiles** a $x_{1/4}$ y $x_{3/4}$ y **intervalo intercuartiles** a la diferencia $x_{3/4} - x_{1/4}$.

Se observara que para una distribución F_n discreta o empírica, un cuantil para un α dado no es única en general (es un intervalo). Se define entonces como x_α al valor tal que

$$\mathbb{P}(X < x_\alpha) \leq \alpha \leq \mathbb{P}(X \leq x_\alpha)$$

Se llaman quintiles a los valores $x_{k/5}$ para $k = 1, ..., 5$, deciles a los valores $x_{k/10}$ para $k = 1, ..., 10$. Estos valores son generalmente utilizados para estudiar la asimetría de una distribución.

Capítulo 3

ESTIMACIÓN PUNTUAL

3.1 EL PROBLEMA DE LA ESTIMACIÓN

En el estudio de la duración de las ampolletas de $100W$ de la marca ILUMINA (ejemplo 2.1.1), sabemos que la *duración* no es constante: Varía de una ampolleta a otra. Queremos entonces conocer el comportamiento de la variable *duración* que denotaremos X y su función de distribución

$$F(x) = I\!P(X \leq x)$$

Otro problema sería explicar la variabilidad de la duración de las ampolletas y si algunos de los factores tienen incidencia sobre la duración, cómo por ejemplo, la frecuencia con la cual se enciende la ampolleta, la humedad ambiental, etc.

En el experimento que se realiza para estudiar la duración de las ampolletas, el orden con el cual se obtienen los datos de duración sobre una muestra aleatoria simple no tiene importancia. Se puede entonces considerar los datos como realizaciones de variables aleatorias independientes de la misma distribución F desconocida, llamada *función de distribución de la población*, que describe la variabilidad de la duración de las ampolletas.

Se quiere encontrar entonces una función F que coincida mejor con los datos de duración obtenidos sobre una muestra de las ampolletas. Este problema de **modelamiento de los datos muestrales** es el objetivo de la inferencia estadística.

¿Cómo podemos encontrar la función de distribución de población F?

Como lo vimos en el estudio de la función de distribución empírica, esperamos que la distribución de la muestra sea lo más parecida a la distribución de la población. Pero esto nunca la sabremos pues ignoramos si la muestra es realmente "representativa" y no conocemos la distribución de la población. Una manera de proceder consiste en hacer supuestos sobre la función de distribución de la población, lo que constituirá el modelo estadístico.

Vimos en los capítulos anteriores las condiciones que permiten obtener una muestra "representativa" de la población, y vimos también que la media muestral parece ser bastante útil para *estimar*

la media de la población. Pero la duración de vida media es insuficiente para caracterizar completamente la distribución de la variable *duración*. Algunas ampolletas durarán más y otras menos, pero: ¿Cuanto más ó cuanto menos?

En el ejemplo anterior un cierto conocimiento del problema puede sugerir que una distribución *Gamma* de función de densidad:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad si \;\; t > 0$$

es un buen modelo para la duración de vida de las ampolletas.

El problema de la inferencia estadística se reduce entonces en encontrar la función $Gamma(\alpha, \beta)$ que coincide mejor con los datos observados en la muestra. Es decir, se tienen que buscar solamente los parámetros α y β de la función *Gamma* que *ajusten mejor* los valores muestrales. Este es el problema de la estimación puntual que es una de las maneras de inferir a partir de la muestra los parámetros de la población. Veremos varios métodos de estimación puntual.

En el ejemplo 2.1.5 el fabricante efectúa un control de calidad de una muestra aleatoria pequeña con n piezas (generalmente $n \ll N$). Se define la v.a. X con el valor 1 si la pieza es defectuosa y 0 en el caso contrario. Sean $x_1, x_2, ..., x_n$ los valores obtenidos sobre la muestra aleatoria. El modelo estadístico es un proceso de Bernoulli:

$$x_i \sim \mathcal{B}(\theta) \qquad (0 \leq \theta \leq 1)$$

donde el parámetro desconocido θ es la probabilidad de que una pieza sea defectuosa. El fabricante y el cliente quieren saber si θ es mayor que 2%. Se consideran en este caso dos posibilidades para el modelo estadístico: $\mathcal{B}(\theta)$ con $\theta \leq 2\%$ y $\mathcal{B}(\theta)$ con $\theta > 2\%$.

Según el conocimiento que se tiene de F o los supuestos sobre F, se tiene distintos métodos de inferencia estadística.

- Si se sabe que F pertenece a una familia de funciones $\mathcal{F}(\theta)$ que dependen de un parámetro ó un vector de parámetros θ, el problema consiste en *estimar* solamente el parámetro desconocido θ. Cuando se define un valor para θ a partir de los valores muestrales, se habla de **estimación puntual**. Otra forma de estimar un parámetro consiste en buscar no sólo un valor para θ, sino un intervalo, en el cual se tenga una alta probabilidad de encontrar al parámetro θ. Se habla del método de **estimación por intervalo** que permite asociar a la estimación puntual una precisión.

- Si no se supone que F pertenece a una familia conocida de funciones de distribución, pero se hace supuestos más generales sobre la forma de la función de distribución, se habla de una **estimación no parámetrica**.

- Si queremos verificar que el conjunto de valores muestrales proviene de una función de distribución F de parámetro θ con una condición sobre θ, se usa la teoría de **test de hipótesis paramétrica** para verificar si se cumple la condición sobre θ.

- Si queremos verificar que el conjunto de valores muestrales proviene de una familia de funciones de distribución dada, se usa la teoría de **test de hipótesis no parámetrica**.

En cada uno de los casos anteriores se define un **modelo estadístico** que se toma como base para la inferencia estadística.

En el caso del problema de estimación, el modelo es una familia de funciones de distribución y se estiman entonces los parámetros desconocidos del modelo. En el caso del test de hipótesis, se plantean dos o más modelos estadísticos alternativos y se busca cual es el más adecuado de acuerdo con los datos observados.

3.2 ESTIMACIÓN DE PARÁMETROS

En el problema de estimación puntual el modelo estadístico esta definido por una familia de distribuciones de donde se supone provienen los valores muestrales y el modelo tiene solamente algunos elementos desconocidos que son los **parámetros del modelo**. Se trata entonces de encontrar los parámetros desconocidos del modelo utilizando los valores muestrales. La elección de la familia de distribuciones se hace a partir de consideraciones teóricas ó de la distribución de frecuencias empíricas.

El el ejemplo 2.1.1 de las ampolletas, hicimos el supuesto que $F(x)$ pertenece a la familia de las distribuciones $Gamma(\alpha, \beta)$, en los ejemplos 2.1.2 de la talla de las chilenas y 2.1.3 de las bolsas de azúcar, la distribución normal $\mathcal{N}(\mu, \sigma^2)$ y los ejemplos 2.1.4 del candidato a la eleción y 2.1.5 de las piezas defectuosas, un modelo de Bernoulli $\mathcal{B}(p)$.

Los parámetros α, β, μ, σ^2 ó p son constantes desconocidas.

Definición 3.2.1 *Un modelo estadístico parámetrico es una familia de distribuciones de probabilidad indiciado por un parámetro θ (que puede ser un vector). El conjunto de los valores posibles de θ es el espacio de parámetro Ω. Denotaremos $F_\theta(x)$ a la función de distribución (acumulada).*

Por ejemplos:

$\mathcal{N}(\mu, 1)$	$\Omega = I\!R$
$\mathcal{N}(\mu, \sigma)$	$\Omega = I\!R \times]0, +\infty[$
$Exp(\beta)$	$\Omega =]0, +\infty[$
$\mathcal{B}(p)$	$\Omega = [0, 1]$
$Poisson(\lambda$	$\Omega =]0, +\infty[$
$Uniforme(]\theta_1, \theta_2[)$	$\Omega = I\!R \times I\!R$ (sujeto a $\theta_1 < \theta_2$)

En el ejemplo 2.1.4 el candidato encarga un estudio de opinión a un estadístico, que toma una muestra aleatoria pequeña de n votantes. Se define la v.a. X que toma el valor 1 si la persona i interrogada declara que su intención de voto es para el candidato y 0 en el caso contrario. Sean $x_1, x_2, ..., x_n$ los valores obtenidos sobre la muestra aleatoria. El modelo estadístico es entonces el siguiente:

$$x_i \sim Bernoulli(\theta) \quad (0 \leq \theta \leq 1)$$

donde el parámetro desconocido es la probabilidad θ que un elector vote por el candidato.

En el ejemplo 2.1.2, si $X_1, X_2, ..., X_N$ son las tallas de todas las chilenas mayores de 15 años, la media de la población es igual a $\mu = \sum X_i/N$. Dado el gran tamaño grande de esta población, se obtiene la talla de una muestra aleatoria de tamaño pequeño n. Sean $x_1, x_2, ..., x_n$. Si suponemos que la distribución de población de X es normal, el modelo es

$$x_i \sim \mathcal{N}(\mu, \sigma^2) \quad (\mu \in I\!R), \ (\sigma^2 \in I\!R^+)$$

donde μ y σ^2 son ambos desconocidos.

Distinguiremos el caso de función de distribución continua y discreta.

Definición 3.2.2 *Sea la variable $X : \mathcal{P} \longrightarrow \mathcal{Q}$*
a) Un modelo estadístico parámetrico es continuo si para todo $\theta \in \Omega$ la función de distribución $F_\theta(x)$ es continua con función de densidad que denotaremos $f_\theta(x)$.
b)Un modelo estadístico paramétrico es discreto si para todo $\theta \in \Omega$ la función de distribución $F_\theta(x)$ es discreta con función de probabilidad (masa) que denotaremos $p_\theta(x)$.

La función de distribución de la talla de las mujeres chilenas o de la duración de vida de la ampolleta es continua y la distribución de las maquinas defectuosas es discreta.

Sean $X_1, ..., X_n$ los valores muestrales obtenidos sobre una muestra aleatoria simple de una v.a. X de función de densidad $f_\theta(x)$ (o probabilidad $p_\theta(x)$), en que θ es desconocido. Se busca elegir entonces un valor para θ a partir de los valores muestrales, es decir una función $\delta : \mathcal{Q}^n \longrightarrow \Omega$, que es un estadístico (una función de los valores muestrales) llamado **estimador** de θ. El valor tomado por esta función sobre una muestra particular de tamaño n es una **estimación**.

Procediendo así, tratamos de **estimar el valor del parámetro**, que es una constante, a partir de un estadístico que es aleatorio.

El problema es que no hay una regla única que permita construir estos estimadores. Por ejemplo, en una distribución de población simétrica la media y la mediana empíricas son ambas estimaciones posibles para la esperanza. Para elegir entonces entre varios estimadores de un mismo parámetro hay que definir criterios de comparación. Presentemos a continuación algunas propiedades razonables para decidir si un estimador es aceptable.

Cabe destacar que las propiedades de consistencia, eficiencia y suficiencia para un buen estimador fueron introducida por R. A. Fisher (parráfo 1.1).

3.3 PROPIEDADES DE LOS ESTIMADORES

Un buen estimador $\hat{\theta}$ para θ sera aquel que tiene un error de estimación $|\hat{\theta} - \theta|$ lo más pequeño posible. Pero como esta diferencia es aleatoria, hay diferentes maneras de verla. Por ejemplos:

- $|\hat{\theta} - \theta|$ es pequeña con alta probabilidad.
- $|\hat{\theta} - \theta|$ es nulo en promedio.
- $|\hat{\theta} - \theta|$ tiene una varianza pequeña.

3.3.1 Estimadores consistentes

Un estimador depende del tamaño de la muestra a través de los valores muestrales; los estimadores $\hat{\theta}_n$ asociados a muestras de tamaño n ($n \in \mathbb{N}$) constituyen sucesiones de variables aleatorias. Un buen estimador debería converger en algún sentido hacia θ cuando el tamaño de la muestra crece. Tenemos que usar las nociones de convergencia de variables aleatorias.

Definición 3.3.1 *Se dice que un estimador $\hat{\theta}_n$ de un parámetro θ es* **consistente** *cuando converge en probabilidad hacia θ: Dado $\varepsilon > 0$ y $\eta > 0$ pequeños, $\exists\, n_{\varepsilon,\eta}$, dependiente de ε y η tal que*

$$I\!P(|\hat{\theta}_n - \theta| \leq \epsilon) > 1 - \eta \quad \forall n \geq n_{\varepsilon,\eta}$$

Se escribe $\hat{\theta}_n \xrightarrow{prob} \theta$.

Los momentos empíricos de una v.a. real son estimadores consistentes de los momentos teóricos correspondientes. Más aún la convergencia es casi-segura y la distribución asintótica de estos estimadores es normal.

3.3.2 Estimadores insesgados

Definición 3.3.2 *Se dice que un estimador $\hat{\theta}$ de θ es* **insesgado** *si y solo si $E(\hat{\theta}) = \theta$.*

Es decir que los errores de estimación tienen un promedio nulo.

Vimos que la media muestral $\bar{X}_n$ es un estimador insesgado de la media poblacional si la muestra es aleatoria simple, pero la varianza muestral $S_n^2 = \frac{1}{n}\sum(x_i - x_n)^2$ no es un estimador insesgado para la varianza poblacional σ^2:

$$E(S_n^2) = \frac{n-1}{n}\sigma^2$$

Sin embargo, la diferencia $|E(S_n^2) - \sigma^2| = \sigma^2/n$, que es **el sesgo**, tiende a cero.

Definición 3.3.3 *Se dice que el estimador $\hat{\theta}$ es* **asintóticamente insesgado** *cuando $E(\hat{\theta}) \xrightarrow{n\to\infty} \theta$.*

Se puede construir un estimador insesgado a partir de S_n^2: $\tilde{\sigma}^2 = \frac{1}{n-1}\sum(x_i - X_n)^2$. Observemos que $\tilde{\sigma}^2 = (\frac{n}{n-1})^2\sigma^2$, es decir que el estimador insesgado $\tilde{\sigma}^2$ tiene mayor varianza que S_n^2.

En efecto si $\hat{\theta}_n^2$ es un estimador sesgado de θ, eso no implica nada sobre su varianza.

Consideramos entonces la varianza del error de estimación llamado **error cuadrático medio**:

$$E(\hat{\theta}_n - \theta)^2 = Var(\hat{\theta}_n) + (sesgo)^2$$

En efecto,

$$E(\hat{\theta}_n - \theta)^2 = E[(\hat{\theta}_n - E(\hat{\theta}_n) + E(\hat{\theta}_n) - \theta)^2]$$
$$E(\hat{\theta}_n - \theta)^2 = E[(\hat{\theta}_n - E(\hat{\theta}_n))^2] + [E(\hat{\theta}_n) - \theta)]^2$$

Si $[E(\hat{\theta}_n) - \theta)]^2 \longrightarrow 0$ entonces $\hat{\theta}_n$ converge en media cuadrática hacia θ
($\hat{\theta}_n \xrightarrow{m.c.} \theta$).

Proposición 3.3.4

$$[E(\hat{\theta}_n - \theta)^2 \longrightarrow 0] \Longleftrightarrow [Var(\hat{\theta}_n) \to 0 \quad y \quad E(\hat{\theta}_n) \to \theta]$$

Como la convergencia en media cuadrática implica la convergencia en probabilidad se tienen los dos resultados siguientes:

Proposición 3.3.5 *Si $\hat{\theta}_n$ es un estimador consistente de θ y $E(\hat{\theta}_n)$ es finito, entonces $\hat{\theta}_n$ es asintóticamente insesgado.*

Proposición 3.3.6 *Si θ_n es un estimador de θ tal que $Var(\hat{\theta}_n) \to 0 \quad y \quad E(\hat{\theta}_n) \to \theta$, entonces $\hat{\theta}_n$ es un estimador consistente de θ.*

Nota 3.3.7 *En la última proposición la condición es suficiente pero no necesaria.*

Ejercicio: Compare los errores cuadráticos medio de $S_n^2 = \frac{1}{n}\sum(x_i - x_n)^2$ y $\tilde{\sigma}^2 = \frac{1}{n-1}\sum(x_i - \bar{X}_n)^2$. Se muestra en la figura 3.1 la variación del error cuadrático medio en función de la tamaño del la muestra para los dos estimadores cuando $\sigma^2 = 1$.

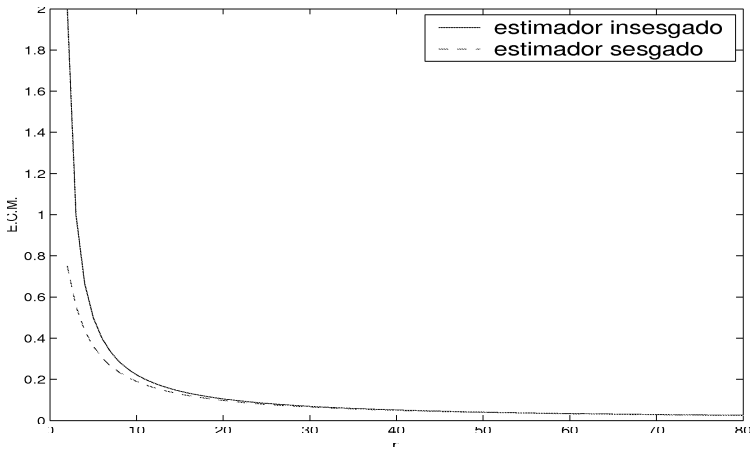


Figura 3.1: Error cuadrático medio en función de n

En resumen, un estimador puede ser insesgado pero con una varianza elevada y entonces poco preciso. Otro estimador puede tener un sesgo y una varianza pequeños, lo que produce un error cuadrático medio pequeño (ver figura 3.2 donde el centro del blanco representa el parámetro a estimar y los otros puntos son diferentes estimaciones obtenidos de un estimador).

Otra manera de ilustrar el problema entre sesgo y precisión esta dada en las figuras 3.3 cuando se supone que el estimador se distribuye como una distribución normal. Cuando la distribución del estimador esta centrada en el parámetro buscado, el estimador es insesgado; cuando la distribución esta poco dispersa, el estimador es preciso.

En la figura izquierda, ambos estimadores son insesgados, entonces se prefiere el estimador representado por la línea continua. En la figura derecha, se prefiere el estimador representado por la línea continua también: aún si es sesgado, es mejor que el otro que es insesgado: globalmente sus valores son más cercanos al valor a estimar.

3.3.3 Estimador eficiente

Vimos que si $x_1, ...x_n$ son valores muestrales i.i.d de una población $\mathcal{N}(\mu, \sigma^2)$, la media muestral $\bar{x}$ es un estimador insesgado de μ y que su varianza es igual a $\frac{\sigma^2}{n}$. Nos preguntamos entonces si existen

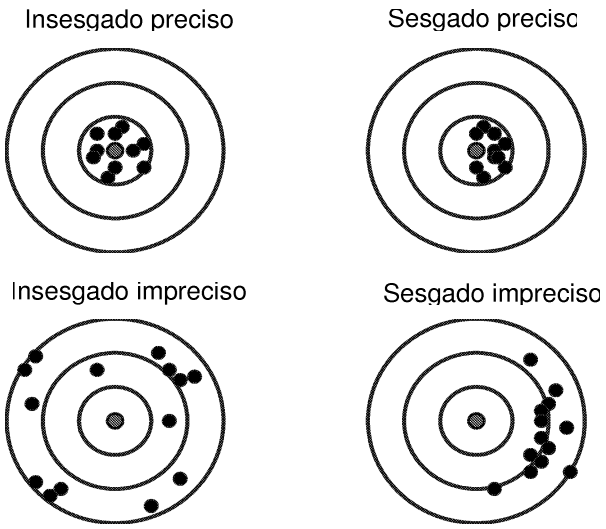


Figura 3.2: Sesgo y varianza

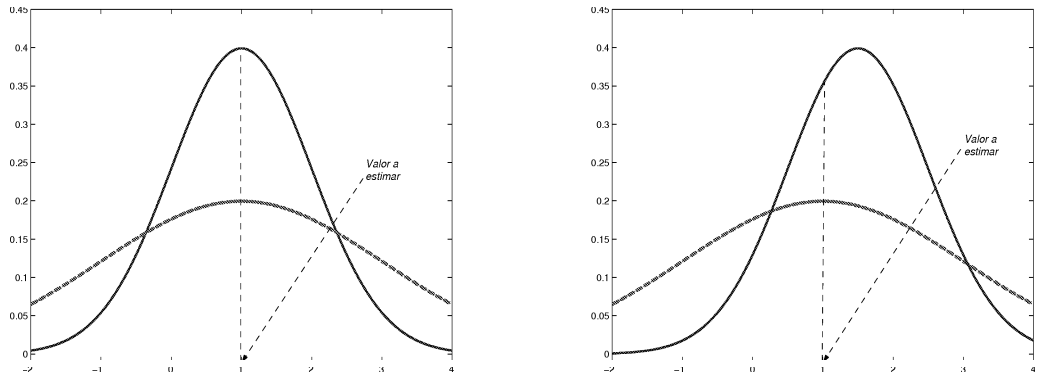


Figura 3.3: Sesgo y varianza

otros estimadores insesgado de μ de menor varianza. Es decir queremos encontrar entre todos los estimadores insesgados el que tenga la menor varianza. Esto no es siempre fácil.

Aquí vamos a dar, bajo ciertas condiciones, una manera que permite verificar si un estimador insesgado dado tiene la varianza más pequeña. Tal propiedad se llama **eficiencia** del estimador.

Vamos a establecer una desigualdad (CRAMER-RAO), que nos permite dar una cota inferior a la varianza de un estimador insesgado. Esta cota se basa en la cantidad de información de Fisher.

Definición 3.3.8 *Se llama cantidad de información de Fisher dada por X sobre el parámetro θ a la cantidad*

$$I(\theta) = E[(\frac{\partial \ln(f)}{\partial \theta})^2]$$

Se puede dar dos otras formas a la cantidad de Información de Fisher:

Teorema 3.3.9

$$I(\theta) = Var(\frac{\partial \ln(f)}{\partial \theta})$$

Demostración Sea S el dominio de X y f_θ la función de densidad de la variable X, entonces como $\int_S f_\theta(x)dx = 1, \forall\theta \in \Omega$, se tiene $\int_S f'_\theta(x)dx = 0, \forall\theta \in \Omega$. Además $\frac{\partial \ln f_\theta}{\partial\theta} = \frac{f'_\theta}{f}$, luego $E(\frac{\partial \ln f_\theta}{\partial\theta}) = 0, \forall\theta \in \Omega$ y $I(\theta) = Var(\frac{\partial \ln f_\theta}{\partial\theta})$. ■

El teorema siguiente nos da otra expresión para $I(\theta)$ que a menudo es más fácil de calcular.

Teorema 3.3.10 *Si el dominio s de X no depende de θ, entonces*

$$I(\theta) = -E[(\frac{\partial^2 \ln f_\theta}{\partial\theta^2})]$$

si esta cantidad existe.

Demostración Si $\frac{\partial^2 \ln f_\theta}{\partial\theta^2}$ existe $\forall\theta$, como $E(\frac{\partial \ln f_\theta}{\partial\theta}) = 0$ y $\frac{\partial^2 \ln f_\theta}{\partial\theta^2} = \frac{f_\theta f''_\theta - (f'_\theta)^2}{f_\theta^2} = \frac{f''_\theta}{f_\theta} - (\frac{\partial \ln f_\theta}{\partial\theta^2})$. Luego $\frac{\partial^2 \ln f_\theta}{\partial\theta^2} = \int_S f''_\theta(x)dx - I(\theta)$, y se deduce que $I(\theta) = -E[(\frac{\partial^2 \ln f_\theta}{\partial\theta^2})]$. ■

Sea una m.a.s. $x_1, x_2, ..., x_n$, de función de densidad o función de probabilidad $f_\theta(x)$ en donde θ es un parámetro desconocido del conjunto Ω. Sea L_θ la función de verosimilitud de la muestra.

Definición 3.3.11 *Se llama cantidad de información de Fisher dada por una muestra aleatoria $x_1, x_2, ..., x_n$ sobre el parámetro θ a la cantidad*

$$I_n(\theta) = E[(\frac{\partial \ln(L_\theta)}{\partial\theta})^2]$$

Nuevamente se tienen las dos otras formas de expresar $I_n(\theta)$:

$$I_n(\theta) = Var[(\frac{\partial \ln(L_\theta)}{\partial\theta})] \qquad I_n(\theta) = -E[(\frac{\partial^2 \ln(L_\theta)}{\partial\theta^2})]$$

Teorema 3.3.12 *Si los valores muestrales son independientes y $I(\theta)$ es la cantidad de información de Fisher dada para cada x_i sobre el parámetro θ, entonces*

$$I_n(\theta) = nI(\theta)$$

Si $x_1, x_2, ..., x_n$son los valores muestrales obtenidos de una variable X de función de densidad o función de probabilidad $f_\theta(x)$, se tiene la desigualdad de CRAMER-RAO:

Teorema 3.3.13 *Si el dominio S de X no depende del parámetro θ, para todo estimador T insesgado de θ se tiene:*

$$Var(T) \geq \frac{1}{I_n(\theta)}$$

Además si T es un estimador insesgado de $h(\theta)$ una función de θ, entonces $Var(T) \geq \frac{(h'(\theta))^2}{I_n(\theta)}$

Demostración Como $E(\frac{\partial ln(L_\theta)}{\partial \theta}) = 0$,

$$Cov(T, \frac{\partial ln(L_\theta)}{\partial \theta}) = E(T\frac{\partial ln(L_\theta)}{\partial \theta}) = \int t\frac{\partial ln(L_\theta)}{\partial \theta}L_\theta dx = \int t\frac{\partial L_\theta}{\partial \theta}dx$$

$$Cov(T, \frac{\partial \ln(L_\theta)}{\partial \theta}) = \frac{\partial}{\partial \theta}\int tL_\theta dx = \frac{\partial}{\partial \theta}E(T) = h'(\theta))$$

De la desigualdad de Schwarz, se obtiene

$$(Cov(T, \frac{\partial ln(L_\theta)}{\partial \theta}))^2 \leq Var(T)Var(\frac{\partial ln(L_\theta)}{\partial \theta})$$

Es decir

$$(h'(\theta))^2 \leq Var(T)I_n(\theta)$$

■

Nota 3.3.14 *La desigualdad de Cramer-Rao puede extenderse al error cuadrático medio de los estimadores sesgados: Si el dominio S de X no depende del parámetro θ y $b(\theta) = E(T) - \theta$ es el sesgo de T, para todo estimador T de θ se tiene:*

$$E[(T-\theta)^2] \geq \frac{(1 + \frac{\partial b(\theta}{\partial \theta})^2}{I_n(\theta)}$$

Sea $X \sim \mathcal{N}(\mu, \sigma^2)$ con σ^2 varianza conocida. Como $I_n(\mu) = \frac{n}{\sigma^2}$, todo estimador T insesgado de μ tiene una varianza al menos igual a $\frac{\sigma^2}{n}$. Por tanto se deduce que la media $\bar{x}$ es eficiente.

Si ahora se supone que σ^2 es desconocida la cota de CRAMER-RAO nos indica que todo estimador insesgado de σ^2 tendra una varianza al menos igual a $\frac{2\sigma^2}{n}$. El estimador $\frac{1}{n-1}\sum(x_i - \bar{x})^2$, que es insesgado para σ^2, tiene una varianza igual $\frac{2\sigma^2}{n-1}$, que es mayor que la cota. Sin embargo este estimador es función de un estadístico insesgado suficiente por lo tanto es eficiente (ver el parráfo siguiente). Lo que no muestra que la cota de Cramer-Rao no sea precisa en el caso de la varianza.

3.3.4 Estimador suficiente

Generalmente los valores muestrales proporcionan alguna información sobre el parámetro θ. Pero tomar todos los valores muestrales separadamente puede dar informaciones redondantes. Es la razón por la cual se resumen los valores muestrales en un estadístico (como la media muestral o la varianza muestral). Pero en este resumen no debemos perder información en lo que concierne al parámetro θ. El concepto de *estadístico suficiente* proporciona una buena regla para obtener estimadores que cumplan este objetivo, eliminando de los valores muestrales la parte que no aporta nada al conocimiento del parámetro θ y resumiendo la información contenida en los valores muestrales en un solo estadístico que sea relevante para θ.

En el ejemplo 2.1.5, se busca deducir de las observaciones de una muestra aleatoria de n piezas de una máquina una información sobre la proporción θ de piezas defectuosas en el lote total. Es más simple considerar el número de piezas defectuosas encontradas en la muestra en vez de la sucesión de resultados $x_1, x_2, ..., x_n$. El conocimiento de los valores individuales no procura

ninguna información aditiva para la proporción θ que $\sum_{i=1}^{n} x_i$. En el ejemplo 2.1.4, el conocimiento del voto de cada encuestado no aporta más información para determinar la proporción de votos del candidato en la elección que la cantidad de votos recibidos por el candidato en la muestra. En estos dos ejemplos se reducen los n datos a un sólo valor, que es función de estos datos (la suma de los valores muestrales), sin perder información para determinar el parámetro θ de la Bernoulli.

Supongamos el caso $n = 2$ y el estadístico $T = X_1 + X_2$, con $X_i \sim \mathcal{B}(\theta)$. Buscamos la distribución condicional de $X = (X_1, X_2)$ dado T. El estadístico T toma 3 valores:

$$T = \begin{cases} 0 & si \quad X = (0,0) \quad con \quad probabilidad \quad 1 \\ 1 & si \quad X = (0,1) \quad o \quad X = (1,0) \quad con \quad probabilidad \quad 1/2 \\ 2 & si \quad X = (1,1) \quad con \quad probabilidad \quad 1 \end{cases}$$

La distribución condicional de $X = (X_1, X_2)$ dado T no depende de θ y la distribución de $X^* = (X_1^*, X_2^*)$ obtenida de la distribución condicional de X dado T es igual a la distribución de $X = (X_1, X_2)$. En consecuencia, si $d(X) = d(X_1.X_2)$ es un estimador de θ, $d(X^*)$ da una regla al menos igual de buena que $d(X)$. Lo que significa que basta buscar un estimador basado solamente en $T = X_1 + X_2$. Se dice que $T = X_1 + X_2$ es un estadístico suficiente para θ.

En los ejemplos 2.1.2 y 2.1.3, la media muestral $\bar{X}_n$ permite simplificar la información dada por los n valores muestrales. Pero nos preguntamos si se pierde información usando la media muestral para estimar la media μ de la población.

Observemos que si suponemos la varianza conocida e igual a 1, la función de densidad conjunta, llamada también **función de verosimilitud** puede escribirse como función únicamente de la media muestral y del tamaño n de la muestra:

$$f_\theta(x_1, x_1, ..., x_n) = (\frac{1}{2\pi})^{\frac{n}{2}} exp(-\frac{n}{2}(\bar{x}_n - \theta)^2)$$

Es decir que la única información relevante para estimar θ está dada por la media muestral. En este caso se dice que la media muestral es un estadístico suficiente. Un estadístico suficiente, que se toma como estimador del parámetro θ, debería contener toda la información que llevan los valores muestrales sobre θ.

Definición 3.3.15 *Un estadístico $T(x_1, ..., x_n)$, función de los valores muestrales y con valor en Ω, se dice* **suficiente** *para θ si la distribución conjunta de los valores muestrales condicionalmente a $T(x_1, ..., x_n)$ no depende de θ.*

Un estadístico suficiente para un parámetro θ no es necesariamente único. Buscaremos un estadístico que sea una mejor reducción de los datos.

Definición 3.3.16 *Se dice que un estadístico T es suficiente minimal si la distribución condicional de cualquier otro estadístico suficiente dado T no depende de θ.*

No es siempre fácil detectar si un estadístico es suficiente y menos encontrar un estadístico suficiente minimal. Los dos siguientes teoremas permiten enunciar condiciones para que un estadístico sea suficiente.

Teorema 3.3.17 *Principio de factorización*
Si $T(x_1, x_1, ..., x_n)$ es suficiente para θ y $g(T(x_1, x_1, ..., x_n); \theta)$ es la densidad de $T(x_1, x_1, ..., x_n)$, entonces

$$f_\theta(x_1, x_1, ..., x_n) = g(T(x_1, x_1, ..., x_n; \theta)h(x_1, x_1, ..., x_n|T(x_1, x_1, ..., x_n))$$

El principio de factorización nos permite reconocer si un estadístico es suficiente, pero no permite construir uno ó saber si existe uno. El siguiente teorema permite buscar estadísticos suficientes para una clase de distribuciones llamadas exponenciales.

Teorema 3.3.18 *Theorema de Darmois-Koopman*
Si X es una variable real cuyo dominio de variación no depende del parámetro θ, una condición necesaria y suficiente para que exista un estadístico suficiente es que la función de densidad de X sea de la forma:

$$f(x; \theta) = b(x)c(\theta)exp\{a(x)q(\theta)\}$$

Además $T_n(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} a(X_i)$ es un estadístico suficiente minimal.

Si $X \sim \mathcal{N}(\theta, 1)$ y si $x_1, ..., x_n$ es una muestra aleatoria de X

$$f_n(x_1, ..., x_n; \theta) = \frac{1}{(2\pi)^{n/2}} exp(-\frac{1}{2}\sum x_i^2) exp(-\frac{n\theta^2}{2} + n\theta\bar{X}_n)$$

El término $exp(-\frac{1}{2}\sum x_i^2)$ no depende de θ y el término $exp(-\frac{n\theta^2}{2} + n\theta\bar{X}_n)$ depende de θ y $\bar{X}_n$.
$n\bar{X}_n = \sum x_i$ es un estadístico suficiente y toda función biyectiva de $\bar{X}_n$ lo es también, en particular $\bar{X}_n$.

Un último resultado importante, que permite construir estimadores insesgados mejores es.

Teorema 3.3.19 *Theorema de Rao-Blackwell*
Si T(X) es un estadístico suficiente para θ y si b(X) es un estimador insesgado de θ, entonces

$$\delta(T) = E(b(X)|T)$$

es un estimador insesgado de θ basado sobre T mejor que b(X).

No es fácil encontrar buenos estimadores insesgado, de varianza minimal; de hecho estas dos propiedades pueden ser antagónicas en el sentido que al buscar eliminar el sesgo se aumenta la varianza. Por otro lado la búsqueda de estimadores insesgados de mínima varianza esta relacionada con la existencia de estadísticos suficientes.

A continuación daremos los métodos usuales de estimación puntual.

3.4 MÉTODO DE LOS MOMENTOS

Vimos en el capítulo anterior que la media muestral $\bar{X}_n \xrightarrow{c.s.} E(X) = \mu$. Más generalmente si el momento $\mu_r = E(X^r)$ existe, entonces por la ley de los grandes números:

$$m_r = \frac{1}{n}\sum X_i^r \xrightarrow{c.s.} \mu_r \qquad (I\!P(\lim_{n\to\infty} m_r = \mu_r) = 1)$$

Luego una método de estimación consiste en hacer coincidir el momento μ_r de orden r del modelo estadístico con el momento empírico m_r obtenido de la muestra.

Ejemplos:

- Caso de la normal $\mathcal{N}(\mu, \sigma^2)$: El método de los momentos produce como estimador de la media μ, $\hat{\mu} = \bar{x}_n$ y como estimador de la varianza $\sigma^2 = m_2 - \bar{x}_n^2 = s_n^2$.
- Caso de una Bernoulli $\mathcal{B}(\theta)$: Como $E(X) = \theta$, el estimador de los momentos de θ es $\bar{x}_n$.
- Caso de una $Poisson(\lambda)$: Como $E(X) = \lambda$, el estimador de los momentos de λ es $\bar{x}_n$.
- Caso de una uniforme en $[0, \theta]$: Como $E(X) = \frac{\theta}{2}$, el estimador de los momentos es $\hat{\theta} = 2\bar{x}_n$. Un inconveniente de este estimador es que algunos valores muestrales podrían ser mayor que $\hat{\theta}$.

La ventaja del método es que es intuitivo y, en general, basta calcular el primer y segundo momento. Pero tiene que existir estos momentos y no ofrece tanta garantía de buenas propiedades como el estimador de máxima verosimilitud.

3.5 MÉTODO DE MÁXIMA VEROSIMILITUD

Sean $x_1, x_2, ..., x_n$ una muestra aleatoria simple de una v.a. de densidad $f_\theta(x)$ en que $\theta \in \Omega$, el espacio de parámetros.

Definición 3.5.1 *Se llama* **función de verosimilitud** *a la densidad conjunta (ó función de probabilidad) del vector de los valores muestrales; para todo vector observado* $\underline{x} = (x_1, x_2, ..., x_n)$ *en la muestra, se denota* $f_\theta(x_1, x_2, ..., x_n) = f_\theta(\underline{x})$.

Cuando los valores muestrales son independientes, se tiene:

$$f_\theta(\underline{x}) = f_\theta(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

El estimador de máxima verosimilitud es un estadístico $T(x_1, ..., x_n)$ función de los valores muestrales que maximiza la función f_θ.

Tal estimador puede entonces no ser único, o bien no existir.

Cuando este estimador existe, tiene algunas propiedades interesantes que se cumplen bajo condiciones bastante generales:

- Es consistente.
- Es asintóticamente normal;
- No es necesariamente insesgado, pero es generalmente asintóticamente insesgado;
- Es función de un estadístico suficiente, cuando existe uno;

- Entre todos los estimadores asintóticamente normales, tiene la varianza asintóticamente más pequeña (es eficiente).
- Tiene la propiedad de **invarianza**.

Proposición 3.5.2 *(Propiedad de Invarianza) Si $\hat{\theta}$ es el estimador de máxima verosimilitud del parámetro θ y si $g : \Omega \longrightarrow \Omega$ es biyectiva, entonces $g(\hat{\theta})$ es el estimador de máxima verosimilitud de $g(\theta)$.*

Demostración En efecto si $\tau = g(\theta)$, como g es biyectiva, $\theta = g^{-1}(\tau)$; si $f_\theta(\underline{x}) = f_{g^{-1}(\tau)}(\underline{x})$ es máxima para $\hat{\tau}$ tal que $g^{-1}(\hat{\tau}) = \hat{\theta}$. $\hat{\tau}$ es necesariamente el estimador de máxima verosimilitud y como g es biyectiva, $\hat{\tau} = g(\hat{\theta})$. ■

Veremos a continuación, que el estimador de máxima verosimilitud de σ se puede obtener directamente ó como la raíz del estimador de máxima verosimilitud de σ^2. Eso se debe a la propiedad de **invarianza** del estimador de máxima verosimilitud transformación funcional. Veamos algunos ejemplos.

Sean en el ejemplo 2.1.5, $x_1, x_2, ..., x_n$ los valores muestrales.

$$x_i \sim Bernoulli(\theta) \quad (0 \leq \theta \leq 1)$$

$$f_\theta(\underline{x}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$\max_{\theta\in[0,1]} f_\theta(\underline{x}) \Longleftrightarrow \max_{\theta\in[0,1]} Log f_\theta(\underline{x})$$

$$Log f_\theta(\underline{x}) = \sum_{i=1}^{n}[x_i Log\theta + (1-x_i)Log(1-\theta)]$$

$$\frac{dLog f_\theta(\underline{x})}{d\theta} = \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1-\theta} = 0$$

Luego el estimador de máxima verosimilitud $\hat{\theta}$ de θ es la proporción de piezas defectuosas observada $\sum x_i/n$.

$\hat{\theta} = \sum x_i/n$ ($\hat{\theta} \in |0,1|$) es un estimador del parámetro θ insesgado, consistente y suficiente.

Sean $x_1, x_2, ..., x_n$ las tallas obtenidas sobre la muestra de mujeres chilenas mayores de 15 años en el ejemplo 2.1.2.

Se supone que $x_i \sim \mathcal{N}(\mu, \sigma^2)$ con μ y σ^2 desconocidos.

$$f_\theta(\underline{x}) = (\frac{1}{2\pi\sigma^2})^{\frac{n}{2}} exp\{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\}$$

$Log f_\theta(\underline{x})$ es máximo cuando μ es igual a la media muestral $\bar{x}_n$ y σ^2 es igual a la varianza muestral S_n^2.

El estimador $(\bar{x}_n, S_n^2)$ es suficiente para (μ, σ^2). El estimador $\bar{x}_n$ de la media poblacional μ es insesgado, consistente y de mínima varianza. El estimador S_n^2 de la varianza de la población es asintóticamente insesgado y consistente.

Nota 3.5.3 *Si se supone la varianza poblacional σ^2 conocida, el estimador de máxima verosimilitud de μ queda igual a la media muestral $\bar{x}_n$. Además Se puede buscar el estimador de la varianza o bien de su raíz σ. El resultado no cambia.*

Sea $x_i \sim Uniforme|0,\theta| \qquad \theta > 0,\ f_\theta(\underline{x}) = 1/\theta^n \quad si \quad 0 \leq x_i \leq \theta \quad \forall i.$

Cuando $\theta \geq x_i$ para todo i, $f_\theta(\underline{x})$ es no nulo y es decreciente en θ; luego $f_\theta(\underline{x})$ es máxima para el valor más pequeño de θ que hace $f_\theta(\underline{x})$ no nulo: el estimador de máxima verosimilitud de θ es entonces $\hat{\theta} = max\{x_1, x_2..., x_n\}$.

El método de los momentos produce un estimador bien diferente. En efecto, como $E(X) = \theta/2$, el estimador de los momentos es $\tilde{\theta} = 2\bar{x}_n$.

En este ejemplo, una dificultad se presenta cuando se toma el intervalo $|0,\theta|$ abierto, dado que no se puede tomar como estimador el máximo $\hat{\theta}$; en este caso el estimador de máxima verosimilitud no existe. Puede ocurrir que no es único también. Si se define el intervalo $|\theta, \theta+1|$, es decir el largo del intervalo es conocido e igual a 1, la función de verosimilitud es:

$$f_\theta(\underline{x}) = 1 \quad si \quad \theta \leq x_i \leq \theta + 1 \quad \forall i$$

es decir: $f_\theta(\underline{x}) = 1 \quad si \quad max\{x_1, ..., x_n\} - 1 \leq \theta \leq min\{x_1, ..., x_n\}$. Por lo cual todo elemento del intervalo $|max\{x_1, ..., x_n\} - 1, min\{x_1, ..., x_n\}|$ maximiza la verosimilitud.

Aquí el estimador de los momentos, que es igual a $\bar{x}_n - 1/2$, es bien diferente también.

Se deja como ejercicio estudiar las propiedades de estos estimadores.

3.6 EJERCICIOS

1. Sea X_i, $i = 1, ..., n$ una muestra aleatoria simple de una v.a. X de función de distribución Gamma(α, β). Estime $\mu = E(X)$ por el método de máxima verosimilitud. Muestre que el estimador resultante es insesgado, convergente en media cuadrática y consistente.

2. Sea una m.a.s. $x_1, ...x_n$ de una v.a. X de función de densidad $f_\theta(x) = \theta x^{\theta-1} \boldsymbol{I}_{[0,1]}$. Encuentre el estimador de máxima verosimilitud $\hat{\theta}$ de θ y pruebe que $\hat{\theta}$ es consistente y asintóticamente insesgado.

3. Sean dos preguntas complementarias: A="vota por Pedro" y A'="no vota por Pedro". Se obtiene una muestra aleatoria simple de n personas que contestan a la pregunta A ó A'; lo único que se sabe es que cada persona ha contestado A con probabilidad θ conocida y A' con probabilidad $(1-\theta)$. Se definen:
p: la probabilidad que una persona contesta "SI" a la pregunta (A ó A')
π: la proporción desconocida de votos para Pedro en la población.

a) Dé la proporción π en función de p y θ.
b) Dé el estimador de máxima verosimilitud de p y deduzca un estimador $\hat{\pi}$ para π. Calcule la esperanza y la varianza de $\hat{\pi}$.
c) Estudie las propiedades de $\hat{\pi}$; estudie en particular la varianza $\hat{\pi}$ cuando $\theta = 0.5$.

4. Se considera la distribución discreta: $I\!P(X = x) = a_x \theta^x / h(\theta)$, con $x = 0, 1, 2, ...$, en donde h es diferenciable y a_x puede ser nulo para algunos x.
Sea $\{x_1, x_2, ..., x_n\}$ una m.a.s. de esta distribución.

a) Dé las expresiones de $h(\theta)$ y $h'(\theta)$.
b) Dé el estimador de máxima verosimilitud de θ en función de h y h'.
c) Muestre que el estimador de máxima verosimilitud es el mismo que el del método de los momentos.
d) Aplique lo anterior para los casos siguientes:
i) $X \sim Binomial(N, p)$ (N conocido)
ii) $X \sim Poisson(\lambda)$.

5. Sean T_i, $i = 1, ..., I$ estimadores del parámetro θ tales que : $E(T_i) = \theta + b_i$, $b_i \in R$
Se define un nuevo estimador T de θ como $T = \sum_{i=1}^{I} \lambda_i T_i$

a) Dé una condición sobre los λ_i para que T sea insesgado.
b) Suponga que $b_i = 0 \; \forall i$ (estimadores insesgados). Plantee el problema de encontrar los coeficientes λ_i para que la varianza de T sea mínima.
c) Suponiendo que los T_i son no correlacionados , resuelva el problema planteado.
d) Sean X_{ij}, $i = 1 \dots M, j = 1 \dots n_i$ M m.a.s. independientes entre si, de variables aleatorias X^i con distribuciones normales de varianza común σ^2.
Sea $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$, el estimador insesgado de la varianza calculado en la muestra i.
Demuestre que $S^2 = \frac{1}{\sum_{i=1}^{M} n_i - M} \sum_{i=1}^{M} (n_i - 1) s_i^2$ es el estimador lineal insesgado de varianza mínima para σ^2.

Capítulo 4

INTERVALOS DE CONFIANZA DE NEYMANN

Vimos en el capítulo anterior métodos de estimación puntual. Pero no podemos esperar que la estimación que producen coincide exactamente con el verdadero valor del parámetro θ. En este capítulo buscaremos construir intervalos que proporcionen una precisión de las estimaciones encontradas. La precisión esta dada por la diferencia δ entre el estimador $\hat{\theta}$ y el parámetro θ. Si bien no conocemos a δ podemos dar como vimos anteriormente su media (el sesgo) y su varianza (el error cuadrático medio) o bien un cota u tal que es poco probable que el error $|\delta|$ sobrepase a u:

$$I\!P(|\delta| \geq u) = I\!P(|\hat{\theta} - \theta| \geq u) = \alpha$$

donde α es una probabilidad pequeña fijada a priori y $1 - \alpha$ se llama *nivel de confianza*.

Se puede escribir también $I\!P(\hat{\theta} - u \leq \theta \leq \hat{\theta} + u) = 1 - \alpha$. Si α es pequeño, se puede decir que es altamente probable de encontrar el parámetro θ en el el intervalo $|\hat{\theta} - u, \hat{\theta} + u|$, dicho de otra manera, **el intervalo $|\hat{\theta} - u, \hat{\theta} + u|$ cubre el parámetro θ con alta probabilidad ($1 - \alpha$)**.

Se dice entonces que $|\hat{\theta} - u, \hat{\theta} + u|$ es un intervalo de confianza para θ de nivel de confianza igual a $1 - \alpha$.

Observemos que aquí el intervalo $|\hat{\theta} - u, \hat{\theta} + u|$ es aleatorio y la función de distribución (ó de probabilidad) del error δ depende de la función de distribución (ó probabilidad) del estimador $\hat{\theta}$.

Apliquemos lo anterior a algunos casos.

4.1 INTERVALO PARA UNA MEDIA

Dada la función de distribución de población de la v.a. $X \sim \mathcal{N}(\theta, \sigma^2)$ y $x_1, x_2, ..., x_n$ los valores obtenidos sobre una muestra aleatoria simple de esta población, entonces la media muestral $\bar{x}_n \sim \mathcal{N}(\theta, \frac{\sigma^2}{n})$. Para construir el intervalo de confianza tenemos que considerar si la varianza σ^2 es conocida o no.

4.1.1 Caso de la varianza poblacional conocida

Si se supone que $X \sim \mathcal{N}(\theta, \sigma^2)$, se conoce la distribución del error del estimador $\bar{x}_n$ de θ: $\delta = \bar{x} - \theta \sim \mathcal{N}(0, \frac{\sigma^2}{n})$.

Para obtener el valor u dado un nivel de confianza $1-\alpha$, observamos que $\frac{\sqrt{n}}{\sigma}\delta = \frac{\sqrt{n}}{\sigma}(\bar{x}-\theta) \sim \mathcal{N}(0,1)$.

Luego usando la tabla de la distribución normal $\mathcal{N}(0,1)$ (o un programa computacional) obtenemos el valor a tal que:

$$I\!P(-a \leq Z \leq a) = 1 - \alpha$$

si $Z \sim \mathcal{N}(0,1)$, deducimos que

$$I\!P(\bar{x} - a\frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{x} + a\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

El intervalo $[\bar{x} - a\frac{\sigma}{\sqrt{n}}, \bar{x} + a\frac{\sigma}{\sqrt{n}}]$ define un intervalo de confianza para la media θ de nivel de confianza $(1-\alpha)$.

Queremos que el largo del intervalo de confianza sea pequeño porque es un indicador de la precisión de la estimación, pero acá el nivel de confianza también lo es. Aquí el largo del intervalo es igual a $2a\frac{\sigma}{\sqrt{n}}$. Para n y α dados, mientras más grande es la variabilidad en la población, más largo será el intervalo. Para tener un intervalo más pequeño, es decir, con más precisión sin cambiar el nivel de confianza y la varianza σ^2, se tendrá que aumentar el tamaño n de la muestra. Si se quiere aumentar la precisión sin cambiar el tamaño de la muestra se tiene que aumentar el nivel de confianza, es decir disminuir α y tener un intervalo más grande.

Por ejemplo, para $\alpha = 0.05$, se obtiene el intervalo $[\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}]$. Pero con $\alpha = 0.01$, el intervalo es $[\bar{X} - 2.56\sigma/\sqrt{n}, \bar{X} + 2.56\sigma/\sqrt{n}]$. El largo del intervalo aumenta.

4.1.2 Caso de la varianza poblacional desconocida

Si no se supone que la varianza σ^2 es conocida, se tiene que usar un estadístico cuya distribución muestral no dependa de σ^2. Eso nos lleva a usar un estadístico t de Student definida en 2.4.4. En efecto $Z = \frac{\sqrt{n}(\bar{x}-\theta)}{\sigma} \sim \mathcal{N}(0,1)$ y $U = \frac{n}{\sigma^2}s_n^2 = \frac{\sum(x_i-\bar{x})}{\sigma^2} \sim \chi^2_{n-1}$ y son independientes, luego $\frac{Z}{\sqrt{u/(n-1)}} \sim t_{n-1}$. El estadístico depende del error δ, por lo tanto de θ, pero no de la varianza σ^2 desconocida:

$$T = \frac{\bar{x} - \theta}{s_n/\sqrt{(n-1)}}$$

que sigue una distribución t Student a $n-1$ grados de libertad.

Procedemos como en el caso de la varianza conocida. Buscamos el valor a para una probabilidad pequeña α dado tal que

$$I\!P(-a \leq t_{n-1} \leq a) = 1 - \alpha$$

Se deduce el intervalo

$$[\bar{x} - a\frac{s_n}{\sqrt{n-1}}, \bar{x} + a\frac{s_n}{\sqrt{n-1}}]$$

para θ de nivel de confianza $1-\alpha$.

4.2 INTERVALO PARA LA VARIANZA

Si los valores muestrales $x_1, x_2, ..., x_n$ son i.i.d. de la $\mathcal{N}(\theta, \sigma^2)$, el estimador de máxima verosimilitud de σ^2 es $\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \theta)^2$ si la media θ es conocida y $s_n^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ si θ es desconocida. El estadístico que se utilizará en la construcción del intervalo de confianza depende entonces de si la media poblacional es conocida o no.

4.2.1 Caso de la media poblacional conocida

En este caso se usa el estadístico $U_n = n\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_n^2$. Considerando que si

$$I\!P(u_1 \leq U_n \leq u_2) = 1 - \alpha$$

entonces

$$I\!P(n\frac{\hat{\sigma}^2}{u_2} \leq \sigma^2 \leq \frac{\hat{\sigma}^2}{u_1}) = 1 - \alpha$$

expresión que nos permite construir un intervalo para σ^2 de nivel de confianza igual a $1 - \alpha$.

El problema aquí es que nos sabemos como elegir el par de cotas (u_1,u_2) que nos son únicas con el mismo nivel de confianza $1 - \alpha$. Considerando que el largo del intervalo para un nivel de confianza $1 - \alpha$ dado esta relacionado con la precisión de la estimación, lo mejor sería encontrar el intervalo más pequeño entre todos los intervalos de mismo nivel de confianza. Pero este problema no tiene una solución simple. En la práctica lo que lleva a simplificar el problema es tomar las cotas u_1 y u_2 tales que:

$$I\!P(U_n \leq u_1) = I\!P(U_n \geq u_2) = \frac{\alpha}{2}$$

Por ejemplo, con una muestra de tamaño $n = 20$ para $\alpha = 0.05$, $U_1 = 9.59$ y $U_2 = 34.17$, se obtiene el intervalo $[20\frac{\hat{\sigma}^2}{34.17}, 20\frac{\hat{\sigma}^2}{9.59}]$. Pero con $\alpha = 0.01$, el intervalo es $[20\frac{\hat{\sigma}^2}{39.90}, 20\frac{\hat{\sigma}^2}{7.43}]$.

El largo del intervalo aumenta cuando se disminuye α y disminuye cuando se aumenta n.

4.2.2 Caso de la media poblacional desconocida

Ahora se usa el estadístico $U_{n-1} = \frac{\sum(X_i - X)^2}{\sigma^2} \sim \chi_{n-1}^2$.

Considerando que si

$$I\!P(u_1 \leq U_{n-1} \leq u_2) = 1 - \alpha$$

entonces

$$I\!P(n\frac{s_n^2}{u_2} \leq \sigma^2 \leq \frac{s_n^2}{u_1}) = 1 - \alpha$$

expresión que nos permite construir un intervalo para σ^2 de nivel de confianza igual a $1 - \alpha$ cuando la media poblacional es desconocida. Como en el caso anterior se elige u_1 y u_2 de manera que:

$$I\!P(U_{n-1} \leq u_1) = I\!P(U_{n-1} \geq u_2) = \frac{\alpha}{2}$$

o sea

$$\mathbb{P}(\frac{\sum(x_i-\bar{x})^2}{u_2} \leq \sigma^2 \leq \frac{\sum(x_i-\bar{x})^2}{u_1}) = 1-\alpha$$

Por ejemplo, con una muestra de tamaño $n = 20$ para $\alpha = 0.05$, $U_1 = 8.91$ y $U_2 = 32.85$, se obtiene el intervalo $[20\frac{\hat{\sigma}^2}{32.85}, 20\frac{\hat{\sigma}^2}{8.91}]$. Pero con $\alpha = 0.01$, el intervalo es $[20\frac{\hat{\sigma}^2}{38.58}, 20\frac{\hat{\sigma}^2}{8.84}]$.

4.3 LA DIFERENCIA DE DOS MEDIAS

Sean dos poblaciones normales $\mathcal{N}(\mu_1, \sigma_1^2)$ y $\mathcal{N}(\mu_2, \sigma_2^2)$. Se considera una muestra aleatoria simple de tamaño n_1 de la primera población y una muestra aleatoria simple de tamaño n_2 de la segunda población, las dos muestras siendo independientes. Si $\bar{x}_1$ y $\bar{x}_2$ son las medias muestrales respectivas, $d = \bar{x}_1 - \bar{x}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$.

Si las varianzas son conocidas entonces un intervalo para δ la diferencia de las medias de las poblaciones esta dado por:

$$[\bar{x}_1 - \bar{x}_2 - u\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \quad \bar{x}_1 - \bar{x}_2 + u\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

con u determinado a partir de las tablas de la distribución normal según el nivel de confianza $1-\alpha$.

Si las varianzas no son conocidas, para encontrar un estadístico que sirva y cuya distribución no dependa de estas varianzas, hay que hacer alguno supuesto suplementario. En efecto si tomamos como estimador de la varianza de la diferencia $\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$ con s_1^2 y s_2^2 las varianzas muestrales sesgadas, $\frac{n_1 s_1^2}{\sigma_1^2} + \frac{n_2 s_2^2}{\sigma_2^2} \sim \chi^2_{n_1+n_2-2}$ y

$$\frac{(\bar{s}_1 - \bar{s}_2 - (\mu_1 - \mu_2))/\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}{\sqrt{(\frac{n_1 s_1^2}{\sigma_1^2} + \frac{n_2 s_2^2}{\sigma_2^2})/(n_1+n_2-2)}} \sim t_{n_1+n_2-2}$$

que depende de la varianzas desconocidas σ_1^2 y σ_2^2.

Si se supone que estas varianzas son proporcionales: $\sigma_2^2 = k^2\sigma_1^2$, entonces se tiene un estadístico que no depende de σ_1^2 y σ_2^2:

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{(\frac{k^2 n_1 s_1^2 + n_2 s_2^2}{k^2(n_1+n_2-2)})(\frac{k^2 n_1 + n_2}{n_1 n_2})}} \sim t_{n_1+n_2-2}$$

Usualmente si toma $k = 1$. Se obtiene el valor de u tal que

$$\mathbb{P}(-u \leq t_{n_1+n_2-2} \leq u) = 1-\alpha$$

En este caso el intervalo para $\mu_1 - \mu_2$ es:

$$[\bar{x}_1 - \bar{x}_2 - u\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}, \quad \bar{x}_1 - \bar{x}_2 + u\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

Sea, por ejemplo, $n_1 = 15$, $n_2 = 17$, $\bar{x}_1 = 4$, $\bar{x}_2 = 5$, $S_1^2 = 2.0$, $S_2^2 = 2.5$ y $\alpha = 0.05$. En este caso $u = 2.042$ y el intervalo es

$$[-1 - 2.042\sqrt{\frac{72.5}{30}}\sqrt{\frac{32}{255}}, \quad -1 + 2.042\sqrt{\frac{72.5}{30}}\sqrt{\frac{32}{255}} = [-2.1247, \quad 0.1247]$$

Se observa que el intervalo cubre apenas el valor 0, lo que indica que las medias son bastante distintas.

4.4 EL COCIENTE DE DOS VARIANZAS

En el párrafo anterior para construir el estadístico t de Student se hizo el supuesto que las varianzas σ_1^2 y σ_2^2 son proporcionales. Este supuesto podra verificarse usando un intervalo de confianza para el cociente de las dos varianzas $\frac{\sigma_1^2}{\sigma_2^2}$.

Sean dos poblaciones normales $\mathcal{N}(\mu_1, \sigma_1^2)$ y $\mathcal{N}(\mu_2, \sigma_2^2)$ y una muestra aleatoria simple de cada población, ambas tomadas de manera independiente. Nos interesamos en el cociente de las dos varianzas: $\frac{\sigma_1^2}{\sigma_2^2}$.

El estadístico $n_1 s_1^2/\sigma_1^2 \sim \chi^2_{n_1-1}$ y el estadístico $n_2 s_2^2/\sigma_2^2 \sim \chi^2_{n_2-1}$, siendo estos independientes. Vamos a definir una nueva distribución para el cociente de dos varianzas empíricas llamada distribución F de Fisher.

Definición 4.4.1 *El cociente de una variable χ^2 a r grados de libertad y de una variable χ^2 a s grados de libertad, independientes entre si, sigue una distribución llamada F de Fisher a r y s grados de libertad. Se denota $F_{r,s}$.*

Mostramos que $F_{r,s}$ tiene una función de densidad igual a:

$$h(y) = \frac{\Gamma(\frac{r+s}{2})}{\Gamma(\frac{r}{2})\Gamma(\frac{s}{2})} \frac{r^{r/2} s^{s/2} y^{(r/2)-1}}{(ry+s)^{(r+s)/2}} \quad \forall y > 0$$

Sean $U \sim \chi_r^2$ y $V \sim \chi_s^2$ independientes entre si. Como U y V son independientes, se puede calcular fácilmente la función de densidad conjunta de (U,V):

$$f(u,v) = \frac{u^{(r/2)-1} e^{-u/2}}{2^{r/2}\Gamma(r/2)} \frac{v^{(s/2)-1} e^{-v/2}}{2^{s/2}\Gamma(s/2)}$$

Con el cambio de variables:$(U,V) \longrightarrow (Y,Z)$ con $U = rYZ/s$ y $V = Z$, obtenemos la densidad conjunta de (Y,Z):

$$g(y,z) = \frac{(r/s)z}{2^{(r+s)/2}\Gamma(r/2)\Gamma(s/2)} (r/s)^{(r/2)-1} y^{(r/2)-1} z^{(r+s-1)/2} e^{-1/2(ry/s+1)z}$$

Se deduce la densidad marginal de Y:

$$f(y) = \int_0^\infty g(y,z)dz = \frac{\Gamma(\frac{r+s}{2} r^{r/2} s^{s/2} y^{(r/2)-1}}{\Gamma(r/2)\Gamma(s/2)(ry+s)^{(r+s)/2}}$$

Nota 4.4.2 *Observamos que si* $Y \sim F_{r,s}$ *entonces* $1/Y \sim F_{s,r}$. *Además* $\frac{rY/s}{1+rW/s} \sim \beta eta((r-2)/2,(s-2)/2)$.

Aquí el estadístico $\frac{n_1 S_1^2/(n_1-1)\sigma_1^2}{n_2 S_2^2/(n_2-1)\sigma_2^2} \sim F_{n_1-1,n_2-1}$, lo que permite construir un intervalo de confianza para el cociente σ_1^2/σ_2^2.

Si $I\!P(a \leq F_{n_1-1n2-1} \leq b) = 1-\alpha$, entonces el intervalo es

$$[\frac{n_1 s_1^2(n_2-1)}{bn_2 s_2^2(n_1-1)}, \quad \frac{n_1 s_1^2(n_2-1)}{an_2 s_2^2(n_1-1)}]$$

4.5 INTERVALO PARA LA PROPORCIÓN

Sea la proporción θ de piezas defectuosas en un lote de piezas fabricadas por la industria 2.1.5. El número de piezas defectuosas encontradas en una muestra aleatoria simple de tamaño n sigue una distribución binomial $B(n,\theta)$. Construir un intervalo de confianza para una proporción es más complicado que construirlo para una media o varianza. Cuando n es pequeño hay que usar la distribución binomial (tablas y ábacos fueron calculados para determinar valores de θ_1 y θ_2 para los diferentes valores de k y n y de nivel de confianza $1-\alpha$).

Cuando n es grande, se puede usar la aproximación a la distribución normal $\mathcal{N}(n\theta, n\theta(1-\theta))$, pero subsiste un problema ya que la varianza depende también de θ. Una manera de proceder considerando la proporción empírica $\hat{p} \approx \mathcal{N}(\theta, \frac{\theta(1-\theta)}{n})$. Se tiene entonces:

$$I\!P(|\ \frac{\sqrt{n}(\hat{p}-\theta)}{\sqrt{\theta(1-\theta)}}\ | \leq u) = 1-\alpha$$

Lo que equivale a:

$$I\!P(n(\hat{p}-\theta)^2 - u^2\theta(1-\theta) \leq 0) = 1-\alpha$$

Las soluciones de la ecuación:

$$(n+u^2)\theta^2 - (2n\hat{p}+u^2)\theta + n\hat{p}^2 = 0$$

siendo $\frac{2n\hat{p}+u^2 \pm \sqrt{u^4+4n\hat{p}u^2-4nu^2\hat{p}^2}}{2(n+u^2)}$, se obtiene el intervalo:

$$[\frac{n}{n+u^2}(\hat{p}+\frac{u^2}{2n}) - u\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{u^2}{4n^2}}, \quad \frac{n}{n+u^2}(\hat{p}+\frac{u^2}{2n}) + u\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{u^2}{4n^2}}.$$

Para n grande, se puede aproximar por:

$$[\hat{p} - u\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + u\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

4.6 EJERCICIOS

1. Sea una m.a.s. $\{x_1, ...x_n\}$ de una distribución normal de media θ desconocida y varianza σ^2 conocida.

a) Dé el número mínimo n del tamaño de la muestra para que un intervalo de confianza I a 95% tenga un largo L a lo más igual a 0.016 σ.
b) Sea $L = \sigma/5$. Dé el nivel de confianza $1-\alpha$ cuando n=10, 20, 30 y 100.
c) Repetir b) con σ^2 desconocido. Comente.
d) Dé el intervalo de confianza de largo mínimo para θ con un nivel de confianza de 95%, cuando $\sigma^2 = 4$.

2. Una empresa desea estimar el promedio de tiempo que necesita una secretaria para llegar a su trabajo. Se toma una m.a.s. de 36 secretarias y se encuentra un promedio de 40 minutos. Suponiendo que el tiempo de trayecto proviene de una $\mathcal{N}(\mu, \sigma^2)$, con $\sigma = 12$, dé un intervalo de confianza para la media μ.

3. Se dispone de 10 muestras de sangre tomadas en las mismas condiciones a una misma persona. Se obtiene para cada una la dosis de Colesterol (en gramos) *245, 248, 250, 247, 249, 247, 247, 246, 246, 248*. Cada medida puede considerarse como una realización particular de la variable "tasa de Colesterol" $X \sim \mathcal{N}(\mu, \sigma^2)$.

a) Dé un intervalo de confianza para μ al 95% suponiendo $\sigma^2 = 1.5$.
b) Dé un intervalo de confianza para μ al 95% suponiendo σ^2 desconocido.
c) Construya un intervalo de confianza para σ^2 al 95% .

4. Se tienen 2 muestras de tamaños n_1 y n_2 de una misma v.a. X medida sobre dos poblaciones distintas. Se asume que para ambas poblaciones X sigue una distribución Normal con medias μ_1, μ_2 y varianzas σ_1^2, σ_2^2, respectivamente.

a) Construya un intervalo de confianza para $\mu_1 - \mu_2$, suponiendo que $\sigma_2^2 = k^2\sigma_1^2$ en que k es una constante conocida.
b) Muestre que los extremos del intervalo anterior convergen en probabilidad si los tamaños de las muestras crecen.
c) Se supone ahora la constante k desconocida. Dé un método para construir un intervalo de confianza para la constante k.
d) ¿ Que inconveniente cree ud. que tiene este método?

5. Considere una v.a. $X \sim \mathcal{N}(\mu, 1)$ y una m.a.s. de X con una sóla observación x. Dada una constante $a > 0$, se define el intervalo aleatorio: $C_a(x) = |min(0, x-a), max(0, x+a)|$.

a) Muestre que $\mathbb{P}(\mu \in C_a(x)/\mu = 0) = 1 \ \forall x$.
b) Muestre que $C_a(x)$ es un intervalo de confianza para μ de nivel de confianza $1-\alpha = 95\%$, cuando a=1.65.
c) Sea $\pi(\mu) = 1$ $(\forall \mu)$ una distribución a priori para μ. Deducir la distribución a posteriori de μ dado x.
d) Sea Φ la función de distribución de la normal $\mathcal{N}(0,1)$. Muestre que se encuentra una probabilidad condicional

$$\mathbb{P}(\mu \in C_a(x)/x) = \begin{cases} \Phi(-x) - \Phi(-a) & \text{si } x < -a \\ \Phi(a) - \Phi(-a) & \text{si } -a < x < a \\ \Phi(a) - \Phi(-x) & \text{si } x > a \end{cases}$$

e) Deducir que, para a=1.65, la probabilidad condicional $I\!P(\mu \in C_a(x)/x) \geq 0.90$ y que $\lim_{a\to\infty} I\!P(\mu \in C_a(x)/x) = 1$.

Capítulo 5

TESTS DE HIPOTESIS

5.1 ¿COMÓ UN JUEZ SENTENCIA?

Sir Ronald Fisher (párrafo 1.1) describe en su escrito Experimental Design (1935) lo siguiente: Una dama británica (Lady, en el texto) declara ser capaz de distinguir si la leche fue puesta antes o después del té en su taza. Fisher propone entonces un experimento para comprobar lo que dice esta dama. Se prepara 4 tazas de té en las cuales se puso la leche antes del té y 4 otras tazas en las cuales se puso la leche después del té. Se presentan al azar las 8 tazas a la Dama, que las prueba todas y decide cuáles tuvieron la leche antes y cuáles tuvieron la leche después. Si acierta a las 8 tazas, ¿diría Ud que fue solamente suerte? ¡Nosotros, diríamos que no! En efecto tenemos repuestas posibles y una sola repuesta correcta. Por lo tanto, si la dama ha contestado al azar, tenía una probabilidad de 1/70 de dar la repuesta correcta. ¿Es eso suerte?

En el capítulo 3, se presentaron métodos que permiten encontrar los valores de los parámetros desconocidos de la distribución de población y en el capítulo anterior, la estimación por intervalo permite dar una cierta indicación sobre la **precisión** de la estimación puntual. Tales estimaciones puntuales y por intervalo, que fueron obtenidas a partir de valores muestrales, permiten formarse una opinión sobre la población y entonces darse una **hipótesis** de trabajo.

Ejemplos:

- Antes de apostar "cara" o "sello" en el lanzamiento de una moneda, se tiene que postular que la moneda está equilibrada. La hipótesis de trabajo es entonces que el parámetro p=probabilidad de sacar "cara" de la distribución de Bernoulli es

$$p = 0.5$$

- Un agricultor se compromete a entregar a una fábrica de azúcar remolacha con un cierto porcentaje p_o de glucosa; la hipótesis de trabajo es entonces

$$p = p_o \qquad o \qquad p \geq p_o$$

- Los hombres chilenos pretenden ser más altos que los argentinos en promedio; si μ_1 y μ_2 son las tallas promedias respectivas de los hombres chilenos y argentinos, la hipótesis de trabajo es

$$\mu_1 \geq \mu_2$$

- En el ejemplo sobre la estimación puntual de la talla promedia μ_1 de los hombres chilenos, se hizo la hipótesis de trabajo que la v.a. X talla de los hombres chilenos sigue una distribución

$$F \sim Normal$$

En los cuatro casos se procederá de la misma manera: se tiene una hipótesis de trabajo y un experimento que nos proporciona una muestra de observaciones; se trata de decidir si la hipótesis planteada es compatible con lo que se puede aprender del estudio de los valores muestrales. Se tiene que encontrar un procedimiento para decidir si la muestra que se obtuvo esta en acuerdo con la hipótesis de trabajo. Naturalmente no se espera que, para cualquier muestra, el valor empírico obtenido en la muestra coincide con el valor esperado de la hipótesis; el problema es entonces decidir si la desviación encontrada entre el valor esperado y el valor observado en la muestra es demasiado grande para poner en duda la hipótesis de trabajo. Ahora bien si se pone en duda la hipótesis original, entonces se la rechaza en favor de una **hipótesis alternativa**.

En efecto, en el ejemplo de la moneda, si se encuentra en 100 lanzamientos un 45% de caras, ¿debemos rechazar la hipótesis $p = 1/2$,la moneda esta equilibrada? y si se rechaza esta hipótesis, ¿será a favor de la hipótesis $p \leq 1/2$, la moneda esta cargada en los "sellos"?

Se distingue la hipótesis de trabajo llamándola **hipótesis nula** y la cual se confronta a una **hipótesis alternativa**.

> *¿Con qué grado de desacuerdo uno tiene que abandonar la hipótesis nula a favor de la hipótesis alternativa?*

Para decidir entre las dos hipótesis, se necesita una **regla de decisión**.

Cualquiera regla de decisión debería tratar de minimizar los errores de decisión. Si δ es la regla de decisión adoptada y $\alpha(\delta)$ la probabilidad de equivocarse cuando la hipótesis nula es cierta y $\beta(\delta)$ la probabilidad de equivocarse cuando la hipótesis alternativa es cierta, uno buscara minimizar ambas probabilidades de equivocarse. Pero veremos, a través de un ejemplo, que no se puede disminuir los dos errores al mismo tiempo, en particular para tener $\alpha(\delta)$ nula se hace $\beta(\delta)$ igual a 1 viceversa.

Entonces para decidir entre la hipótesis nula y la hipótesis alternativa se corre el riesgo de obtener resultados falsos y tomar decisiones incorrectas con consecuencias graves. ¿Qué riesgo estamos dispuestos a asumir?. ¿Los dos errores tienen la misma importancia para la decisión a tomar? Se puede entender un poco mejor el procedimiento de un test de hipótesis haciendo el paralelo con las prácticas procesales en el sistema judicial. Frente a un acusado, el juez tiene que tomar la decisión grave entre dos casos: "El acusado es inocente" o "el acusado es culpable". El juez escucha el abogado de la acusación, que recopiló los antecedentes del caso en contra del acusado y escucha la defensa que trata de refutar las evidencias de la acusación.

La acusación debe presentar suficiente evidencia que permitan convencer al juez de la culpabilidad del acusado, al menos más allá de "una duda razonable". El resultado por defecto, si no hay suficiente pruebas de culpabilidad, es que "el acusado es inocente". Solo en el caso de una evidencia abrumadora, el juez declarará "el acusado culpable". El veredicto será automáticamente "inocente". Es el rol de la acusación de tener un caso para culpabilizar el acusado y no a la defensa de probar la inocencia, pero refutar la evidencia presentada por la acusación.

Ahora si el juez declara culpable a un inocente, no prueba que el acusado sea realmente culpable. Pueden ocurrir errores por una mala defensa. Sin embargo, los procedimientos de investigación y

El juez tiene que tomar una decisión grave después de escuchar las evidencias presentadas por la acusación y la defensa:
"El acusado es inocente" o
"El acusado es culpable".

de justicia son establecidos para controlar la probabilidad que tal error ocurra. Por otra parte, si un acusado es absuelto, no significa que es inocente. Significa que la acusación no encontró evidencia adecuada de su culpabilidad.

El juez puede cometer dos tipos de errores, pero no se le da la misma importancia a ambas. El juez va a tratar de minimizar el error de *condenar a un inocente*.

Dada una hipótesis nula H_o, vimos que $\alpha(\delta)$ es la probabilidad de rechazar la hipótesis H_o con la regla δ cuando H_o es cierta. Ahora bien la regla δ se basa en los valores muestrales, es decir en evidencias; si la muestra es de tamaño n y los valores muestrales pertenecientes a Q, una regla de decisión δ consiste en dividir el dominio Q^n del conjunto de todas las muestras de tamaño n en dos partes disjuntas: la parte W en donde las muestras conducen a rechazar la hipótesis nula H_o, es decir evidencias contra el acusado presentadas por la acusación, y la parte $\overline{W}$ en donde no se rechaza H_o, evidencias presentadas por la defensa que recusan las evidencias de la acusación. La parte W se llama **región de rechazo de** H_o o **región crítica del test**, que consiste en las evidencias que permiten declarar culpable al acusado.

Como la región crítica del test es aquella en donde se rechaza H_o, debería tomar en cuenta la hipótesis alternativa.

Una regla de decisión consiste entonces en determinar la región crítica del test en función de las dos hipótesis.

5.2 HIPÓTESIS ESTADÍSTICAS

Las hipótesis estadísticas son muy precisas: se refieren al comportamiento de variables aleatorias. Pero en los ejemplos expuestos en el párrafo anterior, se observó que las hipótesis no son todas del mismo tipo. En los tres primeros ejemplos, la hipótesis concierne solamente a los valores de parámetros de una distribución cuya forma es especificada a priori y no está puesta en duda. Tales hipótesis se llaman **hipótesis parámetricas**. En el último ejemplo "$F \sim Normal$", es la distribución completa que esta puesta en juicio; se habla de **hipótesis no parámetricas**.

Por ejemplo, sea una v.a. X de distribución $F_\theta(x)$, que depende de un parámetro θ. Si Ω es el espacio del parámetro θ y Ω_o un subconjunto de Ω, entonces

$$H : \theta \in \Omega_o$$

es una hipótesis parámetrica, mientras que

$$H : F \sim Normal$$

es una hipótesis no parámetrica.

Se puede clasificar también las hipótesis parámetricas según su grado de especificidad. Cuando en la hipótesis parámetrica $H_o : \theta \in \Omega_o$, Ω_o esta reducido a un sólo valor, entonces se habla de **hipótesis simple**, sino se habla de **hipótesis compuesta**.

5.3 TEST DE HIPÓTESIS PARÁMETRICAS

Trataremos en primer lugar los tests de hipótesis parámetricas para hipótesis simples antes de tratar el caso general apoyandonos en los resultados del caso de las hipótesis simples. Encontrar una regla de decisión es encontrar una región crítica del test. ¿Como hacerlo minimizando los errores de decisión? Para eso usaremos la función de potencia.

5.3.1 Función de potencia

Sea un test de hipótesis sobre el parámetro θ ($\theta \in \Omega$) de la distribución F de una v.a. X.

$$H_o : \theta \in \Omega_o \quad contra \quad H_1 : \theta \in \Omega_1$$

Si una regla de decisión nos condujo a una región crítica W para el test, entonces para cada valor de $\theta \in \Omega$, determinaremos la probabilidad $\pi(\theta)$ que la regla de decisión definida por W nos conduce a rechazar H_o cuando el parámetro vale θ.

Definición 5.3.1 *La FUNCIÓN DE POTENCIA DEL TEST es*

$$\pi(\theta) = I\!P(decidir \quad rechazar \quad H_o|\theta)$$

¡OJO! aquí θ no es una variables aleatoria.

W es la región crítica del test y $\underline{x}$ el vector de los valores muestrales, entonces

$$\pi(\theta) = I\!P(\underline{x} \in W|\theta) \quad \forall\theta \in \Omega$$

Luego la región crítica ideal es aquella que produce una función de potencia tal que:

$$\pi(\theta) = \begin{cases} 0 & \text{si } \theta \in \Omega_c \\ 1 & \text{si } \theta \in \Omega_1 \end{cases}$$

En efecto, para todo $\theta \in \Omega_o$, la decisión de rechazar H_o es una decisión equivocada, entonces $\pi(\theta)$ es **una probabilidad de error de tipo I** (ó riesgo de primer especie). Por otro lado, para todo $\theta \in \Omega_1$, la decisión de rechazar H_o es una decisión correcta, entonces $1-\pi(\theta)$ es **una probabilidad de error de tipo II** (ó riesgo de segundo especie).

Definición 5.3.2 *Se llama TAMAÑO del test a* $sup\{\pi(\theta)|\theta \in \Omega_o\}$

Veamos en un ejemplo que una región crítica ideal, con $\pi(\theta) = 0$ en Ω_o y $\pi(\theta) = 1$ en Ω_1, no existe.

Sea $x_1, x_2, ..., x_n$ una m.a.s. de una v.a. X uniforme en $|0,\theta|$ con $\theta > 0$.
Consideramos la hipótesis nula $H_o : 3 \leq \theta \leq 4$ contra la hipótesis alternativa $H_1 : \theta < 3$ o $\theta > 4$.
Supongamos que una regla de decisión δ nos llevo a decidir de no rechazar a la hipótesis nula H_o cuando $\max\{x_1, x_2, ..., x_n\}$ de una m.a.s. de la v.a. X esta en el intervalo $|2.9,4.1|$ y a rechazar H_o en el caso contrario. Luego la región crítica del test es un subconjunto $W \subset I\!R^n$ tal que $\max\{x_1, x_2, ..., x_n\} < 2.9$ o $> 4.1\}$. La función de potencia del test es entonces:

$$\pi(\theta) = I\!P(\max\{x_1 + x_2, ..., x_n\} < 2.9|\theta) + I\!P(\max\{x_1, x_2, ..., x_n\} > 4.1|\theta)$$

$$\text{Si } \theta \leq 2.9 \Rightarrow \left\{ \begin{array}{l} I\!P(\max\{x_1, x_2, ..., x_n\} < 2.9|\theta) = 1 \\ I\!P(\max\{x_1, x_2, ..., x_n\} > 4.1|\theta) = 0 \end{array} \right\} \Rightarrow \pi(\theta) = 1$$

$$\text{Si } 2.9 < \theta \leq 4.1 \Rightarrow \left\{ \begin{array}{l} I\!P(\max\{x_1, x_2, ..., x_n\} < 2.9|\theta) = (\frac{2.9}{\theta})^n \\ I\!P(\max\{x_1, x_2, ..., x_n\} > 4.1|\theta) = 0 \end{array} \right\} \Rightarrow \pi(\theta) = (\frac{2.9}{\theta})^n$$

$$\text{Si } \theta > 4.1 \Rightarrow \left\{ \begin{array}{l} I\!P(\max\{x_1, x_2, ..., x_n\} < 2.9|\theta) = (\frac{2.9}{\theta})^n \\ I\!P(\max\{x_1, x_2, ..., x_n\} > 4.1|\theta) = 1 - (\frac{4.1}{\theta})^n \end{array} \right\} \Rightarrow \pi(\theta) = 1 + (\frac{2.9}{\theta})^n - (\frac{4.1}{\theta})^n$$

El tamaño del test es igual a $\alpha = Sup\{\pi(\theta)|3 \leq \theta \leq 4\} = \pi(3) = (\frac{2.9}{3})^n$

En los gráficos 5.1, se muestra la función de potencia para los casos $n = 10$ y $n = 50$. Se observa que el tamaño del test $\alpha = 0.10$, es decir que en el intervalo $|3, 4|$ la probabilidad de equivocarse no sobrepasa 10%. Pero el error de tipo II, que es igual a $1 - \pi(\theta)$ cuando $\theta \in \Omega_o$, puede ser muy elevado; entre 3 y 2.9, el error disminuye de 10% a 0%; pero entre 4 y 4.1 es casi igual a 1.

En este ejemplo si queremos disminuir el tamaño del test α, hay que elegir un intervalo $\overline{W}$ más grande o una muestra de tamaño mayor. Pero en ambos casos se aumentara el error de tipo II. Para tratar de acercarnos a la situación ideal, se puede, por ejemplo, buscar minimizar una función de los dos errores, o bien fijarse una cota máxima para el error de tipo I y minimizar el error de tipo II, como en el caso del juez.

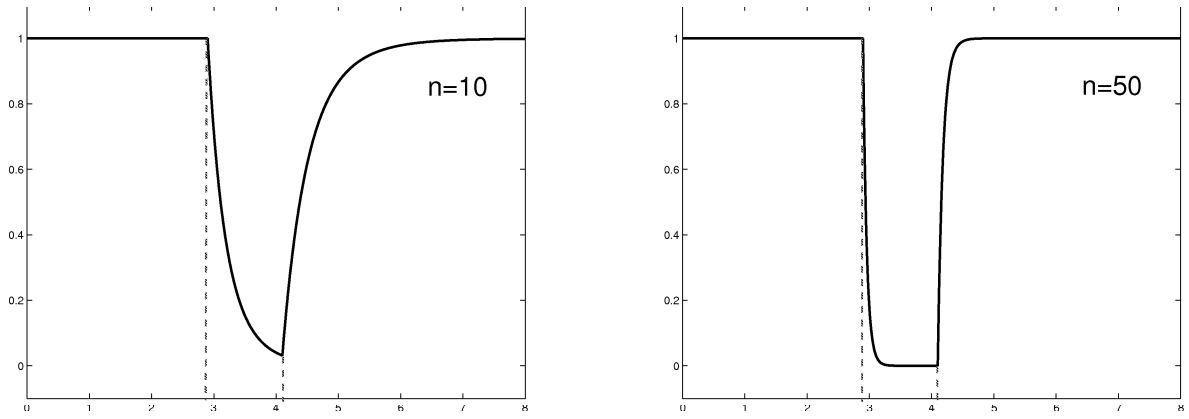


Figura 5.1: Función de potencia para la región crítica $[2.9, 4.1]$

5.3.2 Tests para hipótesis simples

Sean $x_1, x_2, ..., x_n$, los valores muestrales independientes de una v.a. de función de densidad $f_\theta(x)$. Se plantea las hipótesis simples:

$$H_o : \theta = \theta_o \quad contra \quad H_1 : \theta = \theta_1$$

Dada una regla de decisión δ, se tienen los dos errores:

$$\alpha(\delta) = I\!P(rechazar H_o | \theta = \theta_o) \quad (error\ de\ tipo\ I)$$

$$\beta(\delta) = I\!P(no\ rechazar H_o | \theta = \theta_1) \quad (error\ de\ tipo\ II)$$

Tenemos dos estrategias para tratar los dos errores:

(a)Minimizar una función simple de los dos errores, como el promedio ponderado de los errores tipo I y tipo II $a\alpha(\delta) + b\beta(\delta)$, con $a + b = 1$.

(b)Fijarse una cota máxima α_o para el error de tipo I y minimizar el error de tipo II.

Mostraremos en primer lugar como minimizar el promedio de los dos errores. Para el según caso usaremos la solución de (a) para encontrar la forma de construir la región crítica en el caso (b).

Dados dos escalares a y b, buscamos minimizar la función $a\alpha(\delta) + b\beta(\delta)$. Se denota $f_o(\underline{x})$ y $f_1(\underline{x})$ a las funciones de verosimilitud dado H_o y dado H_1 respectivamente:

$$f_o(\underline{x}) = \prod_i f(x_i/\theta_o) \quad y \quad f_1(\underline{x}) = \prod_i f(x_i/\theta_1)$$

Teorema 5.3.3 *Si δ^* es la regla de decisión tal que:*

$$\begin{array}{c} se\ rechaza\ H_o\ cuando\ af_o(\underline{x}) < bf_1(\underline{x}), \\ se\ acepta\ H_o\ cuando\ af_o(\underline{x}) > bf_1(\underline{x}), \end{array}$$

entonces $a\alpha(\delta^*) + b\beta(\delta^*) \leq a\alpha(\delta) + b\beta(\delta)\ \ \forall\delta$

Demostración Si W es la región crítica asociada a una regla de decisión δ,

$$\alpha(\delta) = \int ... \int_W f_o(\underline{x}) dx_1 ... dx_n$$

$$\beta(\delta) = \int ... \int_{\overline{W}} f_1(\underline{x}) dx_1 ... dx_n$$

$$a\alpha(\delta) + b\beta(\delta) = a \int ... \int_W f_o(\underline{x}) dx_1 ... dx_n + b(1 - \int ... \int_W f_1(\underline{x}) dx_1 ... dx_n)$$

Luego $a\alpha(\delta) + b\beta(\delta)$ es mínimo cuando $\int ... \int_W (af_o(\underline{x}) - bf_1(\underline{x})) dx_1 ... dx_n$ es mínimo.

Es decir si: $\begin{cases} af_o(\underline{x}) - bf_1(\underline{x}) < 0\ \ \forall \underline{x} \in W \\ af_o(\underline{x}) - bf_1(\underline{x}) > 0\ \ \forall \underline{x} \in \overline{W} \end{cases}$

entonces δ^* es óptimo para estos valores a y b dados. Se observará que $f_o(\underline{x}) - bf_1(\underline{x}) = 0$ es irrelevante, dado que no cambia el mínimo. ■

Para la segunda estrategía (b) supongamos que α_o la cota máxima de error de tipo I que se quiere aceptar. Daremos en primer lugar definiciones.

Definición 5.3.4 *Se llama RAZÓN DE VEROSIMILITUD de la muestra al cociente*

$$\frac{f_1(\underline{x})}{f_o(\underline{x})}$$

Definición 5.3.5 *Se llama NIVEL DE SIGNIFICACIÓN del test a la cota máxima de error de tipo I aceptada.*

Se tiene entonces que buscar una regla de decisión δ que produzca un error de tipo I $\alpha(\delta) \leq \alpha_o$ y tal que $\beta(\delta)$ sea mínimo. El siguiente lema, que deriva del teorema anterior, nos da la forma de proceder.

Lema 5.3.6 *(NEYMAN-PEARSON)*
Si δ^* es una regla de decisión tal que para algún $k > 0$ fijo,
se rechaza H_o, si $\frac{f_1(\underline{x})}{f_0(\underline{x})} > k$
no se rechaza H_o, si $\frac{f_1(\underline{x})}{f_0(\underline{x})} < k$,
entonces para toda regla δ tal que $\alpha(\delta) \leq \alpha(\delta^*)$ se tiene $\beta(\delta) \geq \beta(\delta^*)$.

Sea $x_1, ..., x_n$ una muestra aleatoria simple de la v.a. $X \sim \mathcal{N}(\mu, \sigma^2)$, μ desconocido y σ^2 conocido. Se estudia $H_o : \mu = 1$ contre $H_1 : \mu = 2$. La razón de verosimilitud se escribe:

$$\frac{f_1(\underline{x})}{f_o(\underline{x})} = exp\{-\frac{1}{2\sigma^2}[\sum(x_i - 2)^2 - \sum(x_i - 1)^2]\}$$

$$\frac{f_1(\underline{x})}{f_o(\underline{x})} = exp\{-\frac{1}{2\sigma^2}[-2\sum x_i + 3n]\}$$

$$\frac{f_1(\underline{x})}{f_o(\underline{x})} = exp\{\frac{\sum x_i}{\sigma^2} - \frac{3n}{2\sigma^2}\}$$

La regla de decisión que minimiza al error promedio $a\alpha(\delta) + b\beta(\delta)$ consiste en rechazar H_o si

$$\frac{f_1(\underline{x})}{f_o(\underline{x})} > \frac{a}{b}$$

es decir: $\bar{x} > \frac{3}{2} + \frac{\sigma^2}{n} ln(\frac{a}{b})$

Si $\sigma^2 = 2$ y $n = 20$, la región crítica $\mathcal{R}$, que es de la forma $\{\bar{x} > c\}$ depende de a y b:

si $a = b$, $c = 3/2$, pero si $a > b$, $c > 3/2$ y si $a < b$, $c < 3/2$.

En particular, si $a = 2/3$ y $b = 1/3$, $\mathcal{R} = \{\bar{x} > 1.57\}$, pero si $a = 1/3$ y $b = 2/3$, $\mathcal{R} = \{\bar{x} > 0.143\}$.

El error de tipo I $\alpha(\delta)$ es $\pi(1) = I\!P(\bar{x} > C/\mu = 1)$. Como $\bar{x} \sim \mathcal{N}(1, \sigma^2/n)$ bajo H_o, $\alpha(\delta) = 1 - \Phi(\frac{c-1}{\sigma/\sqrt{n}})$, en que $\Phi(x)$ es la función de distribución de $\mathcal{N}(0, 1)$.

El error $\beta(\delta)$ de tipo II es $1 - \pi(2) = 1 - I\!P(\bar{x} > c/\mu = 1) = I\!P(\bar{x} < c/\mu = 2) = \Phi(\frac{c-2}{\sigma/\sqrt{n}})$

Si $a = b$, como $c = 3/2$, para $n = 20$, se obtiene $\alpha(\delta) = \beta(\delta) = 1 - \Phi(1.58) = 0.057$, pero con $n = 100$, $\alpha(\delta) = \beta(\delta) = 1 - \Phi(3.53) \simeq 0$.

Si se obtuvo una media muestral $\bar{x} = 1.30$ para una muestra aleatoria de tamaño 20, no se rechaza $H_o : \mu = 1$ con un error de tipo 1 de 0.057 cuando se toma a=b; si se toma $a = 0.3$ y $b = 0.7$, se rechaza H_o a favor de H_1 con un error de tipo 1 igual a 0.11.

Si tenemos un nivel de significación fijado en $\alpha_o = 0.05$, entonces se obtiene una región crítica $\mathcal{R} = \{\bar{x} > c\}$ tal que

$$I\!P(\bar{x} > c|\mu = 1) = 0.05$$

Como $\sqrt{n}(\bar{x} - 1) \sim \mathcal{N}(0, 1)$

$$I\!P(\bar{x} > c|\mu = 1) = 1 - \Phi(\sqrt{n}(c-1)/\sqrt{2}) = 0.05$$

Como $\Phi(1.65) = 0.95$, se obtiene que $\sqrt{n}(c-1)/\sqrt{2} = 1.65$, es decir que $c = 1.52$ y $\mathcal{R} = \{\bar{x} > 1.52\}$. En este caso no se rechaza H_o.

5.3.3 Tests Uniformemente Más Potentes (U.M.P.)

Vamos extender ahora los resultados del lema de Neyman-Pearson a hipótesis compuestas.

Sean las hipótesis compuestas $H_o : \theta \in \Omega_o$ contra $H_1 : \theta \in \Omega_1$ $(\Omega_o \bigcap \Omega_1 = \varnothing)$.

Si nos fijamos un nivel de significación α_o, buscamos una regla de decisión δ tal que la función de potencia cumpla:

$$\pi_\delta(\theta) \leq \alpha_o \ \forall\theta \in \Omega_o \quad y \quad \pi_\delta(\theta) \text{ sea máxima } \forall\theta \in \Omega_1.$$

Ahora bien no es siempre posible encontrar un test δ que satisfaga esta condición. En efecto si $\Omega_1 = \{\theta_1, \theta_2\}$, un test δ podrá tener una potencia máxima para θ_1 pero no necesariamente para θ_2.

En el ejemplo anterior, si tomamos como hipótesis alternativa dos valores $H_1 = \mu \in \{0, 2\}$, entonces como $H_o : \mu = 1$, la región crítica más potente para $\mu = 0$ sera de la forma $\mathcal{R} = \{\bar{x} < c\}$, que, como lo vimos, no es la región crítica más potente para $\mu = 2$.

Definición 5.3.7 *Si un test δ maximiza la función de potencia para todo valor θ de la hipótesis alternativa $H_1 : \theta \in \Omega_1$, se dice que el test δ es uniformemente más potente (U.M.P.); es decir que δ^* es un test U.M.P. al nivel de significación α_o si $\alpha(\delta) \leq \alpha_o$ y si para todo otro test δ tal que $\alpha(\delta) \leq \alpha_o$, se tiene $\pi_\delta(\theta) \leq \pi_{\delta^*}(\theta)$ $\forall\theta \in \Omega_1$.*

Observamos en el ejemplo que la razón de las verosimilitud dado $\mu = \mu_2$ y $\mu = \mu_1$ se escribe:

$$\frac{f_n(\underline{x}|\mu_2)}{f_n(\underline{x}|\mu_1)} = exp\{\frac{n(\mu_2 - \mu_1)}{\sigma^2}(\bar{x} - \frac{1}{2}(\mu_2 - \mu_1))\}$$

Se observa que $\dfrac{f_n(\underline{x}|\mu_2)}{f_n(\underline{x}|\mu_1)}$ depende de $\underline{x}$ sólo a través de la media muestral $\bar{x}$; además crece en función de $\bar{x}$ si $\mu_1 < \mu_2$. Este cociente es monótono con respecto a $\bar{x}$.

Definición 5.3.8 *Se dice que $f_n(\underline{x}/\theta)$ tiene una razón de verosimilitud monótona para un estadístico $g(\underline{x})$ si y sólo si $\forall\theta_1, \theta_2$ tal que $\theta_1 < \theta_2$, el cociente $\dfrac{f_n(\underline{x}|\theta_2)}{f_n(\underline{x}|\theta_1)}$ depende del vector $\underline{x}$ a través de la función $g(\underline{x})$ y el cociente es una función creciente de $g(\underline{x})$ $\forall\underline{x}$.*

En el ejemplo anterior $f_n(\underline{x}|\mu)$ tiene una razón de verosimilitud monótona en $\underline{x}$.

Veamos el ejemplo de una Bernoulli de parámetro p, tomando $y = \sum x_i$, $f_n(\underline{x}|p) = p^y(1-p)^{n-y}$. Si $0 < p_1 < p_2 < 1$

$$\frac{f_n(\underline{x}|p_2)}{f_n(\underline{x}|p_1)} = \frac{(p_2(1-p_1))^y}{(p_1(1-p_2))^y}\frac{(1-p_2)^n}{(1-p_1)^n}$$

cociente que depende de $\underline{x}$ a través de y. Además es una función creciente de y y por lo tanto tiene una razón de verosimilitud monótona en $\sum x_i$.

Definición 5.3.9 *Un test sobre las hipótesis $H_o : \theta \leq \theta_o$ contra $H_1 : \theta > \theta_o$, se dice unilateral y un test sobre las hipótesis $H_o : \theta = \theta_o$ contra $H_1 : \theta \neq \theta_o$, se dice bilateral.*

Vamos a mostrar que si $f_n(\underline{x}|\theta)$ tiene una razón de verosimilitud monótona en algún estadístico T, entonces existe un test U.M.P. para las hipótesis $H_o : \theta \leq \theta_o$ contra $H_1 : \theta > \theta_o$

Teorema 5.3.10 *Si $f_n(\underline{x}/\theta)$ tiene una razón de verosimilitud monótona en el estadístico T y si c es la constante tal que $\mathbb{P}(T \geq c|\theta = \theta_o) = \alpha_o$, entonces la regla de decisión que permite rechazar la hipótesis nula si $T \geq c$ es un test U.M.P. para $H_o : \theta \leq \theta_o$ contra $H_1 : \theta > \theta_o$ al nivel de significación α_o.*

Demostración Sea θ_1 tal que $\theta_1 > \theta_o$, $\alpha(\delta) = \mathbb{P}(\text{ rechazar } H_o|\theta = \theta_o) = \pi_\delta(\theta_o)$
y $\beta(\delta) = \mathbb{P}(\text{ aceptar } H_o|\theta = \theta_1) = 1 - \pi_\delta(\theta_1)$.

Del lema de Neyman-Pearson, se deduce que entre todos los procedimientos δ de error de tipo I $\alpha(\delta) < \alpha_o$, el valor de $\beta(\delta)$ será mínimo para el procedimiento δ^* que consiste en rechazar H_o cuando $\frac{f_n(\underline{x}|\theta_1)}{f_n(\underline{x}|\theta_o)} \geq k$, k siendo elegido de tal forma que $\mathbb{P}(\text{rechaza } H_o|\theta = \theta_o) \leq \alpha_o$.

Como $\frac{f_n(\underline{x}|\theta_1)}{f_n(\underline{x}|\theta_o)}$ es una función creciente de T, un procedimiento, que rechaza H_o cuando el cociente es al menos igual a k, es equivalente al procedimiento que rechaza H_o cuando T es al menos igual a una constante c.

La constante c se elige de tal forma que $\mathbb{P}(\text{rechazar } H_o/\theta = \theta_o) \leq \alpha_o$.

Ahora bien esto es cierto para todo $\theta_1 > \theta_o$ y por lo tanto este procedimiento es U. M. P. para $H_o : \theta = \theta_o$ contra $H_1 : \theta > \theta_o$.

Por otro lado, la función de potencia es no decreciente en θ y por lo tanto si $\pi(\theta_o|\delta) \leq \alpha_o$, entonces $\pi(\theta|\delta) \leq \alpha_o \ \forall\theta \leq \theta_o$. ■

Cuando $f_n(\underline{x}|\theta)$ no tiene una razón de verosimilitud monótona, el test de razón de verosimilitud permite resolver una gran cantidad de problemas:

Si $H_o : \theta \in \Omega_o$ contra $H_1 : \theta \in \Omega_1$, se define

$$\lambda(\underline{x}) = \frac{Sup f_n(\underline{x}|\theta \in \Omega_1)}{Sup f_n(\underline{x}|\theta \in \Omega_o)}$$

El test de razón de verosimilitud consiste en rechazar H_o si $\lambda(\underline{x}) > k$ y en no rechazar H_o si $\lambda(\underline{x}) < k$.

El problema es encontrar entonces la distribución de $\lambda(\underline{x})$. El siguiente teorema da una solución aproximada a este problema.

Teorema 5.3.11 *Si θ es un vector de parámetros de dimensión r y si la hipótesis nula es función lineal de θ, $H_o : H\theta = 0$, en que $H \in \mathcal{M}_{sxr}$, entonces $-2ln\lambda(\underline{x})$ tiene una distribución asintótica χ^2_s.*

5.4 TESTS PARAMÉTRICOS USUALES

Veamos algunos tests usuales que se basan en los resultados anteriores.

5.4.1 Test sobre una media con la varianza conocida

Sea una v.a. $X \sim \mathcal{N}(\mu, \sigma^2)$ en que la varianza σ^2 es conocida e igual a 36^2 y una muestra aleatoria de tamaño $n = 9$.

Sean las hipótesis $H_o : \mu = 180$ contra $H_1 : \mu > 180$ y un nivel de significación de $\alpha = 0.05$.

De lo anterior, se deduce que la región crítica más potente es de la forma $\mathcal{R} = \{\bar{x} > c\}$ con c determinado por:

$$I\!P(\bar{x} \geq c | \mu = 180) = 0.05$$

Como $\bar{x} \sim \mathcal{N}(\mu, 144)$, $(\bar{x} - \mu)/12 \sim \mathcal{N}(0, 1)$.

Si la hipótesis nula $H_o : \mu = 180$ es cierta $(\bar{x} - 180)/12 \sim \mathcal{N}(0, 1)$. Luego la constante c se determina de:

$$I\!P(\bar{x} - 180)/12 \geq (c - 180)/12) = 0.05$$

Utilizando las tablas estadísticas se obtiene $(c - 180)/12 = 1.65$. Finalmente $c = 200$.

La región crítica $\{\bar{x} \geq 200\}$ es U. M. P. para todo $\mu > 180$ de la hipótesis alternativa.

El error de tipo II depende de μ. Como lo muestra la tabla 5.1 y el gráfico 5.2 , el error de tipo II aumenta cuando el valor de μ es muy cercano al valor 180 de H_o.

μ	180	185	190	200	210	220	230
$\pi(\mu)$	0.05	0.11	0.20	0.50	0.80	0.95	0.994
$1 - \pi(\mu)$	0.95	0.89	0.80	0.50	0.20	0.05	0.006

Tabla 5.1: Potencia y error de tipo II para $H_1 : \mu > 180$

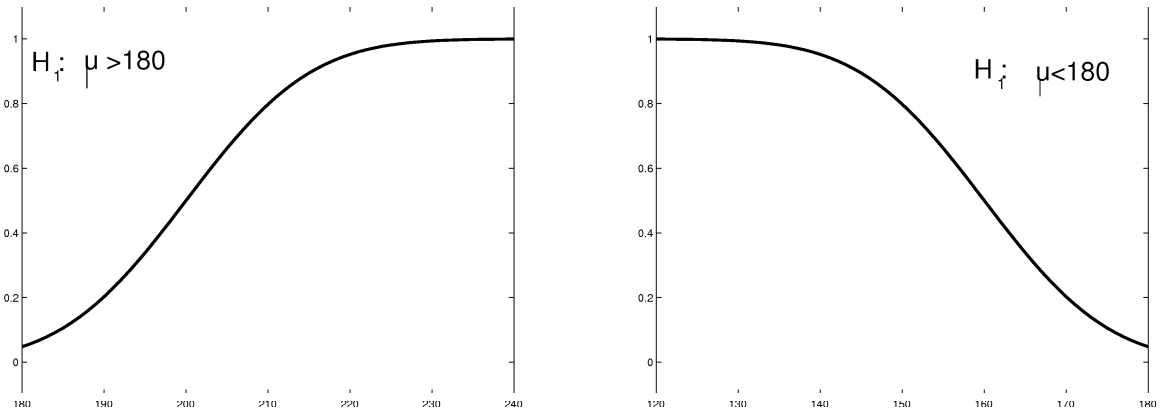


Figura 5.2: Función de Potencia para $H_1 : \mu > 180$ y para $H_1 : \mu < 180$

Sea ahora $H_o : \mu = 180$ contra $H_1 : \mu < 180$ con un nivel de significación de 0.05.

La región crítica más potente es de la forma $\mathcal{R} = \{\bar{x} \leq c\}$ con c determinado por:

$$I\!P(\bar{x} \leq c|\mu = 180) = 0.05$$

La región crítica $\{\bar{x} \leq 160\}$ es U. M. P. para todo $\mu < 180$ de la hipótesis alternativa. La función de potencia esta dada en la tabla 5.2 y el gráfico 5.2.

μ	180	175	170	160	150	140	130
$\pi(\mu)$	0.05	0.11	0.20	0.50	0.80	0.95	0.99
$1 - \pi(\mu)$	0.95	0.89	0.80	0.50	0.20	0.05	0.006

Tabla 5.2: Potencia y error de tipo II para $H_1 : \mu < 180$

Sea finalmente $H_o : \mu = 180$ contra $H_1 : \mu \neq 180$ con un nivel de significación de 0.05.

No existe un test U. M. P. para est hipótesis alternativa; se propone como región crítica

$$\mathcal{R} = \{\bar{x} \leq a\} \cup \{\bar{x} \geq b\}$$

de tal forma que $I\!P(\bar{x} \leq a) = 0.025$ y $I\!P(\bar{x} \geq b) = 0.025$. Obtenemos $a = 156.5$ y $b = 203.5$, que produce una función de potencia presentada en la tabla 5.3 y el gráfico 5.3. Se nota que la potencia es siempre inferior o igual a la potencia de la tabla 5.1 ó 5.2 para todo μ.

μ	140	150	160	170	175	180	185	190	200	210	220
$\pi(\mu)$	0.91	0.69	0.37	0.12	0.07	0.05	0.07	0.12	0.37	0.69	0.91
$1 - \pi(\mu)$	0.09	0.31	0.43	0.88	0.93	0.95	0.93	0.88	0.43	0.31	0.09

Tabla 5.3: Función de Potencia para $H_1 : \mu \neq 180$

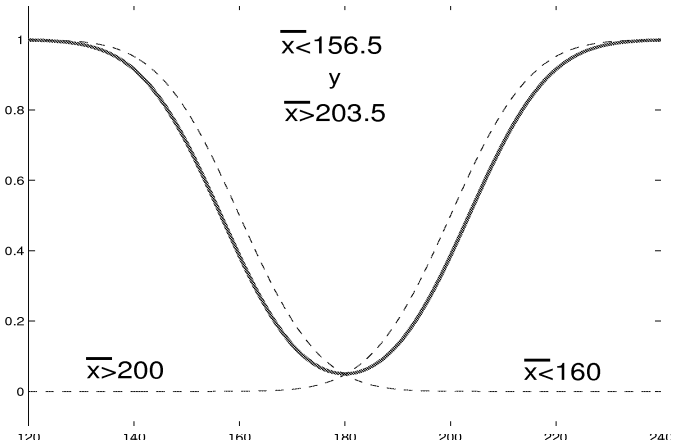


Figura 5.3: Función de Potencia para $H_1 : \mu \neq 180$

Se observara que este test se basa en el supuesto de distribución normal de los valores muestrales. Cuando el tamaño de la muestra es grande, este supuesto es aceptable, pero para muestras pequeñas, es importante comprobar si lo es.

5.4.2 Test sobre una media con la varianza desconocida

Si en el problema anterior suponemos que la varianza es desconocida, se procede de manera parecida al caso de la varianza conocida utilizando la distribución de t Student de la variable $\frac{(\bar{x}-\mu)}{s_n/\sqrt{n-1}}$ que es una Student a $n-1$ grados de libertad.

5.4.3 Test sobre una varianza

Si ahora planteamos las hipótesis:

$$H_o : \sigma^2 \geq \sigma_o^2 \qquad contra \qquad H_1 : \sigma^2 < \sigma_o^2,$$

en donde σ_o^2 es un escalar positivo dado.

A partir del estadístico $\frac{ns_n^2}{\sigma_o^2}$, que sigue una distribución de χ^2 a $n-1$ grados de libertad bajo H_o, se construye la región critica de nivel de significación α:

$$I\!P(\frac{ns_n^2}{\sigma_o^2} \leq c) = \alpha$$

5.4.4 Test de comparación de dos medias

Frecuentemente se está interesado en hacer inferencia sobre la diferencia de dos medias. Por ejemplo, la diferencia de sueldos promedios μ_1 y μ_2 entre dos poblaciones Ω_1 y Ω_2 o la eficiencia de un tratamiento comparando los resultados de la población que utilizó el tratamiento con la que no lo utilizó. Las hipótesis se escriben entonces:

$$H_o : \mu_1 - \mu_2 = d_o$$

$$H_1 : \mu_1 - \mu_2 \neq d_o$$

Es usual tomar $d_o = 0$ y la hipótesis alternativa H_1 puede ser

$$H_1 : \mu_1 - \mu_2 \neq 0 \qquad o \qquad H_1 : \mu_1 - \mu_2 > 0$$

Definamos la v.a. $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ y X $\sim \mathcal{N}(\mu_2, \sigma_2^2)$ como los sueldos de los individuos en las poblaciones Ω_1 y Ω_2 respectivamente. Si la media muestral de X obtenida sobre una muestra de tamaño n_1 en Ω_1 es $\bar{x}_1$ y la media muestral obtenida sobre una muestra de tamaño n_2 en Ω_2 es $\bar{x}_2$, entonces

$$\bar{x}_1 - \bar{x}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$$

Si las varianzas σ_1^2 y σ_2^2 son conocidas, entonces se obtiene una región crítica de nivel de significación $\alpha = 0.05$ para $H_o : \mu_1 = \mu_2$ contra $H_1 : \mu_1 > \mu_2$:

$$I\!P(\bar{x}_1 - \bar{x}_2 \geq 1.96\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2})$$

Si las varianzas son desconocidas, pero se suponen iguales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), entonces se estima esta varianza y se usa un estadístico que sigue una distribución t de Student. Si s_1^2 y s_2^2 son las varianzas empíricas sesgadas para σ_1^2 y σ_2^2 respectivamente, el estimador de σ^2 es:

$$s^2 = (n_1 s_1^2 + n_2 s_2^2)/(n_1 + n_2 - 2)$$

es insesgado para σ^2. Entonces

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2})(\frac{n_1+n_2}{n_1 n_2})}}$$

es una Student a $n_1 + n_2 - 2$ grados de libertad.

La región crítica se define entonces como:

$$I\!P(\bar{x}_1 - \bar{x}_2 \geq t_\alpha \sqrt{(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2})(\frac{n_1 + n_2}{n_1 n_2})})$$

en donde t_α es tal que $I\!P(t_{n_1+n_2-2} \geq t_\alpha) = \alpha$.

Aquí se hizo el supuesto de la igualdad de las varianzas y de la independencia de las dos muestras.

5.4.5 Test para pares de observaciones

Hay situaciones en donde las muestras no son independientes. Es el caso cuando se toman muestras formadas de pares, es decir cuando cada observación de una muestra está relacionada a una observación de la otra muestra. Por ejemplo, se mide la eficacia de un tratamiento comparando un exámen médico sobre la misma población pero antes y después del tratamiento, o bien se considera la diferencia de edad en un grupo de parejas donde una muestra esta formada por las esposas y la otra muestra por sus correspondientes maridos. La dos muestras no son independientes y son del mismo tamaño. Sean (X, Y) las v.a. edades de la mujer y su marido y una muestra de n matrimonios $\{(x_i, y_i), i = 1, 2, ..., n\}$. La diferencia entre las medias empíricas $\bar{x}_n$ e $\bar{y}_n$ es un estimador insesgado de la diferencia $\mu_1 - \mu_2$ entre las dos poblaciones apareadas:

$$E(\bar{x}_n - \bar{y}_n) = E(X - Y) = E(X) - E(Y) = \mu_1 - \mu_2$$

Pero debido a la dependencia entre X e Y la varianza de la diferencia X-Y cambia.

$$\sigma^2_{X-Y} = E(X - Y - (\mu_1 - \mu_2))^2) = E(X - \mu_1)^2 + E(Y - \mu_2)^2 - 2E(X - \mu_1)(Y - \mu_2)$$

$$\sigma^2_{X-Y} = \sigma^2_1 + \sigma^2_2 - 2Cov(X, Y)$$

Como no se conoce en general las varianzas σ^2_1, σ^2_2 y la covarianza $Cov(X, Y)$, se estima la varianza de la diferencia σ^2_{X-Y} considerando que los valores muestrales son las diferencias $d_i = x_i - y_i$ que provienen de una sola muestra:

$$\hat{\sigma}^2_{X-Y} = \frac{\sum (d_i - d_n)^2}{n}$$

en donde $\bar{d}_n = \frac{\sum d_i}{n} = \frac{\sum x_i - y_i}{n}$. Se puede escribir:

$$\hat{\sigma}^2_{X-Y} = \frac{\sum (x_i - y_i)^2}{n} - \bar{d}_n^2$$

El estimador de la varianza de $\bar{d}_n$ es igual entonces a $\frac{\hat{\sigma}^2_{X-Y}}{n}$ y $\frac{\bar{x}_n - \bar{y}_n - (\mu_1 - \mu_2)}{\hat{\sigma}_{X-Y}/\sqrt{n-1}}$ sigue una t de Student a $n - 1$ grados de libertad.

5.4.6 Test de comparación de dos varianzas: la distribución F

Se quiere comparar las varianzas σ_1^2 y σ_2^2 de dos poblaciones normales a partir de muestras aleatorias independientes de cada población. Si $x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_m$ son las muestras aleatorias respectivas, $s_1^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ y $s_2^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$ son las varianzas muestrales, $U = ns_1^2/\sigma_1^2 \sim \chi^2_{n-1}$ y $V = ms_2^2/\sigma_2^2 \sim \chi^2_{m-1}$; además U y V son independientes.

Vimos en el capítulo anterior que $\dfrac{U/(n-1)}{V/(m-1)}$ sigue una distribución F de Fisher a $n-1$ y $m-1$ grados de libertad.

Consideremos entonces el estadístico

$$\frac{ns_1^2/(n-1)}{ms_2^2/(m-1)}$$

que sigue una distribución $F_{n-1,m-1}$ bajo la hipótesis nula $H_o : \sigma_1 = \sigma_2$.

Se define la región crítica de nivel de significación α para $H_o : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 > \sigma_2^2$ como:

$$I\!P(\frac{nsS_1^2/(n-1)}{ms_2^2/(m-1)} > F_\alpha) = \alpha$$

en donde F_α se calcula a partir de la F de Fisher a $n-1$ y $m-1$ g.l.

5.5 TESTS χ^2

Se habla de test χ^2 a todo test que use un estadístico que sigue una distribución χ^2. Aquí trataremos los tests χ^2 obtenidos a partir de una distribución multinomial. Veremos previamente dos distribuciones de vectores aleatorios, la distribución normal multivariada y la distribución multinomial que tiene un comportamiento asintótico de vector normal multivariado. Después de presentar un test para un modelo multinomial, veremos aplicaciones del test para hipótesis no paramétricas.

5.5.1 La distribución normal multivariada

Se puede definir de dos maneras la distribución normal multivariada.

Sea $X = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ X_p \end{pmatrix}$ un vector aleatorio de $I\!R^p$.

Definición 5.5.1 *Sea u: $I\!R^p \longrightarrow I\!R$*
Se dice que X es un vector normal multivariado de orden p de vector de media μ y de matriz de varianza-covarianza Γ (se denota $X \sim \mathcal{N}_p(\mu, \Gamma)$) si y sólo sí $u(X) \sim \mathcal{N}(u(\mu), \Gamma(u,u))$.

Es decir que si X es un vector normal, toda combinación lineal de X es una v.a. normal.

Definición 5.5.2 *Se dice que $X \sim \mathcal{N}_p(\mu, \Gamma)$ si su función característica es*

$$\Psi_X(u) = exp(iu^t\mu - \frac{1}{2}u^t\Gamma u) \qquad \forall u \in I\!R^p$$

Propiedades:

- Tomando como vector u los vectores canónicos, se obtienen las leyes marginales de X, que son normales; pero la recíproca es falsa: un vector formado de variables normales no es necesariamente un vector normal: En efecto Γ tiene que ser semidefinida positiva ya que $\Gamma(u) = u^t \Gamma u$ es la varianza de $u^t X$.

- Sea Y una matriz $(p \times q)$.

$$X \sim \mathcal{N}_p(\mu, \Gamma) \Longrightarrow Y = AX \sim \mathcal{N}_q(A\mu, A\Gamma A^t).$$

- Las v.a. X_i son independientes $\Longleftrightarrow \Gamma$ es diagonal

- Γ es semidefinida positiva
  En efecto $\Gamma(u, u) = u^t \Gamma u$ es la varianza de la v.a. $u(X) = u^t X$.

- Si Γ es de rango r, existe Λ una matriz $(r \times p)$ tal que $\Gamma = \Lambda\Lambda^t$. Entonces:

$$X \sim \mathcal{N}_p(\mu, \Gamma) \Longleftrightarrow X = \mu + \Lambda Y \quad Y \sim \mathcal{N}_r(0, I_r)$$

  es decir que las componentes del vector Y son centradas, normalizadas y independientes entre sí.

- Si Γ es invertible, Λ es invertible también e $Y = \Lambda^{-1}(X - \mu)$.

Este último resultado permite calcular la densidad del vector X. En efecto se puede calcular la densidad del vector $Y \sim \mathcal{N}_p(0, I_p)$:

$$f(Y) = \prod f(Y_i) = (\frac{1}{2\pi})^{p/2} exp(-\frac{1}{2}\sum Y_i^2)) = (1/2\pi)^{p/2} exp(-\frac{1}{2} Y Y^t)$$

Como $YY^t = (\Lambda^{-1}(X - \mu)^t \Lambda^{-1}(X - \mu) = (X - \mu)^t \Gamma^{-1}(X - \mu)$, el Jacobiano de la transformación es $|\Gamma|^{-1/2}$, luego la densidad de X es:

$$h(X) = (\frac{|\Gamma|^{-1/2}}{2\pi})^{p/2} exp(-\frac{1}{2}(X - \mu)^t \Gamma^{-1}(X - \mu))$$

Proposición 5.5.3 *Si el vector aleatorio $X \sim \mathcal{N}(\mu, \Gamma)$ con Γ de rango r, entonces $\| X - \mu \|^2_{\Gamma^{-1}} \sim \chi^2_r$.*

Demostración Acordamos que si $Y \sim \mathcal{N}(0, I_r)$, $\| Y \|^2 = \sum Y_i^2 \sim \chi^2_r$. Como $\Gamma = \Lambda\Lambda^t$, existe Y tal que $X = \mu + \Lambda Y$, con $Y \sim \mathcal{N}(0, I_r)$. Pero se puede escribir $Y = (\Lambda^t\Lambda)^{-1}\Lambda^t(X - \mu)$, luego:

$$\| Y \|^2_{I_p} = YY^t = \| X - \mu \|^2_{\Gamma^{-1}} \sim \chi^2_r$$

■

5.5.2 La distribución multinomial

La distribución multinomial es una generalización de la distribución binomial. En vez de tener dos alternativas en cada experimento, se tienen k alternativas ($k \geq 2$). Por ejemplo, hay seis resultados posibles cuando se tira un dado. Si el "1" tiene probabilidad p_1, el "2" tiene probabilidad p_2,...,el "6" tiene probabilidad p_6, y si hacemos n lanzamientos independientes, los números M_1 de "1", M_2 de "2",..., M_6 de "6" constituyen un vector aleatorio M con una distribución multinonimal de parámetros $n, p_1, p_2, ..., p_6$. Se observa que $\sum M_i = n$ y

$$I\!P(M = m) = I\!P(M_1 = m_1, ..., M_6 = m_6) = \frac{n! p_1^{m_1} p_2^{m_2} ... p_6^{m_6}}{m_1! m_2! ... m_6!}$$

Calculemos la esperanza y la varianza de M.

Si $p = \begin{pmatrix} p_1 \\ p_2 \\ . \\ p_6 \end{pmatrix}$, entonces $E(M) = np$.

Sea el resultado J_i del lanzamiento i: $J_i = e_h$, el h-ésimo vector canónico si el resultado es h. Entonces $M = \sum J_i$.

$E(J_i) = p$ y $E(J_i J_i^t) = \sum_h e_h e_h^t I\!P(J_i = e_h) = \begin{pmatrix} p_1 & 0 & . & 0 \\ 0 & p_2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & p_6 \end{pmatrix} = Diag(p)$

$Var(J_i) = E(J_i J_i^t) - E(J_i)[E(J_i)]^t = Diag(p) - pp^t = \Sigma(p)$

Luego $Var(M) = n\Sigma(p)$.

Por el Teorema del Límite Central, se tiene:

$$\lim_{n \longrightarrow +\infty} I\!P(\frac{M - np}{\sqrt{n}} \leq x) = \Phi(x),$$

en donde Φ es la función de distribución normal multivariada centrada de matriz de covarianza $\Sigma(p)$.

Ejercicio: Muestre que si el vector multinomial es de dimensión k, entonces el rango de la matriz $\Sigma(p)$ es igual a $k - 1$ (Se puede mostrar que el núcleo de $\Sigma(p)$ es de dimensión 1).

Proposición 5.5.4 *Si M es un vector de distribución multinomial $\mathcal{M}(n, p_1, ..., p_k)$, entonces*

$$Q = \sum \frac{(M_i - np_i)^2}{np_i}$$

tiene una distribución asintótica de χ^2_{k-1}.

La demostración se basa en el resultado del ejercicio anterior.

5.5.3 Test de ajuste para un modelo multinomial

Sea un dado que se tira $n = 102$ veces. Se obtiene entonces la distribución empírica (tabla 5.4).

¿Podemos concluir que el dado esta cargado?

Sea la hipótesis nula $H_o : p_i = \frac{1}{6} \quad \forall i$ que el dado es equilibrado.

Entonces, si la hipótesis nula es cierta, $M \sim \mathcal{M}(102, 1/6, ..., 1/6)$.

Calculamos entonces el estadístico Q para construir la región crítica (tabla 5.5).

i	1	2	3	4	5	6	Total
M_i	12	11	22	20	16	21	102

Tabla 5.4: Resultados del experimento

i	M_i	np_i	$M_i - np_i$	$(M_i - np_i)^2/np_i$
1	12	17	-5	1.471
2	11	17	-6	2.118
3	22	17	5	1.471
4	20	17	3	0.529
5	16	17	-1	0.059
6	21	17	4	0.941
Total	102	102	0	6.589

Tabla 5.5: Calculo de Q

Se obtiene $Q = 6.589$, y $\mathbb{P}(\chi^2_5 > 6.589) > 5\%$, por lo cual no se rechaza H_o. Lo que quiere decir que entre los valores observados y los valores esperados las diferencias no son suficientemente significativas como para decidir que el dado esta cargado.

5.5.4 Test de ajuste para una distribución discreta

Se considera el número de accidentes X observados cada fin de semana en una carretera peligrosa (tabla 5.6). Se quiere probar la hipótesis que X sigue una distribución de Poisson de parámetro λ a partir de datos obtenidos sobre un año. En primera instancia supondremos λ conocido e igual a 1.5. Se tiene entonces $H_o : X \sim \mathcal{P}(1.5)$.

N^o accidentes	0	1	2	3	4	5	6 y más	Total
N^o semanas	17	16	10	5	2	1	1	52

Tabla 5.6: Resultados del experimento

Bajo H_o, los números de semanas M_o con "0" accidente, M_1 con "1" accidente, ..., M_6 con "6" o más accidentes sigue una distribución multinomial de parámetros $n = 52$, y $p_o = \mathbb{P}(X = 0)$, $p_1 = \mathbb{P}(X = 1)$, ..., $p_6 = \mathbb{P}(X \geq 6)$.

Calculemos los $p_i = \mathbb{P}(X = i)$ bajo el supuesto que $X \sim \mathcal{P}(1.5)$ y comparemos los valores np_i con los valores observados M_i (tabla 5.7).

i	M_i	np_i	$M_i - np_i$	$(M_i - np_i)^2/np_i$
0	17	11.60	5.400	2.5124
1	16	17.40	0.596	0.0204
2	10	13.05	-3.052	0.7137
3	5	6.53	-1.526	0.3568
4	2	2.45	-0.449	0.0824
5	1	0.73	0.267	0.0971
6	1	0.24	0.766	3.2735
Total	52	52	0	7.0563

Tabla 5.7: Calculo de Q

Obtienemos $Q = 7.0563$, y $I\!P(\chi^2_6 > 7.0563) > 5\%$, por lo cual no se rechaza H_o.

Ahora si se supone que no se conoce el parámetro λ, se puede estimar por $\hat{\lambda} = \bar{X}_n = \sum iM_i/52 = 72/52 = 1.385$ y proceder como antes. Pero ahora el estadístico Q pierde un grado de libertad debido a la estimación.

Con el parámetro $\hat{\lambda}$, Q=5.62 y $I\!P(\chi^2_5 > 5.62) > 5\%$.

5.5.5 Test de ajuste para una distribución continua

Para construir un test χ^2 para una hipótesis relativa a una distribución continua, por ejemplo $H_o : X \sim \mathcal{N}(1, 0.25)$, basta transformar la variable en una variable discreta. Se divide el rango de X en k intervalos disjuntos I_1, I_2, ...,I_k y se cuentan los números de observaciones de la muestra M_i que caen en el intervalo I_i. El vector M formado de los M_i sigue una distribución multinomial de parámetros de probabilidad determinados por la hipótesis nula.

Sea por ejemplo, las temperaturas medias X del mes de septiembre en la Urbe durante 60 años (tabla 5.8). Se quiere probar la hipótesis nula $H_o : X \sim normal$.

Hay diferentes maneras de definir la partición de intervalos de $I\!R$. Una vez fijado el número de intervalos, se pueden elegir del mismo largo o de la misma probabilidad. Tomaremos aquí 10 intervalos equiprobables.

Para calcular las probabilidades, hay que estimar previamente los parámetros μ y σ^2 de la normal:

$$\hat{\mu} = \bar{x}_n = 15.76 \quad \hat{\sigma}^2 = s^2_n = 13.82$$

Luego los intervalos I_j se obtienen (tabla 5.9) de tal forma que:

$$I\!P(X \in I_j) = 0.10 \quad \forall j$$

en donde $X \sim \mathcal{N}(15.76, 13.82)$.

Evaluando se obtiene $Q = 9.35$. El estadístico χ^2 tiene aquí 7 grados de libertad. (Se estimaron dos parámetros). Como $I\!P(\chi^2_7 > 9.35) > 5\%$, no se rechaza la hipótesis de normalidad.

5.2	6.5	7.5	8.2	10.1	10.5	11.6	12.0	12.0	12.8	13.5	13.8
13.9	14.0	14.0	14.2	14.3	14.5	14.7	14.8	15.0	15.0	15.2	15.2
15.3	15.4	15.6	15.8	15.8	15.9	16.0	16.1	16.2	16.4	16.4	16.5
16.5	16.8	16.9	17.0	17.0	17.1	17.1	17.1	17.4	17.6	17.9	18.2
18.5	18.8	18.9	19.4	19.8	20.3	20.9	21.4	21.9	22.5	22.8	23.9

Tabla 5.8: Temperaturas medias

I_i	M_i	np_i	$M_i - np_i$	$(M_i - np_i)^2/np_i$
$]-\infty$,10.96]	6	6	0	0.00
]10.96,12.64]	3	6	-3	1.50
]12.64,13.83]	3	6	-3	1.50
]13.83,14.83]	8	6	2	0.67
]14.83,15.76]	7	6	1	0.17
]15.76,16.69]	10	6	4	2.67
]16.69,17.69]	9	6	3	1.50
]17.69,18.88]	4	6	-2	0.67
]18.88,20.56]	4	6	-2	0.67
]20.56,$+\infty$]	6	6	0	0.00
Total	60	60	0	9.35

Tabla 5.9: Calculo de Q

5.5.6 Test de independencia para 2 variables nominales

Cuando dos v.a. discretas con valores en conjuntos A y B respectivamente son independientes:

$$I\!P(X = i \ \ e \ \ Y = j) = I\!P(X = i)I\!P(Y = j) \quad \forall (i, j) \in A \times B$$

Si A y B son conjuntos finitos ($card(A) = p$, $card(B) = q$), las frecuencias M_{ij} de observaciones obtenidas en una muestra bivariada de tamaño n siguen una distribución multinomial de parámetros n, p en donde p es el vector de las probabilidades $p_{ij} = I\!P(X = i \ \ e \ \ Y = j)$.

Bajo la hipótesis de independencia entre X e Y, se puede estimar estos parámetros p_{ij} a partir de las frecuencias marginales de X e Y:

$$\hat{p}_{ij} = \hat{p}_{i\bullet}\hat{p}_{\bullet j}$$

con $\hat{p}_{i\bullet} = \sum_j M_{ij}/n$ y $\hat{p}_{\bullet j} = \sum_i M_{ij}/n$.

Lo que permite usar el estadístico

$$Q = \sum_{ij} \frac{(M_{ij} - n\hat{p}_{i\bullet}\hat{p}_{\bullet j})^2}{n\hat{p}_{i\bullet}\hat{p}_{\bullet j}}$$

que sigue una distribución asintótica χ^2 a $(p-1)(q-1)$ grados de libertad.

Sea un conjunto de consumidores que dan su apreciación sobre un chocolate. Se quiere estudiar si existe una relación entre la opinión de los consumidores y su nivel socio-económico (NSE).

Se considera la tabla de contingencia obtenida a partir de une encuesta de estudio de mercado sobre 1600 consumidores (tabla 5.10), que presenta las frecuencias M_{ij} para cada NSE i y apreciación j.

NSE	APRECIACION			TOTAL
	MALA	REGULAR	BUENA	
A	140	100	45	285
B	50	225	350	625
C	15	175	500	690
TOTAL	205	500	895	1600

Tabla 5.10: Tabla de contingencia

Las probabilidades p_{ij} se estiman usando las frecuencias marginales de la tabla; por ejemplo, para el NSE A con la apreciación MALA se obtiene $\hat{p}_{11} = 285 \times 205/1600 = 0.0228$ y $n\hat{p}_{11} = 36.51$.

Se obtiene el valor $Q = 521.46$. Como $I\!P(\chi_4 > 521.46) < 5\%$, por tanto, se rechaza la hipótesis de independencia entre el NSE y la apreciación.

Nota: Se puede usar el mismo test para probar la independencia de dos variables continuas transformándolas en variables discretas.

5.6 EJERCICOS

1. El cocinero del casino preparó la masa para hacer 500 empanadas. Ese mismo día, en un grupo de 20 alumnos que almorzaron juntos, alguien propuso contar la cantidad de pasas que cada uno encontrase en su empanada, encontrándose la siguiente distribución:

N^o de pasas	0	1	2	3	4	5	8
N^o de empanadas	1	3	4	5	4	2	1

a) Suponiendo que la distribución de la cantidad de pasas X en una empanada sigue una ley de Poisson, estime el parámetro λ de esta ley.
b) Justifique la hipótesis: H_0: La distribución de la cantidad de pasas en una empanada sigue una ley de Poisson de las dos formas siguientes:
(i) A priori: Buscando la probabilidad de que una empanada tenga exactamente x pasas.
(ii) A posteriori: comparando los resultados esperados bajo la hipótesis con aquellos observados en la muestra.
c) Se decide que las empanadas son *aceptables* si en promedio cada empanada tiene 3.5 pasas; el cocinero afirma que está es la cantidad de pasas por empanadas. Los alumnos, en cambio, objetan que las empanadas tienen en promedio sólo 2.5 pasas.
¿Qué significa la elección de los test de hipótesis siguientes:

$$H_0\text{: } \lambda = 3.5 \text{ vs. } H_1\text{: } \lambda = 2.5 \qquad H_0'\text{: } \lambda = 2.5 \text{ vs } H_1'\text{: } \lambda = 3.5 \text{ ?}$$

d) Dé la región crítica del test H_0 vs. H_1 al nivel de significación $\alpha = 0.05$. Dé la potencia de este test y concluir si las empanadas son *aceptables*.
e) Misma pregunta tomando H_0' vs H_1'.
f) Comparar las dos decisiones anteriores.

2. Se tienen los pesos de diez parejas antes y después de 6 meses de matrimonio:

	antes	72	69	81	71	88	78	68	76	86	95
Hombres	después	77	68.5	85	74.5	90.5	76	71	75	87.5	101
	antes	52	56	61	49	57	63	66	59	67	51
Mujeres	después	54	55	58	50	55	61	64	56	70	50

¿Cuál es la influencia del matrimonio sobre el peso de los hombres y de las mujeres?

3. Se quiere probar si hay una diferencia de ingreso entre hombres y mujeres médicos. Se hizo una encuesta a $n = 200$ médicos seleccionados al azar e independientemente. Se obtuvo la siguiente información:

	Ingresos bajos	Ingresos altos	Total
Hombres	20	100	120
Mujeres	70	10	80
Total	90	110	200

a) Sean p_1 y p_2 las proporciones poblacionales de médicos hombres y mujeres; y sean p'_1 y p'_2 las proporciones poblacionales de médicos con ingresos bajos y altos. Realice los tests

$$H_0 : p'_1 = p_2 \quad vs. \quad H_1 : p'_1 \neq p_2 \qquad H'_0 : p_1 = p'_2 \quad vs. \quad {}'_1 : p_1 \neq p'_2$$

b) Estudie la independencia entre sexo e ingreso.

4. Supóngase que X_1, ..., X_n constituyen una m.a.s. de X con distribución uniforme sobre $|0, \theta|$ y las siguientes hipótesis:

$$H_0 : \theta > 2 \quad vs. \quad H_1 : \theta < 2$$

Sea $Y_n = max\{X_1, ..., X_n\}$ y considérese la región critíca $\{Y_n \leq 1.5\}$.
a) Determínese la función de potencia.
b) Determínese el tamaño del test.

5. Suponga que se desconoce la proporción p de artículos defectuosos en una población de artículos y se desea probar las hipótesis

$$H_0 : p = 0.2 \quad vs \quad H_1 : p \neq 0.2$$

Suponga además que se selecciona una m.a.s. de tamaño 20. Sea Y el número de artículos defectuosos en la muestra y considere un procedimiento que proporciona una región crítica definida por $Y \geq 7$ o $Y \leq 1$.

a) Determine la función de la potencia $\pi(p)$ en los puntos $p \in \{0, 0.1, 0.2, ..., 0.9, 1\}$.
b) Determine el tamaño del test.

6. Sea $x_1, ..., x_n$ una m.a.s. de una distribución normal de media μ desconocida y varianza 1. Sea μ_0 un valor dado. Se tienen las hipótesis

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0$$

Supongamos que el tamaño de la muestra es 25, y considérese que el procedimiento para no rechazar H_0 está dado por $|\bar{X}_n - \mu_0| < c$. Determine el valor de c para que el tamaño del test sea 0.05.

7. Sea $x_1, ..., x_n$ una m.a.s. de una distribución de media θ desconocida y varianza 1, y sean las hipótesis

$$H_0 : \theta = 3.5 \quad vs. \quad H_1 : \theta = 5.0$$

a) Entre los procedimientos para resolver el test anterior tal que $\beta(\delta) \leq 0.05$, descríbase un procedimiento para el que $\alpha(\delta)$ sea un mínimo.
b) Para $n = 4$, encuéntrese el valor mínimo descrito en a).

8. Supóngase que se selecciona una observación X de una $U(0, \theta)$, donde θ es desconocido y se plantean las siguientes hipótesis:

$$H_0 : \theta = 1 \quad vs \quad H_1 : \theta = 2$$

a) Demuestre que existe un procedimiento para resolver el test para el cual $\alpha(\delta) = 0$ y $\beta(\delta) < 1$.
b) Entre todas las soluciones del test para las cuales $\alpha(\delta) = 0$, hállese una para el cual $\beta(\delta)$ sea mínimo.

9. Sea $x_1, ..., x_n$ una m.a.s. de una $Poisson(\lambda)$, con λ desconocido. Sean λ_0 y λ_1 dados, con $\lambda_1 > \lambda_0 > 0$. Se tienen las siguientes hipótesis:

$$H_0 : \lambda = \lambda_0 \quad vs. \quad H_1 : \lambda = \lambda_1$$

Demuéstrese que el valor de $\alpha(\delta) + \beta(\delta)$ se minimiza por un procedimiento que rechaza H_0 cuando $\bar{X_n} > c$ y encuéntrese el valor de c.

10. Sea $X_1, ..., X_n$ una m.a.s. de una distribución con parámetro θ cuyo valor es desconocido. Supóngase además que se desea contrastar las siguientes hipótesis:

$$H_0 : \theta < \theta_0 \quad vs \quad H_1 : \theta > \theta_0$$

Supóngase además, que el procedimiento que se va a utilizar ignora los valores observados en la muestra y, en vez de ello, depende únicamente de una aleatorización auxiliar en la que se lanza una moneda cargada de manera que se obtendrá cara con probabilidad 0.05 y sello con probabilidad 0.95. Si se obtiene una cara, entonces se rechaza H_0, y si se obtiene sello, no se rechaza H_0. Descríbase la función de potencia de este procedimiento.

11. Sea $x_1, ..., x_n$ una m.a.s. de una distribución con parámetro θ desconocido y una función de densidad conjunta $f_n(x|\theta)$ que tiene cociente de verosimilitud monótona en el estadístico $T = r(X)$. Sea θ_0 un valor específico de θ y supóngase que se quieren contrastar las hipótesis

$$H_0 : \theta > \theta_0 \quad vs \quad H_1 : \theta < \theta_0$$

Sea c una constante tal que $P(T \leq c|\theta = \theta_0) = \alpha_0$. Demostrar que el procedimiento que rechaza H_0 si $T \leq c$ es U.M.P. al nivel α_0.

12. Sea $x_1, ..., x_n$ una m.a.s. de una $Poisson(\lambda)$ con λ desconocido. Se quiere contrastar las hipótesis

$$H_0 : \lambda \geq 1 \quad vs \quad H_1 : \lambda < 1$$

Supóngase además que el tamaño de la muestra es $n = 20$. ¿Para qué niveles de significación α_0, con $0 < \alpha_0 < 0.03$ existen tests UMP?

13. Consideremos una observación X de una distribución de Cauchy con un parámetro de localización desconocido θ, esto es, una distribución cuya función de densidad está dada por:

$$f(x/\theta) = \frac{1}{\pi[1+(x-\theta)^2]} \quad (\forall x)$$

Se desean contrastar las hipótesis

$$H_0 : \theta = 0 \quad vs \quad H_1 : \theta > 0$$

Demuestre que no existe un test UMP de estas hipótesis a ningún nivel de significación α_0.

14. Sea $x_1, ..., x_n$ una m.a.s. de una distribución $\mathcal{N}(\mu, 1)$. Supóngase que se desean contrastar las hipótesis

$$H_0 : \mu \leq 0 \quad vs \quad H_1 : \mu > 0$$

Se denota δ^* el test UMP con nivel de significación $\alpha_0 = 0.025$ y $\pi_{\delta^*}(\mu)$ la función de potencia de δ^*.

15. Sea $x_1, ..., x_n$ una m.a.s. de una distribución $U(0, \theta)$ con θ desconocido. Supongamos que queremos contrastar las hipótesis

$$H_0 : \theta = 3 \quad vs \quad H_1 : \theta \neq 3$$

Considere que H_0 se rechaza si $c_2 \leq max\{x_1, ..., x_n\} \leq c_1$ y sea $\pi(\theta|\delta)$ la función de potencia de δ. Determine los valores de c_1, c_2 para que $\pi(3|\delta) = 0.05$ y δ sea insesgado.

Capítulo 6

ASOCIACIÓN ENTRE DOS VARIABLES

6.1 INTRODUCCIÓN

Generalmente un problema estadístico envolucra más de una variables. En una encuesta de opinión para un estudio de mercado o política se hacen varias preguntas cuyas respuestas son interesantes de relacionar. Por ejemplo, se interroga a los votantes no solamente sobre su candidato preferido, sino que también su edad, género, profesión, etc... El análisis de tal encuesta permitirá eventualmente determinar el perfil del electorado de un candidato, lo que orientaá su campaña electoral. El psicólogo querrá comparar las aptitudes mentales (CI) y el rendimiento escalar de un grupo de estudiantes. Estos problemas llaman a medir y describir relaciones entre variables.

Una asociación entre variables expresa el grado de influencia que puede tener una variable sobre otra. Los índices que se pueden definir dependen del tipo de relación que se estudia y de la naturaleza de las variables consideradas. Se presentan en primer lugar índices descriptivos de asociación y en seguida se hacen inferencia sobre estos coeficientes.

6.2 EL COEFICIENTE DE CORRELACIÓN

Si se consideran dos variables X e Y cuantitativas y mediciones sobre un conjunto de individuos $\mathcal{P}$. con valores en $\mathbb{R}$ ó un intervalo de $\mathbb{R}$, una simple representación gráfica en $\mathbb{R}^2$ con un gráfico de dispersión permitirá detectar la existencia y la forma de una eventual relación entre las dos variables. El coeficiente de correlación lineal es el índice de asociación más usual entre dos variables numéricas. Una inspección del gráfico de dispersión es necesaria para asegurarse de que la interpretación es correcta.

Sea $\{(x_i, y_i)|i = 1, ..., n\}$ una muestra aleatoria bivariada del par (X, Y) de variables. Se denotan $\bar{x} = \frac{1}{n}\sum x_i$ y $\bar{y} = \frac{1}{n}\sum y_i$ a las medias empíricas respectivas de x=$(x_1, x_2, ...x_n)$ e y=$(y_1, y_2, ...y_n)$. y $s_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ y $s_y^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$ a las varianzas empíricas respectivas de x e y.

Definición 6.2.1 *Se llama covarianza empírica entre X e Y a:*

$$cov(x, y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$$

Como la covarianza es sensible a los cambios de escala de las dos variables, se elimina este efecto con el coeficiente de correlación lineal, que toma en cuenta de las varianzas s_x^2 de los valores x y s_y^2 de y.

Definición 6.2.2 *Se llama correlación lineal entre x e y a la cantidad:*

$$r_{x,y} = \frac{cov(x,y)}{s_x s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Este coeficiente, que toma como valores extremos $+1$ y -1, mide el grado de relación de tipo lineal que existe entre x e y.

$r_{x,y} = -1$	relación estrictamente lineal de pendiente negativa
$-1 < r_{x,y} < 0$	tendencia lineal negativa
$r_{x,y} = 0$	ausencia de tendencia lineal
$0 < r_{x,y} < +1$	tendencia lineal positiva
$r_{x,y} = +1$	relación estrictamente lineal de pendiente positiva

La tendencia lineal aumenta cuando $r_{x,y}$ tiende a ± 1 (ver gráficos 6.1). Pero cuando $r_{x,y} \neq \pm 1$, hay muchos casos diferentes que pueden producir el mismo valor del coeficiente $r_{x,y}$. De aquí la importancia de tener cuidado en la interpretación de un coeficiente de correlación por que un dato atípico ó aberrante, una mezcla de poblaciones, una relación no lineal pueden cambiar totalmente el valor del coeficiente (ver gráficos 6.2).

Cuando se estudia en conjunto más de dos variables, se presentan los coeficientes de correlación relativos en una matriz cuyo termino general r_{ij} es el coeficiente de correlación lineal de las variables i y j. La matriz de correlación asociada a los datos de 6 variables recolectados sobre 20 países de América Latina (tabla 6.1), se presentan en la tabla 6.2. Acá, el coeficiente de correlación entre la tasa de natalidad y la fecundidad es igual a 0.972.

Si se quiere estudiar otro tipo de relación, se tiene dos alternativas:

- Dada una función f de X, calcular el coeficiente de correlación entre $f(X)$ e Y. Este método es factible cuando se conoce la función f.
- Usar otros índices, como veremos más adelante.

6.3 LA RAZÓN DE CORRELACIÓN

Cuando una de las dos variables es nominal u ordinal, no se puede calcular el coeficiente de correlación lineal. Si Y es la variable cuantitativa, por ejemplo el PNB de todos países y X la variables nominal, el clima con p categorías o modalidades, cada modalidad de X, es decir cada tipo de clima define un grupo o subpoblación de países y los grupos son disjuntos entre si. Conviene aquí usar notaciones que permitan distinguir los valores de la variable Y según la modalidad que toman las observaciones sobre la variable nominal X. Si n_j observaciones toman la modalidad o categoría j de X, se puede escribir y_{1j},...,$y_{n_j j}$ estas n_j observaciones de Y.

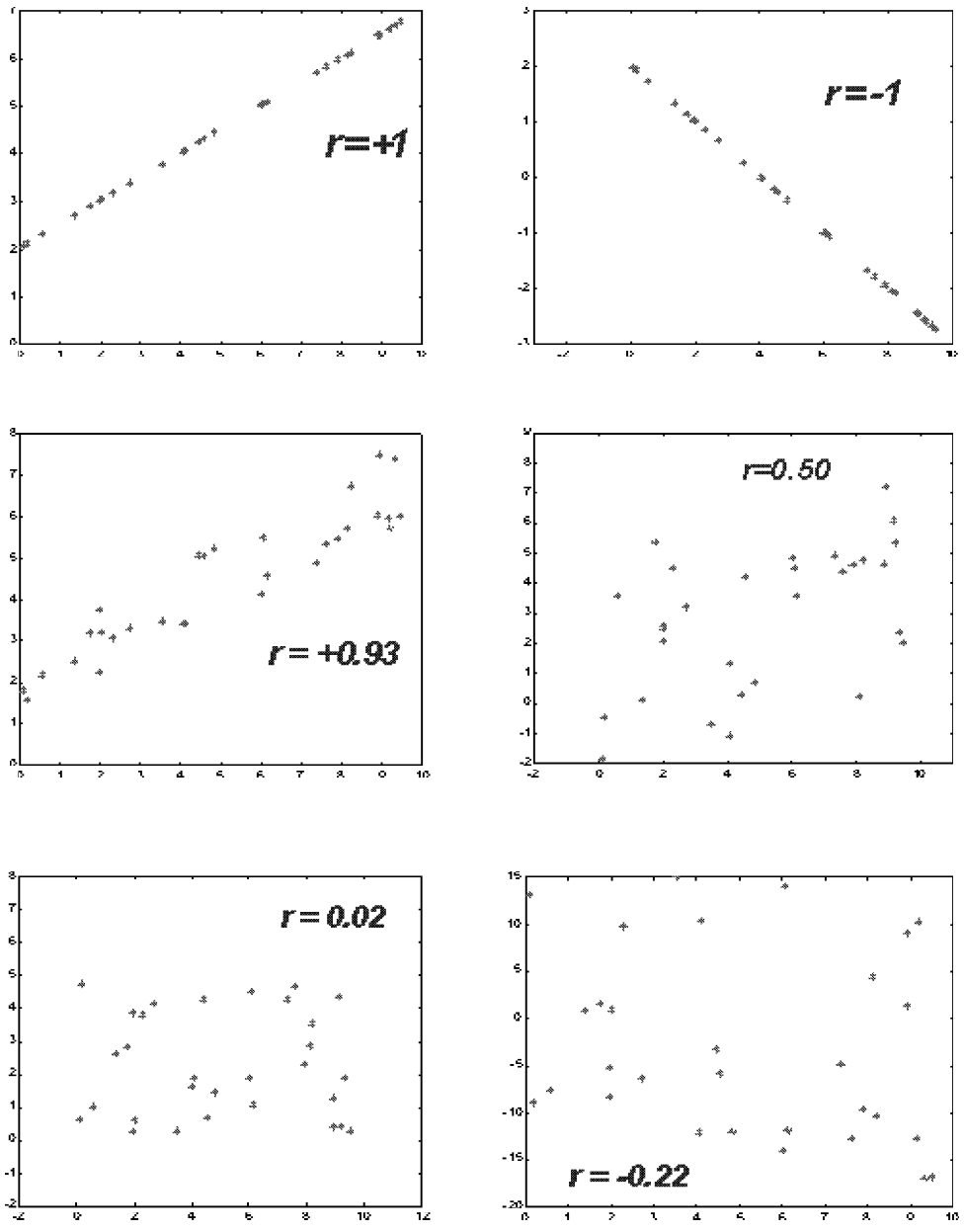


Figura 6.1: Gráfico y coeficiente de correlación lineal

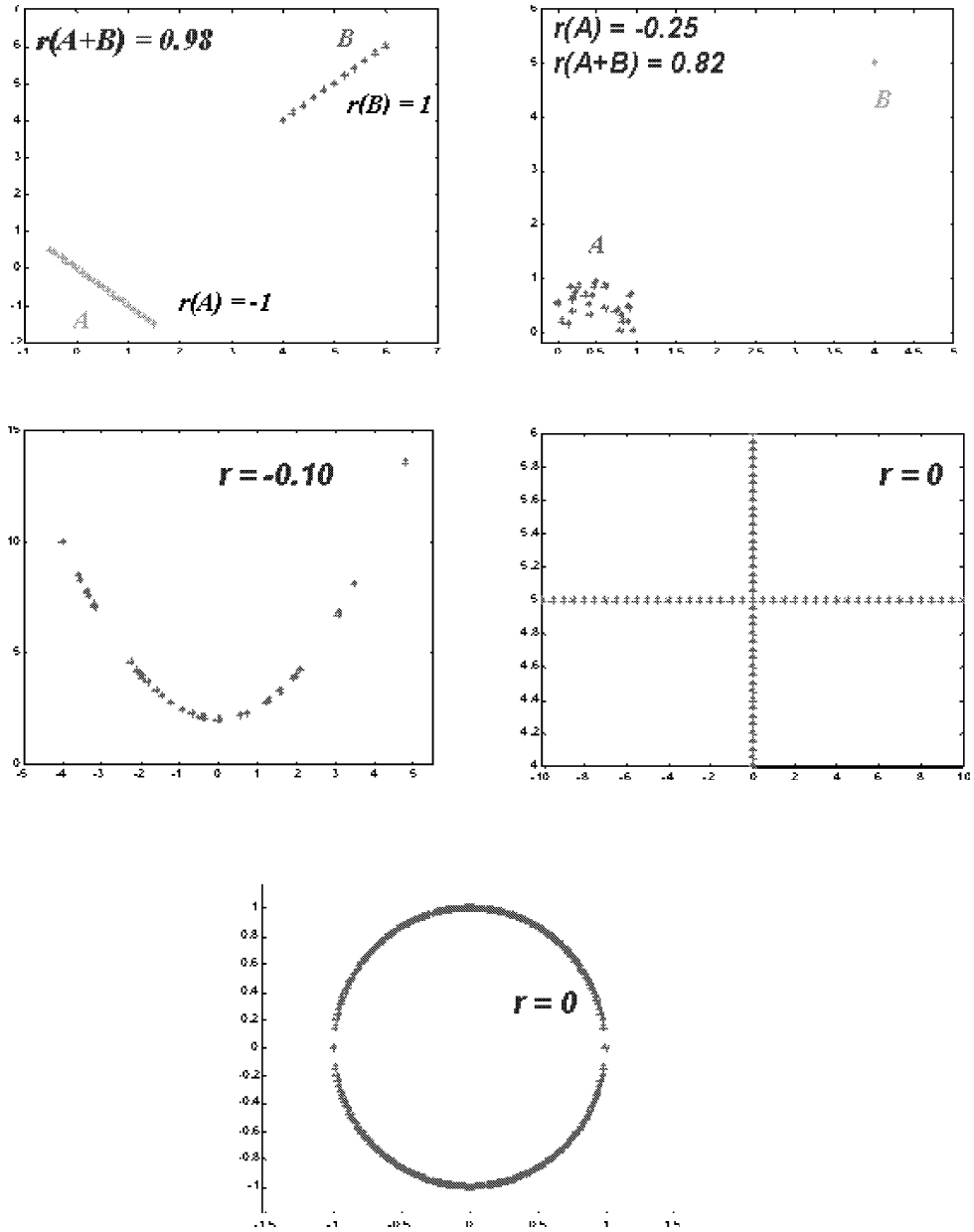


Figura 6.2: ¡OJO! al interpretar el coeficiente de correlación lineal

País	% Pob. Urbana	Tasa natalidad	Tasa mortalidad	Esperanza vida	Fecundidad	Mortalidad infantil
ARGENTINA	86.2	20.3	8.6	71.0	2.8	95.3
BOLIVIA	51.4	34.4	9.3	54.5	4.6	77.5
BRASIL	76.9	26.1	7.5	65.6	3.2	81.1
COLOMBIA	70.3	25.8	5.9	68.8	2.9	86.7
COSTA RICA	53.6	26.3	3.7	74.9	3.1	92.8
CHILE	85.6	22.5	6.4	71.8	2.7	93.4
ECUADOR	56.9	30.9	6.9	66.0	3.9	85.8
EL SALVADOR	44.4	33.5	7.1	64.4	4.0	73.0
GUATEMALA	42.0	38.7	7.6	63.4	5.4	55.1
HAITI	30.3	35.3	11.9	55.7	4.8	53.0
HONDURAS	43.6	37.1	7.2	64.9	4.9	73.1
MEXICO	72.6	27.9	5.4	69.7	3.2	87.3
NICARAGUA	59.8	40.5	6.9	64.8	5.0	81.0
PANAMA	54.8	24.9	5.2	72.4	2.9	88.1
PARAGUAY	47.5	33.0	6.4	67.1	4.3	90.1
PERU	70.2	29.0	7.6	63.0	3.6	85.1
R. DOMINICANA	60.4	28.3	6.2	66.7	3.3	83.3
URUGUAY	85.5	17.1	10.3	72.2	2.3	96.2
VENEZUELA	90.5	28.3	5.4	70.0	3.5	88.1
CUBA	74.9	17.4	6.7	75.4	1.9	94.0

Tabla 6.1: Indicadores demográficos de 20 países de A.L.(PNUD 1992)

Variables	% Pob. % Urbana	Tasa natalidad	Tasa mortalidad	Esperanza vida	Fecundidad	Mortalidad infantil
% Pob. urbana	1.0	-0.739	-0.179	0.588	-0.735	-0.532
T. natalidad	-0.739	1.0	0.101	-0.723	0.972	0.682
T. mortalidad	-0.179	0.101	1.0	-0.609	0.262	0.533
Esperanza V.	0.588	-0.723	-0.609	1.0	0.769	0.951
Fecundidad	-0.735	0.972	0.262	0.769	1.0	0.709
M. infantil	0.532	0.682	0.533	-0.951	0.709	1.0

Tabla 6.2: Matriz de correlación

Decir que el PNB está relacionado al tipo de clima significa que conociendo el tipo de clima de un país se podrá inferir su PNB. Esto podrá darse si los valores del PNB difieren muchos de un grupo a otro.

Construiremos un índice basado en esta variabilidad del PNB.

Si $\bar{y}$ es la media empírica de la variable Y sobre el total de las n observaciones, la varianza de todas estas observaciones es igual a:

$$s_y^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{k=1}^{n_j}(y_{kj} - \bar{y})^2$$

Como se puede distinguir las observaciones según la modalidad que toman sobre la variable X, se puede calcular medias y varianzas en los p grupos inducidos por las modalidades de X.

Si $\bar{y}_j$ es la media de la variable Y sobre las observaciones que toman la misma modalidad j.

la varianza de las observaciones de este grupo es igual a:

$$w_j^2 = \frac{1}{n_j}\sum_{k=1}^{nj}(y_j - \bar{y}_j)^2$$

La variabilidad total de los valores proviene de dos fuentes: la variabilidad al interior de los grupos y la variabilidad entre los grupos.

Consideramos ω^2 la media ponderada de las varianzas de los p grupos: $\omega^2 = \sum_j \frac{n_j}{n} w_j^2$ y b^2 la varianza entre las medias $\bar{y}_j$ de los grupos: $b^2 = \sum_j \frac{n_j}{n}(\bar{y}_j - \bar{y})^2$. Considerando que la media ponderada por los efectivos relativos $\frac{n_j}{n}$ de las medias $\bar{y}_j$ es igual a la media total $\bar{y}$ ($\sum \frac{n_j}{n}\bar{y}_j = \bar{y}$) se puede mostrar que:

$$s^2 = \beta^2 + \omega^2 = \sum_j \frac{n_j}{n}(\bar{y}_j - \bar{y})^2 + \sum_j \frac{n_j}{n} w_j^2$$

Si w^2 es nula, todas las varianzas w_j^2 son nulas y todas las observaciones en un mismo grupo j toman el mismo valor sobre la variable Y, que es igual a la media $\bar{y}_j$ del grupo, y en consecuencias se podrá obtener el valor de un observación sobre la variable Y conociendo su modalidad sobre X. Se observa en este caso una relación funcional de X hacia Y. Esta relación permite estimar el valor de Y para una nueva observación conociendo su valor sobre X.

Al contrario si la varianza entre los grupos b^2 es nula, entonces todas las medias $\bar{y}_j$ son iguales a $\bar{y}$: no se podrá decir nada sobre el valor de Y conociendo la modalidad de X. No se detecto ninguna relación funcional de X hacia y. Se deduce el siguiente índice que permite medir el grado de asociación de tipo funcional de X hacia Y:

$$\eta^2_{Y|X} = \frac{b^2}{s_y^2}$$

Este coeficiente toma valores entre 0 y 1:

$\eta^2_{y\|x} = 1$	relación funcional estricta
$0 < \eta^2_{y\|x} < 1$	tendencia funcional
$\eta^2_{y\|x} = 0$	ausencia de tendencia funcional

La tendencia funcional aumenta con $\eta^2_{y|x}$.

6.3.1 Codificación óptima de una variable nominal

Si se codifica la variable X, atribuyendo un valor numérico a cada una de sus modalidades, podremos usar el coeficiente de correlación lineal como índice de asociación. Pero no se puede codificar de cualquier manera. Una forma natural de hacerlo consiste en buscar la

codificación de las modalidades de X que produzca la mayor correlación lineal con la variable Y.

Si X tiene p modalidades, se le pueden asociar p variables indicativas $\{X^1, X^2, ..., X^p\}$ tales que

$$X^j(k) = \begin{cases} 1 & \text{si el individuo k toma la modalidad j de X} \\ 0 & \text{sino} \end{cases}$$

Se observa que $\sum_{j=1}^{p} X^j(k) = 1 \quad (\forall k)$.

Entonces si a_j es la codificación de la modalidad j $(j = 1, ..., p)$, la variable cuantitativa ξ asociada a esta codificación puede escribirse:

$$\xi(k) = \sum_j a_j X^j(k)$$

Dada $\{(x_i, y_i)|i = 1, ..., n\}$ una muestra de (X, Y), se define la codificación $\{a_j|j = 1, ...p\}$ que maximiza

$$cor(y, \sum_j a_j x^j)$$

Numéricamente el máximo de $cor^2(y, \sum_j a_j x^j)$ es igual a $\eta^2_{y|x}$.

6.3.2 Relación funcional entre dos variables cuantitativas

Cuando un coeficiente de correlación lineal entre X e Y es bajo, significa que las variables X e Y no están ligadas linealmente pero, puede existir otro tipo de relación entre ellas. Ahora bien, vimos que por codificación se puede transformar una variable nominal en una variable cuantitativa, inversamente, se puede transformar una variable cuantitativa en una variable ordinal, por lo tanto nominal particionando el recorrido de los valores de la variable en p intervalos.

Si se transforma X en variable nominal, se puede calcular la razón de correlación $\eta^2_{y|x}$, que permitirá detectar la existencia de una relación funcional de X hacia Y. El valor del coeficiente dependrá de la transformación (número de modalidades construidas). Se observa que ahora se tiene un coeficiente que no es simétrico en las variables como en el caso del coeficiente de correlación lineal. Por lo cual obtendremos resultados distintos según la variable que transformemos, salvo si existe una relación biyectiva entre las dos variables. Además, la razón de correlación es más general que el coeficiente de correlación lineal, y se tiene que $cor^2(x, y) \leq \eta^2_{y|x}$.

Se ilustra en el ejercicio al final del capítulo como estas trasformaciones influyen sobre los coeficientes de asociación.

6.4 VARIABLES NOMINALES

6.4.1 Tabla de contingencia

Los datos obtenidos sobre las dos variables nominales pueden resumirse en una tabla de contingencia. Una tabla de contingencia contiene las frecuencias absolutas conjuntas de las dos variables, es decir las frecuencias obtenidas al cruzar las modalidades de una variable con las modalidades de la otra. En la elección de concejales de 1991, se pueden asociar a cada votante la lista votada y la región. Se puede resumir los resultados en una tabla de frecuencias (Tabla 6.3), que es la única información que se conoce realmente en este caso (por el anonimato de la elección). Esta es una tabla de contingencia. Al pasar de la información individual de los votantes relativa a las dos variables a la tabla de contingencia no se pierde información, salvo la identificación de cada individuo.

Se puede buscar en la tabla si hay mayor concentración de votantes en una región para un partido dado. Vamos a construir un índice que permite medir la existencia de una relación entre las dos variables.

6.4.2 Ji-cuadrado de contingencia

Veamos como leer una tabla de contingencia con ejemplos sencillos (Tablas 6.4 y 6.5 con la variable X en fila y la variable Y en columna. Se observa en la tabla 6.4(a), que las columnas B_1 y B_2 son proporcionales, lo que significa que reparten sus totales en las mismas proporciones entre las modalidades A_1 y A_2. Las modalidades B_1 y B_2 tienen los mismos perfiles. Al observar esta tabla no se ven muchas relaciones entre las dos variables (conociendo una modalidad de una variable, no se puede decir nada sobre la otra variable). No es el caso de la tabla 6.4(b). En efecto, si una observación toma la modalidad B_1, tomará la modalidad A_2 de X; dada A_1, entonces se tendra la modalidad B_3 de Y, pero dada A_2, se tendra B_1 ó B_2. Se tiene entonces una relación funcional de Y hacia X y existe una relación de X hacia Y, pero no es de tipo funcional.

En el caso de la tabla 6.5(d) existe una relación funcional, pero en la tabla 6.5(c) no hay ninguna.

Si denotamos n_{ij}, $(i = 1, .., p, j = 1, ..., q)$ los elementos de una tabla de contingencia, se tienen los márgenes-filas: $n_{i\bullet} = \sum_j n_{ij}$, $i = 1, ..., p$, y los márgenes-columnas $n_{\bullet j} = \sum_i n_{ij}$, $j = 1, ...q$. Se define los perfiles condicionales como:

- Los perfiles condicionales-filas: $\frac{n_{ij}}{n_{i\bullet}}$
- Los perfiles condicionales-columnas: $\frac{n_{ij}}{n_{\bullet j}}$

La variable Y no influye sobre la variable X si y solo si los perfiles condicionales-columnas son todos iguales:

PARTIDO	I	II	III	IV	V	METR.	VI
D.C.	30412	63020	16793	58345	226333	759639	90521
RADICAL	19268	19265	9282	14336	39941	59767	21249
A.H. VERDE	2186	0	0	0	1680	43284	784
SOCIALDEMO	596	0	225	562	2817	6351	0
INDEP	346	73	55	86	1608	16493	2383
PPD	5165	11800	7390	21429	56405	295474	30714
SOCIALISTA	3405	15341	18339	28041	33282	177570	35779
INDEP	385	0	0	0	0	0	122
COMUNISTA	36648	13951	11588	21614	51135	171715	19312
LIBERAL	0	248	378	0	512	0	328
R.N.	12236	12424	16795	54648	96224	311801	54439
NACIONAL	0	0	0	0	422	2325	0
INDEP	3971	11669	4202	9385	40126	88614	7877
U.D.I.	8631	17464	8474	14495	71573	314984	33869
INDEP	587	980	47	0	6905	32008	6340
U.C.C.	6460	15428	5623	10671	73163	181913	26395
INDEP	105	5582	6007	1337	12263	37898	12797
IND IQUIQUE	24757	0	0	0	0	0	0
TOTAL	153888	187245	105198	234979	714389	24999836	342909

PARTIDO	VII	VIII	IX	X	XI	XII	TOTAL
D.C.	114070	223287	118841	121815	11555	13287	1848188
RADICAL	23416	61692	14420	26815	1602	2209	313562
A.H. VERDE	0	2931	1069	585	0	0	52519
SOCIALDEMO	7076	1110	6761	1291	0	0	26789
INDEP	1211	2631	1942	3572	45	27	30472
PPD	27759	64167	25498	29682	1250	8739	585472
SOCIALISTA	38338	94626	18987	51485	3715	20786	539694
INDEP	0	0	0	0	0	0	507
COMUNISTA	18379	50121	9824	13202	2342	2546	421377
LIBERAL	0	0	13842	0	241	0	15549
R.N.	60524	87849	56951	77702	8760	5807	856160
NACIONAL	0	1467	0	0	0	0	4214
INDEP	13644	29665	45384	23587	277	723	279124
U.D.I.	45905	75230	18194	32183	2065	8273	651340
INDEP	4794	16420	2358	3385	1598	731	76153
U.C.C.	47112	72049	21566	50650	1478	4237	516745
INDEP	13977	26376	9356	9066	119	1443	136326
IND IQUIQUE	0	0	0	0	0	0	24757
TOTAL	416205	809891	364993	445020	35047	69348	6378948

Tabla 6.3: Resultados de la elección de consejales de 1991

(a)

	B_1	B_2	B_3	
A_1	50	100	10	160
A_2	100	200	50	350
	150	200	60	

(b)

	B_1	B_2	B_3	
A_1	0	0	50	50
A_2	10	12	0	22
	10	12	50	

Tabla 6.4: Ejemplos de tablas de contingencias

(c)

	B_1	B_2	B_3	
A_1	20	10	7	37
A_2	40	20	14	74
A_3	80	40	28	148
	140	70	49	

(d)

	B_1	B_2	B_3	
A_1	0	20	0	20
A_2	30	0	0	30
A_3	0	0	0	25
	30	20	25	

Tabla 6.5: Más ejemplos de tablas de contingencia

$$\frac{n_{i1}}{n_{\bullet 1}} = \frac{n_{i2}}{n_{\bullet 2}} = ... = \frac{n_{iq}}{n_{\bullet q}} = \frac{n_{i\bullet}}{n} \quad (i = 1, ..., p)$$

De la misma manera la variable X no influye sobre la variable Y si y solo si los perfiles condicionales-filas son todos iguales:

$$\frac{n_{1j}}{n_{1\bullet}} = \frac{n_{2j}}{n_{2\bullet}} = ... = \frac{n_{pj}}{n_{p\bullet}} = \frac{n_{\bullet j}}{n} \quad (j = 1, ..., q)$$

Luego las dos variables X e Y serán independientes si y solo si se cumplen a la vez las dos condiciones anteriores. Se puede demostrar que equivalen a:

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n} \quad \forall (i, j)$$

Considerando las diferencias $n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{n}$, se puede evaluar cuán lejos está la relación entre X e Y de la independencia. Se puede construir un índice que traduzca estas diferencias, tomando en cuenta la importancia de cada una, ponderando por la magnitud de n_{ij} o $\frac{n_{i\bullet} \times n_{\bullet j}}{n}$. Es el índice χ^2 de contingencia:

$$\chi^2 = \sum_{ij} \frac{(n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{n})^2}{\frac{n_{i\bullet} \times n_{\bullet j}}{n}}$$

Este índice es nulo cuando X e Y son independientes y crece al alejarse de la independencia hasta un valor máximo igual a $n \times Min\{p-1, q-1\}$ cuando hay una relación estricta entre los dos variables.

6.4.3 Codificación de las dos variables nominales

Ahora tendremos que codificar ambas variables. Sean a_i, $i = 1, ..., p$ las codificaciones de las modalidades de X y X^i, $i = 1, ..., p$ las variables indicadoras de X; sean b_i, $i = 1, ..., q$ las codificaciones de las modalidades Y e Y^i, $i = 1, ..., q$ las indicadoras de Y.

Se busca codificaciones respectivas de X e Y tales que el coeficiente de correlación lineal de las codificaciones

$$cor(\sum_i a_i x^i, \sum_j b_j y^j)$$

sea máximo.

Esta correlación se usa en una técnica llamada análisis factorial de correspondencias y esta relacionada al ji-cuadrado de contingencia.

6.4.4 Relación entre dos variables cuantitativas

Si transformamos las dos variables cuantitativas en variables nominales podremos usar el χ^2 de contingencia que nos permita detectar una relación de cualquier tipo, no solamente lineal o funcional.

Para hacer las transformaciones se requiere un gran número de observaciones para tener una cantidad suficiente de elementos en cada celda de la tabla de contingencia.

Se observara que las transformaciones producen variables menos precisas que las originales. pero con estas se puede investigar otras relaciones que las lineales.

6.5 VARIABLES ORDINALES

6.5.1 Coeficientes de correlación de rangos

A partir de una variable ordinal, se pueden ordenar las observaciones de manera creciente y deducir una nueva variable que es *el rango*, que indica la posición de cada observación según el orden.

Sean $x_1, ..., x_n$ las realizaciones de la variable ordinal X y $R_{x_1}, ..., R_{x_n}$ los rangos asociados:

$$R_{x_i} < R_{x_j} \Longleftrightarrow x_i < x_j$$

Si R_{x_i} y R_{y_i}, $i = 1, 2, ..., n$, son los rangos asociados a X e Y respectivamente, se define entonces **el coeficiente de rangos de SPEARMAN** R_S de x e y como el coeficiente de correlación lineal empírico entre R_x y R_y.

Si $D_i = R_{x_i} - R_{y_i}$, se obtiene una expresión más práctica:

$$R_S = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)}$$

Se observa entonces que si los rangos inducidos por X e Y son idénticos $R_S = 1$; si son totalmente opuestos R_S = -1.

Si en vez de definir los rangos, se define dos nuevas variables sobre los pares de observaciones:

$$\begin{array}{|ll|}
\hline
S(x_i, x_j) = \text{-}1 & \text{si } x_i \geq x_j \\
S(x_i, x_j) = 1 & \text{si } x_i < x_j \\
S(x_i, x_j) = \text{-}1 & \text{si } x_i \geq x_j \\
S(y_i, y_j) = 1 & \text{si } y_i < y_j \\
S(y_i, y_j) = \text{-}1 & \text{si } y_i \geq y_j \\
\hline
\end{array}$$

Se define entonces **el coeficiente de correlación de rangos de KENDALL**:

$$\tau = \frac{\sum_{i,j} S(x_i, x_j) S(y_i, y_j)}{n(n-1)}$$

El numerador es igual al número de pares de observaciones con el mismo orden menos el número de pares de observaciones con orden contrario. El numerador es igual al número total de pares. Como el coeficiente R_S de Spearman, τ toma valores entre -1 y $+1$ y vale $+1$ si los ordenes son idénticos y -1 cuando son totalmente opuestos.

6.5.2 Relación entre dos variables cuantitativas

A partir de una variable cuantitativa se pueden ordenar las observaciones, y por lo tanto, construir los rangos. Puede ser útil especialmente cuando los valores de las variables no son muy precisos o bien, si se busca la existencia de una relación monótona no lineal entre X e Y. Se pueden aplicar entonces los coeficientes de correlación de rangos anteriores.

6.6 INFERENCIA

Suponiendo que un coeficiente de asociación fue correctamente calculado, es decir que fue calculado sobre una muestra aleatoria simple de una sola población, queremos saber a partir de qué valor se puede decidir la existencia o ausencia de una relación. Para esto se procede mediante un test de hipótesis sobre el valor del coeficiente v de asociación desconocido de la población: $H_0 : v = v_0$, o bien se puede calcular un intervalo de confianza para v. Para eso se requiere conocer la distribución del coeficiente de asociación v en la muestra.

6.6.1 Coeficiente de correlación lineal

¿Cuando se obtiene un coeficiente de correlación lineal r pequeño podemos admitir que la correlación ρ en la población es nula ó si r es grande, podemos concluir que existe una relación lineal?

Para responder a la pregunta se procede mediante un test de hipótesis sobre el valor del coeficiente de correlación ρ desconocido de la población: $H_0 : \rho = \rho_0$, o bien se puede calcular un intervalo de confianza para ρ. El problema es que la distribución del coeficiente de correlación r no es siempre fácil de establecer.

Cuando $\rho = 0$ y las dos variables X e Y provienen de una distribución normal bivariada, la distribución del coeficiente r de la muestra es fácil de obtener ye depende del tamaño n de la muestra: existen tablas de la distribución de r en función de n y para $n > 100$ y se puede aproximara a la normal $\mathcal{N}(0, \frac{1}{\sqrt{n-1}})$.

Por ejemplo, si un coeficiente de correlación lineal r es igual a 0.38 sobre una muestra de $n = 52$ observaciones, vamos a rechazar que $\rho = 0$ al nivel de significación de 5% o incluso 1%, dado que $\mathbb{P}(|r| > 0.27) = 0.05$ y $\mathbb{P}(|r| > .35) = 0.01$, pero si $r = 0.32$ con el mismo tamaño $n = 52$, entonces se rechaza al nivel de 5% pero no al nivel de 1%.

Cuando ρ no es nulo, la distribución exacta de r es mucho más complicada de determinar, sin embargo se puede usar una aproximación a partir de $n = 25$: si $Z = 1/2 \ln(\frac{1+r}{1-r})$, la distribución de Z se aproxima a una normal $\mathcal{N}(1/2 \ln(\frac{1+\rho}{1-\rho}), \frac{1}{\sqrt{n-3}})$.

Finalmente, si las dos variables no siguen una distribución normal, se puede usar los resultados anteriores cuando n es mayor que 30, pero si ρ es nulo, no se puede decir que hay independencia, pero sólo que no hay ligazón lineal.

6.6.2 Razón de correlación: ANOVA a un factor

Para estudiar la significatividad de una razón de correlación empírica obtenida sobre n observaciones entre la variable cuantitativa Y con la variable nominal X a p modalidades. se plantea la hipótesis nula $H_0 : \gamma^2 = 0$ donde γ^2 es la razón de correlación en la población. El problema es que no conocemos la distribución de la razón de correlación observado $\eta^2_{Y|X}$.

Si $Y \sim \mathcal{N}(\mu_j, \sigma_j^2)$ cuando X toma la modalidad j, entonces la hipótesis nula puede escribirse $H_o : \mu_1 = \mu_2 = ... = \mu_p$.

Sabemos que si w_j^2 es la varianza observada del grupo j con un efectivo n_j, $\frac{n_j s_j^2}{\sigma_j^2} \sim \chi^2_{n_j-1}$. Suponiendo que el muestreo es aleatorio simple, los grupos son independientes entre si y suponiendo que todas las varianzas σ_j^2 son iguales a σ^2, $\frac{nw^2}{\sigma^2} \sim \chi^2_{n-p}$ donde w^2 es el promedio de las varianzas dentro los grupos.

Por otra parte si b^2 es la varianza observada entre los p grupos, $\frac{nb^2}{\sigma^2} \sim \chi^2_{p-1}$. Además b^2 y w^2 son independientes, dado que cada media $\bar{y}_j$ es independiente de la varianza w_j^2 del grupo y que los grupos son independientes entre si.

Se considera entonces el estadístico que es el cociente de los dos χ^2 divididos respectivamente por sus grados de libertad. Como se cancelan n y σ^2, se obtiene:

$$\frac{b^2/(p-1)}{w^2/(n-p)}$$

que sigue una distribución F de Fisher a $p-1$ y $n-p$ grados de libertad bajo la hipótesis H_o de ausencia de relación de X hacia Y.

Se observará que

$$\frac{b^2/(p-1)}{w^2/(n-p)} = \frac{\eta^2/(p-1)}{1-\eta^2/(n-p)}$$

Consideremos el ejemplo debido a Ronald Fisher (párrafo 1.1) sobre 3 especies de flores de la familia de los "iris": setosa, versicolor y virgínica (variable X con 3 modalidades). Se busca verificar si las 3 especies se distinguen por algunas mediciones. Se usan dos variables: Y_1 el largo del pétalo e Y_2 el ancho del sépalo (tabla 6.6).

Entre X e Y_1 la razón de correlación es igual a $\eta^2_{Y_1|X} = 0.40$ y entre X e Y_2 la razón de correlación es igual a $\eta^2_{Y_2|X} = 0.94$. Claramente el largo del pétalo es diferentes de una especie a otra pero no se puede afirmar nada para el ancho del sépalo. Los resultados del ANOVA para ambos casos se encuentran en las tablas 6.7 y 6.8. La varianza b^2 proviene de la especie y la varianza w^2 se interpreta como un error cuando se supone que las medias de las 3 especias son iguales. Si bien en ambos casos se rechaza la hipótesis nula (p-valor nulo), el valor del F es mucho más pequeño en el caso del ancho del sépalo, lo que indica una relación menos clara.

Especie	efectivo	Y_1		Y_2	
		Media	Varianza	Media	Varianza
Setosa	50	34.280	3.39	14.620	4.30
Versicolor	50	27.700	3.39	42.600	4.30
Virginica	50	29.740	3.39	55.520	4.30

Tabla 6.6: medias y varianzas

Fuente	n*varianza	g.l.	n*varianza/g.l.	F	p-valor
Especie (b^2)	1134.493	2	567.247	49.160	0.000
Error (w^2)	1696.2	147	11.54		
Total (s^2)	2830.693	149			

Tabla 6.7: Ancho del sépalo

6.6.3 Ji-cuadrado de contingencia

¿Si dos variables nominales X e Y son independientes, cuales son los valores más probables del χ^2 de contingencia? Como vimos en el párrafo 5.5.6,

$$Q = \sum_{ij} \frac{(n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{n})^2}{\frac{n_{i\bullet} \times n_{\bullet j}}{n}}$$

Si X e Y son independientes, $n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$ para todo par (i, j); en esta caso el estadístico Q sigue una distribución aproximada de χ^2 a $(p-1)(q-1)$ grados de libertad, si p y q son los números de modalidades de X e Y respectivamente.

6.6.4 Coeficiente de correlación de rangos de Spearman

Cuando X e Y resultan de un ordenamiento sin empates, entonces los rangos inducidos por X ó Y son valores de $\{1, .., n\}$ y los rangos de uno se obtienen por permutación de los rangos del otro.

Si X e Y son independientes, cualquiera sean las leyes de X e Y, las dos permutaciones inducidas son independientes. En este caso, si el ordenamiento de X esta fijado, las $n!$ permutaciones de este ordenamiento son equiprobables. Se tiene tres maneras de obtener la distribución de R_S bajo la hipótesis de independencia:

Fuente	n*varianza	g.l.	n*varianza/g.l.	F	p-valor
Especie (b^2)	43710.28	2	21855.14	1180.16	0.000
Error (w^2)	2722.26	147	18.518776		
Total (s^2)	46432.54	149			

Tabla 6.8: Largo del pétalo

- Si n es muy pequeño, se puede obtener empíricamente la distribución de R_S, calculando los $n!$ valores asociados a las distintas permutaciones.
- Para $n < 100$, existen tablas de la distribución en función de n.
- Para n grande se puede usar la aproximación a la normal $\mathcal{N}(0, \frac{1}{\sqrt{n-1}})$.

El coeficiente de Spearman entre la Esperanza de Vida y la Tasa de Mortalidad de la tabla 1 vale 0.48.

En las tablas de la distribución del coeficiente de Spearman encontramos que $P(|R_S| > 0.447) = 0.05$, lo que nos lleva a rechazar la independencia entre la Esperanza de Vida y la Tasa de Mortalidad.

6.6.5 Coeficiente de correlación de rangos de Kendall

Como en el caso del coeficiente de correlación de Spearman, se puede construir empíricamente la distribución del τ de Kendall cuando n es muy pequeño. Pero a partir de $n > 8$, se puede aproximar a una distribución normal $\mathcal{N}(0, \sqrt{\frac{2(2n+5)}{9n(n-1)}})$

Para las variables Esperanza de Vida y Tasa de Mortalidad de la Tabla 1, obtenemos $\tau = 0.326$.

$$P(|\tau| < 1.96\sqrt{\frac{90}{180 \times 19}}) = P(|\tau| < 0.317) = 0.05$$

Nuevamente encontramos significativa la relación entre las dos variables.

6.7 EJERCICIO

Sea un conjunto I de $n = 300$ individuos, y cuatro variables cuantitativas X, Y, Z y T observadas sobre los 300 individuos. X varia entre -100 y 100, Y varia entre 0 y 10000, Z y T varían entre -1100 y 1100.

1. Los coeficientes de correlación lineal calculados sobre los 300 individuos son: $R_{X,Y} = -0.057$, $R_{Z,T} = 0.991$. Interprete estos coeficientes.
2. Se transforma la variable X en una variable nominal particionando |-100,100| en q intervalos iguales; se llama X_1, X_2, X_3 y X_4 a las variables nominales obtenidas para q=10, 8, 6 y 4 respectivamente. Interprete las razones de correlación obtenidas y concluir: $\eta_{Y/X_1} = 0.96$, $\eta_{Y/X_2} = 0.93$, $\eta_{Y/X_3} = 0.86$ y $\eta_{Y/X_4} = 0.74$,
3. Se transforma la variable Y en una variable nominal con la partición del intervalo |0,10000| en q intervalos iguales; se llama Y_1, Y_2, Y_3 y Y_4 a las variables nominales obtenidas para q=10, 8, 6 y 4 respectivamente. Interprete las razones de correlación obtenidas y concluya: $\eta_{X/Y_1} = 0.038$, $\eta_{X/Y_2} = 0.027$, $\eta_{X/Y_3} = 0.024$ y $\eta_{X/Y_4} = 0.015$,

4. Se calcula los χ^2 de contingencia entre las variables nominales asociadas a X e Y: $\chi^2_{X_1,Y_1} = 853$, $\chi^2_{X_2,Y_2} = 679$, $\chi^2_{X_3,Y_3} = 450$ y $\chi^2_{X_4,Y_4} = 306$. Concluya.

5. Interprete el coeficiente de correlación parcial de Z y T dado X $R_{Z,T|X} = 0.027$. Compare con $R_{Z,T}$ y interprete.

Capítulo 7

REGRESIÓN LINEAL

7.1 ¿PORQUE MODELAR?

Estudiamos en el capítulo anterior como detectar una asociación entre dos variables; generalmente los roles entre las variables no son simétricos - una variable puede influir sobre la otra y la recíproca no es necesariamente cierta - incluso más de una variable pueden intervenir en esta relación. En este caso nos interesaremos no solamente en evaluar la intensidad de la asociación, sino que también en describirla.

Algunas relaciones son conocidas y deterministas como ciertas leyes de la física o de la mecánica, y, dependen de constantes desconocidas que hay que determinar. Estas constantes pueden obtenerse a partir de experimentos que se utilizarán en el modelo ya planteado. El problema que surge entonces en la determinación de las constantes está en los errores de mediciones.

En otros problemas las relaciones no son conocidas y hay que determinar completamente el modelo. En ciencias sociales o economía, por ejemplo, los modelos no son deterministas y contienen una componente aleatoria, lo que dificulta la búsqueda de las relaciones. En este caso se busca descubrir como un conjunto de variables $X^1, X^2, ..., X^p$ influye sobre otra variable Y. Según el contexto, las variables X^j son llamadas **variables explicativas**, **variables independientes ó variables exógenas** y la variable Y es llamada **variable a explicar, variable repuesta, variable dependiente ó variable endógena**. Cuando las variables son cuantitativas, se busca una función real f que permita reconstituir los valores obtenidos sobre una muestra:

$$Y = f(X^1, X^2, ..., X^p)$$

Por una rázon histórica, este modelo se llama **regresión**. El mayor descubrimiento de Galton (párrafo 1.1) fueron sus formulaciones sobre la regresión simple y su relación con la distribución normal bivariada. Hizo un estudio que mostró que la altura de los niños nacidos de padre altos tiende a retroceder o "regresar" hacia la altura promedio de la población. Por lo que utilizó entonces la palabra "regresión" para referirse a este fenómeno.

Ejemplo 7.1.1 La distancia d que una partícula recorre en un tiempo t esta dada por la formula:

$$d = \alpha + \beta t$$

en que β es la velocidad promedio y α es la posición de la partícula en $t = 0$. Si α y β son desconocidos, observando la distancia d en dos tiempos distintos, la solución del sistema de 2 ecuaciones lineales permite obtener α y β. Sin embargo es difícil obtener en general la distancia sin error de medición el que es de tipo aleatorio. Por lo cual se observa una variable aleatoria: $Y = d + \epsilon$ en vez de d, en donde ϵ es el error de medición. En este caso no basta tener dos ecuaciones, sino que observar los valores de la distancia recorrida en varios periodos de tiempo y métodos estadísticos basados en la aleatoriedad del error, los que permitirán estimar α, β y d sobre la base de una relación de tipo lineal.

Ejemplo 7.1.2 Si consideramos el peso y la talla de las mujeres chilenas, es obvio que no existe una relación lineal ni funcional entre la talla y el peso, pero parece existir una cierta *tendencia*. Considerando que el peso P y la talla T son variables aleatorias de distribución conjunta normal bivariada, se plantea el modelo lineal:

$$E(P|T) = \alpha + \beta T$$

en que α y β dependen de los parámetros de la distribución conjunta de P y T. El peso se escribe entonces:

$$P = \alpha + \beta T + \epsilon$$

en que ϵ refleja la variabilidad del peso P entre las chilenas de la misma talla con respecto a la media.

Ejemplo 7.1.3 Para decidir de la construción de una nueva central eléctrica, ENDESA busca estimar el consumo total de electricidad en Chile después del año 2002. Por lo tanto, se construye un modelo que liga el consumo de electricidad con variables económicas y demográficas, estimado a partir de datos de los años anteriores. Se aplica entonces el modelo para predecir el consumo de electricidad según ciertas evoluciones económicas y demográficas.

Ejemplo 7.1.4 Para establecer una determinada publicidad a la televisión, se cuantifica el efecto de variables culturales y socio-económicas sobre la audiencia de los diferentes programas.

Ejemplo 7.1.5 El modelo lineal puede ser generalizado tomando funciones de las variables explicativas y/o de la variable a explicar. En particular para un ajuste polinomial se tiene una variable Y y la variable X con algunas de sus potencias:

$$Y = a_o + a_1 X^1 + ... + a_p X^p$$

en donde X^j es la potencia j de X.

Ejemplo 7.1.6 Se quiere estimar la constante g de la gravitación; para eso se toman los tiempos de caída t de un objeto desde una altura d dada del suelo.

$$d = \frac{1}{2}gt^2$$

Dados los errores de mediciones, varias observaciones son necesarias y se puede considerar este modelo como lineal tomando como variable t^2.

Nos limitaremos en este texto a los modelos lineales, es decir: la variable repuesta se escribe como combinación lineal de las variables explicativas.

Presentaremos dos métodos para estimar las constantes de un modelo lineal. Consideraremos el problema como un problema de ajuste y se propondrán el método de los mímimos cuadrados, que permite estimar los coeficientes del modelo lineal a partir de valores observados y el modelo normal para los errores que permite estimar las constantes a partir del método de máxima verosimilitud lo que permite estudiar las propiedades de los estimadores de las constantes y dar una precisión del ajuste. Finalmente se usará el modelo para hacer predicciones.

7.2 LOS MÍNIMOS CUADRADOS

Sean $\{(y_i, x_i^1, x_i^2, ..., x_i^p) | i = 1, ..., n\}$ los valores obtenidos sobre una muestra $p+1$ dimensional de tamaño n. Se plantea el modelo lineal:

$$y_i = \beta_o + \beta_1 x_i^1 + ... + \beta_p x_i^p \quad \forall i$$

donde $\beta_o, \beta_1, ..., \beta_p$ son las constantes desconocidas o sea los parámetros del modelo.

Como generalmente no existen constantes que cumplan exactamente esta relación para todas las observaciones, se escribe:

$$y_i = \beta_o + \beta_1 x_i^1 + ... + \beta_p x_i^p + \epsilon_i \quad \forall i$$

en donde ϵ_i es el **error** para la observación i debido al modelo. Se buscar entonces minimizar una función de los errores, por ejemplo.

$$\sum_i \epsilon_i^2 \qquad \sum_i |\epsilon_i| \qquad \sum_i Max\{\epsilon_i\}$$

El criterio de los mínimos cuadrados toma la función cuadrática $\sum_i \epsilon_i^2$ cuya solución es fácil de obtener y que tiene una interpretación geométrica simple.

Escribamos matricialmente el modelo aplicado a la muestra de observaciones.

Sea $\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} 1 & x_1^1 & x_1^2 & ... & x_1^p \\ 1 & x_2^1 & x_2^2 & ... & x_2^p \\ & ... & ... & ... & \\ 1 & x_n^1 & x_n^2 & ... & x_n^p \end{pmatrix}$, $\underline{\beta} = \begin{pmatrix} \beta_o \\ \beta_1 \\ ... \\ \beta_p \end{pmatrix}$, $\underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_n \end{pmatrix}$

Entonces, el modelo se escribe:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

El criterio de los mínimos cuadrados consiste entonces en buscar el punto del subespacio vectorial $W = Im(X)$ de $\mathbb{R}^n$ generado por las columnas de la matriz X lo más cercano al punto $\underline{y}$. La solución es la proyección ortogonal del punto $\underline{y}$ sobre W.

En efecto, $\sum_i \epsilon_i^2$ es igual a $\|\underline{\epsilon}\|^2$ el cuadrado de la norma del vector $\underline{\epsilon}$, es decir el cuadrado de la distancia entre los vectores $\underline{y}$ y $X\underline{\beta}$, siendo $X\underline{\beta}$ un vector del subespacio vectorial W. El vector de W solución es entonces la proyección ortogonal de Y sobre W. Si P es el operador lineal de proyección ortogonal sobre el subespacio vectorial W, entonces la solución es $X\underline{\hat{\beta}} = P\underline{y}$. La expresión matricial de P se puede obtener en función de la matriz X: El vector $\underline{y} - P\underline{y}$ es ortogonal a W o sea que $\underline{y} - X\underline{\hat{\beta}}$ es ortogonal a cada columna de X; si se denotan $X_o, X_1, ..., X_p$ las $p+1$ columnas de X, se expresa la ortogonalidad en termino de $p+1$ productos escalares:

$$< \underline{y} - X\underline{\beta}, X_j > \qquad (j = 0, 1, ...p)$$

Matricialmente se escribe: $X_j^t(\underline{y} - X\underline{\beta}) = 0$ $(\forall j)$, y juntando las $p+1$ ecuaciones se obtiene las **Ecuaciones Normales**:

$$X^t X\beta = X^t y$$

Este sistema de ecuaciones lineales tiene una solución única cuando las columnas de X son linealmente independientes o sea si forman una base del subespacio vectorial de W, lo que ocurre cuando X es de rango igual a $p+1$. En este caso la solución de los mínimos cuadrados es igual a:

$$\underline{\hat{\beta}} = (X^t X)^{-1} X^t \underline{y}$$

Se puede obtener el resultado por derivación matricial también.

Observamos que el estimador $\underline{\hat{\beta}}$ de β es lineal en Y.

El operador de proyección ortogonal sobre W se escribe matricialmente como:

$$P = X(X^t X)^{-1} X^t$$

Este operador lineal P es idempotente de orden 2 ($P^2 = P$) y simétrico ($P^t = P$).

Si la matriz X es de rango incompleto (rango inferior a $p+1$), basta encontrar una base de W entre las columnas de X, y reemplazar X por la matriz formada de estas columnas linealmente independientes.

7.3 MÁXIMA VEROSIMILITUD

En el párrafo anterior, se uso un criterio matemático para estimar los coeficientes β_j. Aquí usaremos un modelo probabilístico y el método de máxima verosimilitud para estimarlos. El modelo consiste en la esperanza condicional de la variable repuesta Y dadas las variables explicativas $X^1, X^2, ..., X^p$:

$$E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = X\beta$$

con $Y = E(Y) + \varepsilon = X\beta + \varepsilon$ en donde se supone $\varepsilon \sim N_n(0, \sigma^2 I_n)$. La función de verosimilitud utilizada es la densidad conjunta de los errores:

$$f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\varepsilon^t\varepsilon\right)$$

$$f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n; \beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)^t(Y - X\beta)\right)$$

El estimador de máxima verosimilitud de β verifica las Ecuaciones Normales:

$$\frac{\partial \ln f}{\partial \beta} = 0 \Rightarrow \frac{\partial (Y - X\beta)^t (Y - X\beta)}{\partial \beta} = 0 \Rightarrow (X^t X)\hat{\beta} = X^T Y$$

Calculemos el estimador de máxima verosimilitud de σ^2:

$$\frac{\partial \ln f}{\partial \sigma^2} = 0 \Rightarrow \sigma^2 = \frac{(Y - X\hat{\beta})^t (Y - X\hat{\beta})}{n}$$

y si $\hat{\varepsilon} = Y - X\hat{\beta}$, entonces

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Es decir que la función de verosimilitud es máxima cuando se cumplen las ecuaciones normales: $(X^t X)\hat{\beta} = X^t Y$ y además $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ llamado la varianza residual dado que es la varianza empírica de los $\hat{\in}_i$; en efecto ya que $Y = X\hat{\beta} + \hat{\varepsilon}, \quad \hat{\varepsilon} \in Im(X)$ y

$$X\hat{\beta} \in (Im(X))^{\perp} \Rightarrow \hat{\varepsilon} \perp 1_n \Rightarrow \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 0$$

El estimador de los mínimos cuadrados es igual entonces al estimador de máxima verosimilitud cuando se tiene el supuesto de normalidad $\varepsilon \sim N_n(0, \sigma^2 I_n)$.

7.4 PROPIEDADES DE LOS ESTIMADORES

Las propiedades del estimador $\underline{\hat{\beta}}$ estan ligadas a los supuestos hechos sobre los errores ϵ_i. Supondremos aquí que X es de rango $p+1$ o sea $\underline{\hat{\beta}} = (X^t X)^{-1} X^t \underline{y}$.

- El estimador es insesgado: $E(\underline{\epsilon}) = \underline{0} \Longrightarrow E(\underline{\hat{\beta}}) = \underline{\beta}$
- El estimador es consistente.

- El estimador tiene mínima varianza: Teorema de GAUSS MARKOV:

Teorema 7.4.1 *Si $E(\underline{\epsilon}) = \underline{0}$ y $E(\underline{\epsilon}\underline{\epsilon}^t) = \sigma^2 I_n$, entonces toda combinación lineal $a^t\underline{\hat{\beta}}$ de $\underline{\hat{\beta}}$ tiene mínima varianza entre los estimadores insesgados lineales en $\underline{y}$ de $a^t\underline{\beta}$. Además si $\varepsilon \sim N_n(0, \sigma^2 I_n)$, entonces $\hat{\beta}$ tiene mínima varianza entre todos los estimadores insesgados de β.*

Demostración Hay que comparar las varianzas de $a^t\underline{\hat{\beta}}$ y $a^t\underline{\beta}^*$ en que $\underline{\beta}^*$ es un estimador insesgado de la forma $C\underline{y}$.

$\underline{\beta}^* = \underline{\hat{\beta}} + D\underline{y}$, en que $D = C - (X^tX)^{-1}X^t$.

Como los dos estimadores son insesgados, $E(Dy) = 0$ y luego $DX = 0$.

$Var(\underline{\beta}^*) = Var(\underline{\hat{\beta}}) + Var(D\underline{y}) + 2Cov(\underline{\hat{\beta}}, D\underline{y})$

$Cov(\underline{\hat{\beta}}, D\underline{y}) = \sigma^2(X^tX)^{-1}X^tD^t = 0$

$Var(\underline{\beta}^*) = Var(\underline{\hat{\beta}}) + \sigma^2 DD^t$

Luego, $Var(a^t\underline{\beta}^*) = a^tVar(\underline{\hat{\beta}})a + \sigma^2 a^tDD^ta$

Como $\sigma^2 a^tDD^ta > 0$, $Var(\underline{\beta}^*) > Var(\underline{\hat{\beta}})$.

Si además los errores siguen una distribución normal, el estimador $\hat{\beta}$ es de mínima varianza entre todos los estimadores insesgados de β. En efecto la cantidad de información de la muestra multivariada para el parámetro β es igual a

$$I_n(\beta) = \frac{1}{\sigma^2}X^tX$$

y el estimador $\hat{\beta}$ tiene una matriz de varianza igual a $\sigma^2(X^tX)^{-1}$. Luego se obtiene la igualdad en la desigualdad de Cramer-Rao. ■

- La estimación de σ^2 obtenida por máxima verosimilitud es sesgada:
$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i^1 - ... - \hat{\beta}_p x_i^p$. Entonces, si $Q = I - P \Rightarrow \underline{\hat{\varepsilon}} = Q\underline{y} = Q\underline{\epsilon} \Rightarrow \sum \hat{\varepsilon}_i^2 = \underline{\hat{\varepsilon}}^t\underline{\hat{\varepsilon}} = \underline{\epsilon}^tQ^tQ\underline{\epsilon} = \underline{\epsilon}^tQ\underline{\epsilon} = Traza(Q\underline{\epsilon}\underline{\epsilon}^t)$ Luego $E(\underline{\hat{\varepsilon}}^t\underline{\hat{\varepsilon}}) = Traza(QE(\underline{\epsilon}\underline{\epsilon}^t)) = \sigma^2 Traza(Q)$

$Traza(Q) = Traza(I - X(X^tX)^{-1}X^t) = n - Traza(I_{p+1}) = n - p - 1$

Es decir que $E(\underline{\hat{\varepsilon}}^t\underline{\hat{\varepsilon}}) = (n - p - 1)\sigma^2$

Se obtiene entonces un estimador insesgado de σ^2 tomando:

$$\hat{\sigma}^2 = \frac{\underline{\hat{\varepsilon}}^t\underline{\hat{\varepsilon}}}{n - p - 1} = \frac{1}{n - p - 1}(\underline{y} - X\underline{\hat{\beta}})^t(\underline{y} - X\underline{\hat{\beta}})$$

7.5 INTERVALO DE CONFIANZA PARA LOS COEFICIENTES

Para cada parámetro β_j del modelo lineal, se puede construir un intervalos de confianza utilizado:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim t_{n-r}$$

en donde $\hat{\sigma}_j^2$ es la estimación de $Var(\hat{\beta}_j) = \sigma^2 (X^t X)^{-1}_{jj}$; es decir $\hat{\sigma}_j^2 (X^t X)^{-1}_{jj}$. El intervalo de confianza de nivel de confianza igual a $1 - \alpha$ es:

$$\left| \hat{\beta}_j - t_{n-r}^{\alpha/2} \hat{\sigma}_j, \hat{\beta}_j + t_{n-r}^{\alpha/2} \hat{\sigma}_j \right|$$

7.6 CALIDAD DEL MODELO

Para ver si el modelo es válido, hay que realizar varios estudios: la verificación de los supuestos sobre los errores, la forma y significación de las dependencias y el aporte de cada variable explicativa. Lo que se hará estudiando, mediante gráficos, índices y test, no solamente la calidad del modelo global y el aporte individual de cada variable explicativa, sino que el aporte de un grupo de m variables explicativas también.

7.6.1 Calidad global del modelo

Los residuos $\hat{\varepsilon}_i$ dan la calidad del ajuste para cada observación de la muestra. Pero es una medida individual que depende de la unidad de medición. Un índice que evita este problema está dado por:

$$\frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2}$$

que representa el cuadrado del coseno del ángulo del vector Y con el vector $\hat{Y}$ en $\mathbb{R}^n$ (Figura **??**).

Se pueden comparar las siguientes varianzas:

- Varianza residual: $\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$.
- Varianza explicada por el modelo: $\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$.

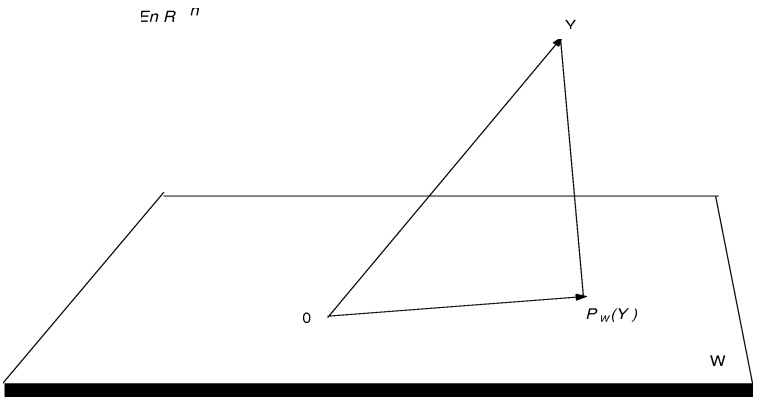


Figura 7.1: Proyección del vector Y en W

- Varianza total: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Un índice estadísticamente más interesante es el **coeficiente de correlación múltiple** R o su cuadrado, el **coeficiente de determinación**:

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

que compara la varianza explicada por el modelo con la varianza total. El coeficiente de correlación múltiple R es el coeficiente de correlación lineal entre Y e $\hat{Y}$. El valor de R está comprendido entre 0 y 1.

Cuando $R = 0$, el modelo obtenido es: $\hat{y}_i = \bar{y}$ $(\forall i)$ ($\bar{y}$ es la media muestral de los valores y_i), y en consecuencia las variables no explican nada en el modelo. En cambio cuando R es igual a 1, el vector Y pertenece al subespacio vectorial W, es decir que existe un modelo lineal que permite escribir las observaciones y_i exactamente como combinación de las variables explicativas. Cuando R es cercano a 1, el modelo es bueno siendo que los valores estimados $\hat{y}_i$ ajustan bien los valores observados y_i.

Para el caso general se tiene:

$$Corr(Y, \hat{Y}) = \frac{||\hat{Y} - \bar{y}1_n||}{||Y - \bar{y}1_n||} = \max_{Z=X\beta} Corr(Y, Z)$$

en donde 1_n es el valor de la bisectriz de $\mathbb{R}^n$ de componentes todas iguales a 1.

Si se plantea la hipótesis global $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \iff H_0 : E(y_i) = \beta_0$ $(\forall i)$, esta hipótesis significa que los valores de las p variables explicativas no influyen en los valores de

Y. Como $\hat{\varepsilon} \sim N_n(0, \sigma^2(I_n - P))$ e $\hat{Y} \sim N_n(X\beta, \sigma^2 P)$, si r es el rango de la matriz X, se tiene:

$$\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sigma^2} = \frac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-r}.$$

Como $\hat{Y}|_{H_0} \sim N_n(\beta_0 1_n, \sigma^2 P) \iff \hat{\beta}_0 = \bar{y}$, se tiene:

$$\sum_{i=1}^{n} \left(\frac{y_i - \beta_0}{\sigma} \right)^2 \sim \chi^2_{r-1} \quad \text{y} \quad \sum_{i=1}^{n} \left(\frac{\hat{y}_i - \bar{y}}{\sigma} \right)^2 \sim \chi^2_{r-1}$$

Además $\dfrac{\sum_{i=1}^{n} \hat{y}_i^2}{\sigma^2}$ y $\displaystyle\sum_{i=1}^{n} \left(\frac{\hat{y}_i - \bar{y}}{\sigma} \right)^2$ son independientes. Se tiene entonces que bajo la hipótesis nula H_0:

$$F = \frac{\dfrac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{r-1}}{\dfrac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-r}} \sim F_{r-1,n-r}$$

en donde $F_{r-1,n-r}$ sigue una distribución de Fisher a $r-1$ y $n-r$ grados de libertad. Se puede expresar F en función del coeficiente de correlación múltiple R:

$$F \frac{(n-r)R^2}{(r-1)(1-R^2)}.$$

La región crítica para la hipótesis nula $H_0 : E(Y|X) = \beta_0 1_n$ contra la hipótesis alternativa $H_1 : E(Y|X) = X\beta$ con un nivel de significación α está definida por

$$I\!P(F_{r-1,n-r} > c_\alpha) = \alpha.$$

Se rechaza H_0, por lo tanto se declara el modelo globalmente significativo cuando se encuentra un valor F en la muestra mayor que c_α.

En la práctica, se define la **probabilidad crítica** o **p-valor** que es el valor p_c tal que $I\!P(F_{r-1,n-r} > F) = p_c$. Si el valor de la probabilidad crítica p_c es alta, no se rechaza H_0, es decir que se declara el modelo como poco significativo.

7.6.2 Medición del efecto de cada variable en el modelo

Cuando las variables explicativas son independientes, el efecto asociado a una variable X_j se mide con $X_j\hat{\beta}_j$. Se observará que el modelo lineal es invariante ante el cambio de las escalas de medición.

Consideremos la hipótesis nula $H_0 : \beta_j = 0$. Como $\hat{\beta}_j \sim N(\beta_j, \sigma_j^2)$ en donde $\sigma_j^2 = Var(\hat{\beta}_j)$ ($\sigma_j^2 = \sigma^2 (X^t X)^{-1}_{j,j}$ en el caso del modelo con rango completo), $\frac{\hat{\beta}_j - \beta_j}{\sigma_j} \sim N(0,1)$. Por otra parte, como $\frac{(n-r)\hat{\sigma}}{\sigma^2} \sim \chi^2_{n-r}$, se deduce que

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim t_{n-r}.$$

Bajo la hipótesis nula $H_0 : \beta_j = 0$,

$$\frac{\hat{\beta}_j}{\hat{\sigma}_j} \sim t_{n-r}.$$

Si la probabilidad crítica o $p-valor$ $I\!P\left(|t_{n-r}| > \frac{\hat{\beta}_j}{\hat{\sigma}_j}\right) = p_c$ es grande, no se rechaza H_0 y si es pequeña se rechaza H_0, lo que en este caso muestra un efecto significativo de la variables X_j sobre Y.

Estos tests individuales sobre los efectos tienen validez cuando las variables explicativas son relativamente independientes. Cuando esto ocurre, es decir, cuando una variable X_j puede tener un efecto sobre Y distinto combinado con otras variables, hay entonces que eliminar los efectos de las otras variables. Para eso se puede usar el **coeficiente de correlación parcial**.

7.6.3 Coeficiente de correlación parcial

El efecto de una variable X sobre la variable Y puede estar afectado por una tercera variable Z cuando Z tiene efecto sobre X también. El estudio se basa entonces en las dos relaciones del tipo lineal:

$$X = \alpha Z + \vartheta$$

$$Y = \gamma Z + \eta.$$

Una vez eliminada la influencia de la variable Z sobre las variables X e Y se mide solamente a partir de los restos:

$$X - \alpha Z = \vartheta$$

$$Y - \gamma Z = \eta.$$

Definición 7.6.1 El coeficiente de correlación parcial entre X e Y bajo Z constante es el coeficiente de correlación entre los errores ϑ y γ:

$$\rho(X, Y|Z) = Corr(\vartheta, \gamma)$$

Se observa que si X y Z son muy correlacionados entonces la correlación parcial entre X e Y es muy pequeña. En efecto X aporta casi ninguna información nueva sobre Y (o vice-versa) cuando Z es conocida.

Se puede generalizar a más de 2 variables Z_j, $j = 1, 2, \ldots, q$. Si

$$X = \sum_{j=1}^{q} \alpha_j Z_j + \vartheta \quad Y = \sum_{j=1}^{q} Z_j + \gamma$$

entonces se define el coeficiente de correlación parcial entre X e Y, dadas las variables Z_j, por:

$$\rho(X, Y|Z_1, Z_2, \ldots, Z_q) = Corr(\vartheta, \gamma).$$

Si las variables Z_j no tienen efecto sobre X e Y, es decir, las correlaciones $Corr(X, Z_j)$ y $Corr(Y, Z_j)$ son todas nulas, entonces $\rho(X, Y|Z_1, Z_2, \ldots, Z_q) = Corr(X, Y)$.

Se generaliza también la matriz de correlación parcial con más de dos variables. Definimos para eso la matriz de varianza-covarianza del vector X dado el vector Z fijo:

$$Var(X|Z) = \Gamma_{XX} - \Gamma_{XZ}\Gamma_{ZZ}^{-1}\Gamma_{ZX}.$$

Se tiene una intertretación geométrica del coeficiente parcial $\rho(X, Y|Z)$ mediante los triángulos esféricos: El ángulo (A) del triángulo esférico(ABC) está definido por el ángulo entre las dos tangentes en A a los lados del triángulo esférico (Gráfico **??**). El ángulo (A) es entonces igual a la proyección del ángulo entre OX y OY sobre el plano ortogonal a OZ. Los ángulos siendo relacionados a los arcos, se tiene:

$$\cos(A) = \frac{\cos(a) - \cos(b)\cos(c)}{\sin(b)\sin(c)}.$$

Luego:

$$\rho(X, Y|Z) = \frac{Corr(X, Y) - Corr(X, Z)Corr(Y, Z)}{\sqrt{1 - Corr^2(X, Z)}\sqrt{1 - Corr^2(Y, Z)}}$$

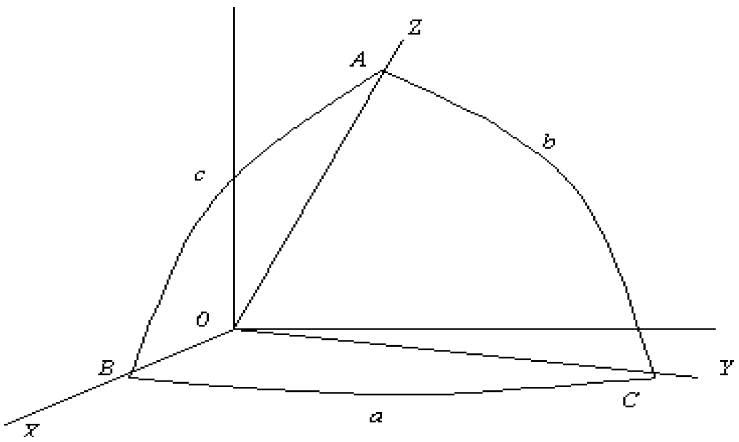


Figura 7.2: Representación esférica del coeficiente de correlación parcial

7.6.4 Efecto de un grupo de variables

Vimos que el efecto global de todas las variables explicativas y los efectos individuales. Veremos aquí el efecto de un grupo de k variables, sean $X_{j_1}, X_{j_2}, \ldots, X_{j_k}$ $(k \leq p)$, entre las p variables. El efecto de estas variables se mide considerando la hipótesis nula $H_0 : \beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0$ contra $H_1 : E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.

Sean $X_{j_{k+1}}, X_{j_{k+2}}, \ldots, X_{j_p}$ el restante de las P variables. Bajo H_0, el modelo se escribe: $Y = \gamma_0 + \gamma_{j_{k+1}} X_{j_{k+1}} + \cdots + \gamma_{j_p} X_{j_p} + \varepsilon_0$. Se tiene la varianza residual bajo H_1 menor que la varianza residual bajo H_0:

$$\sum_i \hat{\varepsilon}_i^2 \leq \sum_i \hat{\varepsilon}_{0,i}^2$$

Se puede estudiar el cociente de las dos varianzas residuales $\dfrac{\sum_i \hat{\varepsilon}_{0,i}^2}{\sum_i \hat{\varepsilon}_i^2}$ o su complemento $\dfrac{\sum_i \hat{y}_{0,i}^2}{\sum_i \hat{\varepsilon}_i^2}$

en donde $\hat{y_{0,i}} = y_i - \hat{\varepsilon}_{0,i}^2$ son las componentes del estimador $E(Y|X)$ bajo H_0.

Bajo la hipótesis H_0

$$Q = \frac{\dfrac{\sum_i (\hat{y}_i - \hat{y}_{0,i})^2}{k}}{\dfrac{\sum_i \hat{\varepsilon}_i^2}{n-r}} \sim F_{k,n-r}.$$

Lo que conduce a un test de región crítica de la forma $Q \geq c_\alpha$.

Considerando otra forma de escribir el problema. Sea la hipótesis nula $H_0 : E(Y) = X_0\beta \in W_0$, con X_0 de rango s, contra $H_1 = X\beta \in W$.

La hipótesis H_0 equivale a $(X - X_0)\beta = 0$ lo que corresponde a $k = p - s + 1$ ecuaciones independientes $\underbrace{D}_{k\times(p+1)}\beta = 0$, en que D es de rango k. Para que el test tenga sentido, $D\beta$ tiene que ser estimable, es decir que el estimador $D\beta$ no debe depender de una solución particular $\hat{\beta}$ de las ecuaciones normales.

Sean $\hat{Y}$ e Y^* las proyecciones Y sobre W y W_0 respectivamente y $E(Y) = \mu_0$ bajo H_0 y $E(Y) = \mu$ bajo H_1.

$$\|Y - \mu_0\|^2 = \|Y - Y^* + Y^* - \mu_0\|^2 = \|Y - Y^*\|^2 + \|Y^* - \mu_0\|^2$$

$$\|Y - \mu\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \mu\|^2$$

Sean $S^2 = \dfrac{\|Y - Y^*\|^2}{\|Y - \hat{Y}\|^2}$ y $R^2 = \dfrac{\|\hat{Y} - Y^*\|^2}{\|Y - \hat{Y}\|^2}$. Bajo H_0, se tiene $\dfrac{n-p-1}{k}R^2 \sim F_{k,n-r}$. La región crítica es de la forma $\dfrac{n-r}{k}R^2 > C$.

Se puede plantear el test de razón de verosimilitudes también: $\Lambda = \dfrac{\max_{H_0} L}{\max L}$. La región crítica se escribe $S > C'$ Este test coincide con el test F.

Se observará que $\frac{\|Y-Y^*\|^2}{n-s}$ y $\frac{\|\hat{Y}-Y^*\|^2}{k}$ son ambos estimadores insesgados de σ^2 bajo H_0.

Cuando la varianza σ^2 es conocida, la razón de verosimilitudes es igual a:

$$\Lambda = \frac{\max_{H_0} L}{\max L} = \exp\left\{-\frac{1}{2\sigma^2}\|\hat{Y}-y^*\|^2\right\}.$$

La región crítica del test se escribe entonces $\|\hat{Y}-Y^*\|^2 > \sigma^2\chi^2_k$. Se puede construir un test a partir de $D\hat{\beta} \sim N(D\beta, \sigma^2\Gamma)$, en que Γ depende solamente de D y X. Bajo H_0, $\frac{\hat{\beta}^t D^t \Gamma^{-1} D\hat{\beta}}{\sigma^2} \sim \chi^2_k$. Pero este test no equivale en general al test de razón de verosimilitudes basado en $\|\hat{Y}-Y^*\|^2$.

7.7 HIPÓTESIS LINEAL GENERAL

Sea la hipótesis nula $H_0 : A\beta = c$ contra la hipótesis alternativa $H_1 : A\beta \neq c$, en donde $A \in M_{k,p+1}$ es conocida y de rango k. $A\beta$ tiene que ser estimable, es decir no debe depender de una solución de las ecuaciones normales. Se supondrá aquí un modelo de rango completo.

Sea $\hat{\beta} = (X^tX)^{-1}X^tY$ el estimador de máxima verosimilitud sin restricción y $\hat{\beta}_0$ el estimador bajo $H_0 : A\beta = c$. Se obtiene $\hat{\beta}_0$ usando los multiplicadores de Lagrange:

$$Q = (Y - X\beta)^t(Y - X\beta) + 2\lambda(A\beta - c)$$

$$\frac{\partial Q}{\partial \beta} = 0 \Rightarrow X^tX\hat{\beta}_0 = X^tY + A^t\Lambda \Rightarrow \hat{\beta}_0 = (X^tX)^{-1}(X^tY + A^t\lambda) = \hat{\beta} + (X^tX)^{-1}A^t\lambda.$$

Utilizando la restricción $A\hat{\beta}_0 = c$, obtenemos que $\lambda = [A(X^tX)^{-1}A^t]^{-1}(c - A\hat{\beta})$

$$\Rightarrow \hat{\beta}_0 = \hat{\beta} + (X^tX)^{-1}A^t[A(X^tX)^{-1}A^t]^{-1}(c - A\hat{\beta})$$

Sean P_0 y P los proyectores asociados respectivamente a $X\hat{\beta}_0$ y $X\hat{\beta}$, es decir tales que $P_0Y = X\hat{\beta}_0$ y $Py = X\hat{\beta}$. Entonces

$$P_0Y = PY + X(X^tX)^{-1}A^t[A(X^tX)^{-1}A^t]^{-1}(c - A\hat{\beta}).$$

Sea la varianza residual del modelo sin restricción: $V = (Y - X\hat{\beta})^t(Y - X\hat{\beta})$ y la varianza residual bajo $H_0 : T = (Y - X\hat{\beta}_0)^t(Y - X\hat{\beta}_0)$. Como $T \geq V$,consideramos $U = T - V$ que compararemos a V.

Proposición 7.7.1 *La diferencia de las varianzas residuales con y sin restircción es:*

$$U = (A\hat{\beta} - c)^t[A(X^tX)^{-1}A^t]^{-1}(A\hat{\beta} - c)$$

y bajo la hipótesis nula $\frac{U}{\sigma^2} \sim \chi^2_k$.

Demostración

$$U(Y - X\hat{\beta}_0)^t(Y - X\hat{\beta}_0) - (Y - X\hat{\beta})^t(Y - X\hat{\beta}) = Y^t(P - P_0)Y.$$

Como $P_0Y = PY + X(X^tX)^{-1}A^t[A(X^tX)^{-1}A^t]^{-1}(c - A\hat{\beta})$ y $U = Y^t(P - P_0)^t(P - P_0)Y \Rightarrow$ $U = (A\hat{\beta} - c)^t[A(X^tX)^{-1}A^t]^{-1}(A\hat{\beta} - c)$.

Por otro lado como A es de rango igual a k, $A\hat{\beta} \sim N_k(A\beta, \sigma^2 A(X^tX)^{-1}A^t)$, luego $\dfrac{U}{\sigma^2} \sim \chi^2_k$.
■

Como $\hat{\beta}$ es independiente de $V = \sum_i \hat{\varepsilon}_i^2$, el estadístico del test es:

$$\frac{U/k}{V/(n-p)} \sim F_{k,n-p}$$

7.8 ANÁLISIS DE LOS RESIDUOS

Se supone que el efecto de numerosas causas no identificadas está contenido en los errores, lo que se traduce como una perturbación aleatoria. De aquí los supuestos sobre los errores, que condicionan las propiedades del estimador. Es importante entonces comprobar si los supuestos se cumplen.

La mejor forma de chequear si los errores son aleatorios de medias nulas, independientes y de la misma varianza, consiste en estudiar los residuos

$$\forall i = 1, 2, \ldots, n : \hat{\varepsilon}_i = y_i - \sum_j \hat{\beta}_j x_{i,j}$$

considerndolos como muestra i.i.d. de una distribución normal.

Se puede usar el gráfico $(Y_i, \hat{\varepsilon}_i)$, que debería mostrar ninguna tendencia de los puntos, o bien construir test de hipótesis sobre los errores. En el gráfico de la izquierda (gráfico 7.3) se puede ver los residuos aleatorios independientes de Y, lo que no es el caso de los residuos del gráfico de la derecha.

Si el supuesto que los errores son $N(0, \sigma^2)$ no se cumple, tenemos que estudiar el efecto que esto tiene sobre la estimación de los parámetros y sobre los tests de hipótesis, además tenemos que detectar si este supuesto es cierto o no y corregir eventualmente la estimación de los parámetros y tests.

Vimos donde interviene el supuesto de normalidad en la estimación de los parámetros del modelo y en los tests de hipótesis para verificar la significación de las variables en el modelo. Este tema se relaciona con el concepto de la *robustez* (ver MILLER[9]).

La teoría de estimación y de test de hipótesis se basa en supuestos sobre la distribución de población. Por lo tanto si estos supuestos son inexactos, la estimación o la conclusión del test sera distorsionada. Se buscan entonces métodos que sean lo menos sensibles a la inexactitud de los supuestos. Se habla entonces de robustez del método.

Se divide el estudio en tres partes: la normalidad, la independencia y la igualdad de las varianzas de los errores.

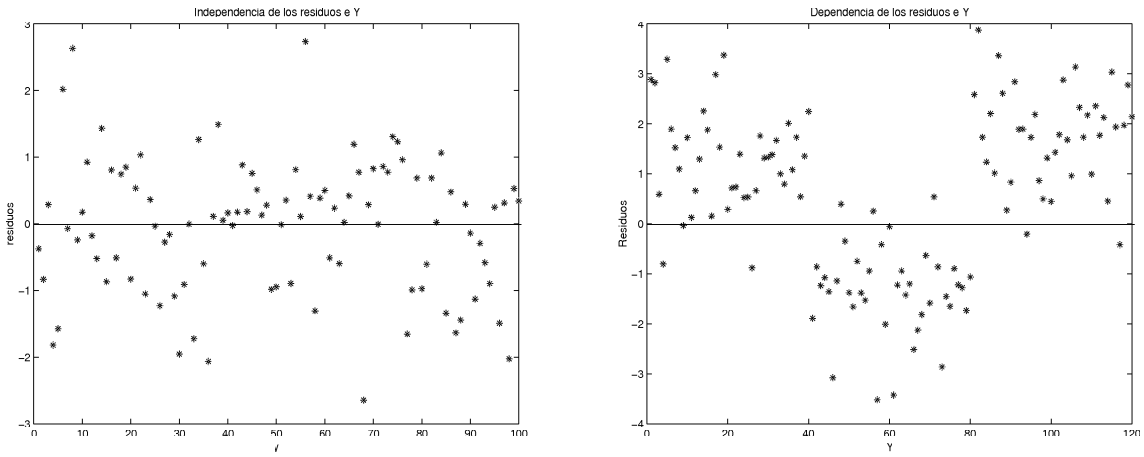


Figura 7.3: Gráficos de residuos

7.8.1 Estudio de la normalidad de los errores

Si no se cumple la normalidad de los errores, los efectos sobre la estimación o tests relativos a los parámetros son pequeos, pero son más importantes sobre los tests relativos a coeficiente de correlación. El problema es más agudo en presencia de observaciones atípicas.

Tenemos entonces que verificar la hiptesis nula $H_o : \varepsilon_i \sim N(0, \sigma^2)$ o sea si $u_i = \frac{\varepsilon_i}{\sigma}$, $H_o : u_i \sim N(0, 1)$. Esto sugiere de comparar la función de distribución empírica F_n de los residuos normalizados con la función de distribución de la $N(0, 1)$. Sea F la función de distribución de la $N(0, 1)$, que es invertible.

Entonces si los u_i provienen de $N(0, 1)$, $F^{-1}(F_n(u_i)) \approx u_i$. Consideramos entonces los estadísticos de orden de los u_i, que son los residuos normalizados ordenados de menor a mayor: sea $u_{(1)} \leq u_{(2)} \leq ... \leq u_{(n)}$. La función de distribución empírica es entonces:

$$F_n(u) = \frac{card\{u_{(i)} \leq u\}}{n}$$

Se define los cuantiles empíricos $q_i = F^{-1}(F_n(u_{(i)}$. Si F_n se parece a F, los puntos (u_i, q_i) deberían ser colineales (sobre la primera bisectriz). Este gráfico se llama *probit* o *recta de Henri* (gráfico 7.4).

Si los puntos en ell gráfico probit aparecen como no lineal, se rechaza la normalidad de los errores y se puede corregir utilizando la regresión no paramétrica basada o bien otras alternativas según la causa de la no normalidad (no simetría, observaciones atípicas, etc..

7.9 PREDICCIÓN

Si se tiene una nueva observación para la cual se conocen los valores de las variables explicativas, sean $x_{0,1}, x_{0,2}, \ldots, x_{0,p}$, pero se desconoce el valor Y_0 de la variables respuesta, se puede

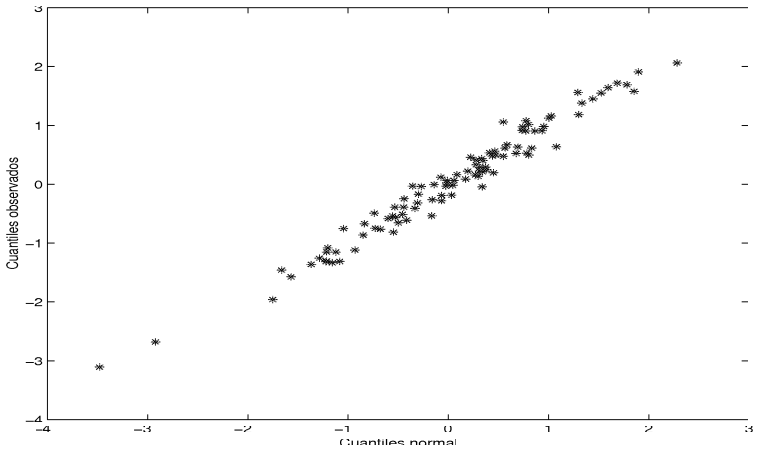


Figura 7.4: Recta de Henri

entonces usar el modelo para inferir un valor para Y_0 a través de su modelo esperado:

$$\mu_0 = E(y_0) = x_0^t \beta$$

en que $x_0^t = (x_{0,1}\, x_{0,2} \ldots x_{0,p})$.

Si $\hat{\beta}$ es el estimador de β obtenido sobre las antiguas observaciones, se estima μ_0 dados los valores tomados por las variables explicativas por:

$$\hat{\mu}_0 = E(y_0) = x_0^t \hat{\beta}.$$

Se puede calcular un intervalo de confianza para μ_0: la distribución de $\hat{y}_0$ es $N(\mu_0, \sigma^2 x_0^t (X^t X)^{-1} x_0)$, luego $\dfrac{\hat{y}_0 - \mu_0}{\tilde{\sigma}\sqrt{x_0^t (X^t X)^{-1} x_0}} \sim t_{n-p-1}$. Se usa este estadístico para construir un intervalo de confianza de nivel $1-\alpha$ para μ_0:

$$I\!P\left(\hat{y}_0 - t_{n-p-1}^{\alpha/2} \tilde{\sigma}\sqrt{x_0^t (X^t X)^{-1} x_0} \leq \mu_0 \leq \hat{y}_0 + t_{n-p-1}^{\alpha/2} \tilde{\sigma}\sqrt{x_0^t (X^t x)^{-1} x_0}\right) = 1 - \alpha$$

Un problema distinto es de estimar un intervalo para y_0. Hablamos de un intervalo para la predicción. En este caso hay que tomar en cuenta de la varianza aleatoria y_0:

$$y_0 = \hat{y}_0 + \hat{\varepsilon}_0.$$

La varianza de $\hat{\varepsilon}_0$ es igual a: $\sigma^2 + \hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0$, dado que $\hat{y}_0$. Un intervalo de predicción para y_0 se obtiene entonces a partir de $\dfrac{y_0 - \hat{y}_0}{\hat{\sigma}\sqrt{1 + (x_0^t (X^t X)^{-1} x_0)}} \sim t_{n-p-1}$

El intervalo es entonces definido por:

$$I\!P\left(\hat{y}_0 - t_{n-p-1}^{\alpha/2} \hat{\sigma}\sqrt{1 + x_0^t (X^t X)^{-1} x_0} \leq y_o \leq \hat{y}_0 + t_{n-p-1}^{\alpha/2} \hat{\sigma}\sqrt{1 + x_0^t (X^t X)^{-1} x_0}\right) = 1 - \alpha.$$

7.10 EJERCICIOS

1. Cuatro médicos estudian los factores que explican la espera de los pacientes en la consulta. Toman una muestra de 200 pacientes y consideran el tiempo de espera de cada uno el día de la consulta, la suma de los atrasos de los médicos a la consulta este mismo día, el atraso del paciente a la consulta este día (todos estos tiempos en minutos) y el números de médicos que están al mismo tiempo es la consulta este día. Se encuentra un tiempo promedio de espera de 32 minutos con una desviación típica de 15 minutos. Se estudia el tiempo de espera en función de las otras variables mediante un modelo lineal cuyos resultados están dados a continuación:

Variable	**Coeficiente**	**Desv. típica**	**t-Student**	$I\!P(\|X\| > t)$
Constante	22,00	4,42	4,98	0,00
Atraso médico	0,09	0,01	9,00	0,00
Atraso paciente	-0,02	0,05	0,40	0,66
Número de médicos	-1,61	0,82	1,96	0,05
Coef. determinación=0,72	F de Fisher=168	$I\!P(X > F) = 0,000$		

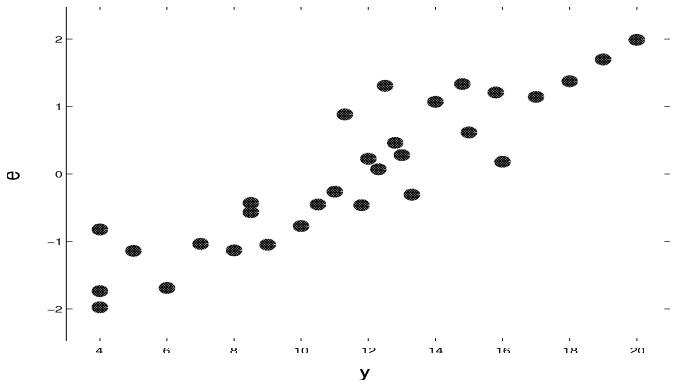
a) Interprete los resultados del modelo lineal. Comente su validez global y la influencia de cada variable sobre el tiempo de espera. Especifique los grados de libertad de las t de Student y la F de Fisher.
b) Muestre que se puede calcular la F de Fisher a partir del coeficiente de determinación. Si se introduce una variable explicativa suplementaria en el modelo, ¿el coeficiente de determinación será más elevado?
c) Dé un intervalo de confianza a 95% para el coeficiente del atraso médico.
d) Predecir el tiempo de espera, con un intervalo de confianza a 95que llega a la hora un día que el consultorio funciona con 4 médicos que tienen respectivamente 10, 30, 0, 60 minutos de atraso.

2. Consideramos el modelo lineal $Y = X\beta + \varepsilon$ con $\varepsilon \sim N_n(0, \sigma^2 I_n)$, $\beta \varepsilon I\!R^{p+1}$, $X \varepsilon M_{n,p+1}(I\!R)$.
a) Escribamos X como: $X = (X_1 \quad X_2)$, con X_1 y X_2 submatrices de X tales que $X_1^t X_2 = 0$ (la matriz nula). El modelo inicial $Y = X\beta + \varepsilon$ se escribe $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ con $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$. Si $\hat{\alpha}_1$ es el estimador de máxima verosimilitud de α_1 en el modelo $Y = X_1\alpha_1 + \varepsilon$ y $\hat{\alpha}_2$ es el estimador de máxima verosimilitud de β es igual a $\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix}$.

(Indicación: se usará el siguiente resultado: si $A \in M_{n,n}(I\!R)$ es una matriz diagonal por bloque, i.e. $A^{-1} = \begin{bmatrix} A_1^{-1} & 0 \\ 0 & A_2^{-1} \end{bmatrix}$, con las submatrices A_1 y A_2 invertibles , entonces A es invertible, y $A^{-1} = \begin{vmatrix} A_1 & 0 \\ 0 & A_2 \end{vmatrix}$).

b) Si $X_1^t X_2 \neq 0$ y si se toma $\hat{\beta} = \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix}$ como estimador de β, que propiedad pierde $\hat{\beta}$ bajo el supuesto usual $E(\varepsilon) = 0$.

3. Consideremos tres variables Y, X, Z observadas sobre una muestra de tamaño $n = 40$. $\{(y_i, x_i, z_i) \quad tq \quad i = 1, \ldots, 40\}$. Se busca explicar Y linealmente a partir de X y Z.



a) Se representan los resultados de modelo lineal: $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, \ldots, 40$:

Variable	Medias	Desv. típica	Estimación	Dev. típ. estimación	t-Student	$I\!P(\|X\| > t)$
Y	11,68	3,46				
Constante			7,06	1,03	6,84	0,00
X	5,854	2,74	0,79	0,16	4,94	0,00

Coef. determinación=0,39	F de Fisher=24,44	$I\!P(X > F) = 0,000$

Interprete estos resultados y efectúe el test de hipótesis $H_0 : \beta = 0$.
b) Dé una estimación insesgada para σ^2 la varianza de los errores de este modelo.
c) Comente el gráfico de los residuos en función de los y_i.
d) Se tiene una nueva observación que toma sobre la variable X el valor $x_0 = 6,50$. Dé una estimación $\hat{y}_0$ del valor y_0 que toma sobre la variable Y.
e) Se presentan los resultados del modelo lineal: $y_i = \delta + \gamma z_i + \varepsilon_i$:

Variable	Medias	Desv. típica	Estimación	Dev. típ. estimación.	t-Student	$I\!P(\|X\| > t)$
Y	11,68	3,46				
Y	11,68	3,46				
Constante			11,68	0,36	32,54	0,00
Z	0,00	2,65	1,00	0,14	7,27	0,00

Coef. determinación=0,58	F de Fisher=52,78	$I\!P(X > F) = 0,000$

Se tiene $\sum_i x_i z_i = 0$ y $\sum_i z_i = 0$.

Muestre que si $X_1 = (1_n|X)$ es una matriz formada del vector de unos y del vector de los x_i y $X_2 Z$ el vector formado de los z_i, se tiene $X_1^t X_2 = 0$. Usando los resultados del ejercicio 2 deduzca las estimaciones de los parámetros del modelo $y_i = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$.

4. Se requiere ajustar una función escalón $y = f(t)$ con f constante en los intervalos en que $j = 0, \ldots, K$ y $a_0 < a_1 < \cdots < a_K$. Para ello se observan datos $\{(t_i, y_i) \quad i = 1, \ldots, n\}$. Se asume que los y_i son mutuamente independientes y que la distribución de los y_i es $N(f(t_i), \sigma^2)$.

a) Formule el problema anterior como modelo lineal.
b) Obtenga la función ajustada por mínimos cuadrados.
c) Concluya un intervalo de confianza para $\int_{a_0}^{a_K} f(t)dt$.

5. Sea $Y \in \mathbb{R}^n$ un vector aleatorio con $E(Y) = \mu$ y $Var(Y) = \sigma^2 I_n$. Se considera el modelo lineal $Y = X\beta + \varepsilon$, en que $X \in M_{n,p}$ es de rango completo. Llamaremos W al subespacio de $\mathbb{R}^n$ conjunto imagen de X e $\hat{Y}$ al estimador de mínimos cuadrados de $\mu = E(Y)$.

a) Sea $a \in W$ y Δ_a la recta generada por a. Se define $H_0 = \{z \in W \quad tq \quad a^t z = 0\}$ el suplemento ortogonal de Δ_a en W. Se tiene entonces la descomposición en suma directa ortogonal de W: $W = H_a \oplus \Delta_a$. Muestre que el estimador de mínimos cuadrados Y^* de μ en H_a se escribe como: $Y^* = \hat{Y} - \left(\frac{a^t\hat{Y}}{a^t a}\right) a$.

b) Si $b \in \mathbb{R}^n$, muestre que $Var(b^t Y^*) = Var(b^t\hat{Y}) - \sigma^2\frac{(b^t b)^2}{a^t a}$.

c) Suponiendo que los errores son normales, dé la distribución de $\frac{\sum_i \varepsilon_i^{*2}}{\sigma^2}$, en que $\varepsilon_i^* = Y_i - Y_i^*$.

d) Se considera el caso particular $a = I_n$. Dé la distribbución de $\frac{\sum_i Y_i^{*2}/p}{\sum_i \varepsilon_i^{*2}/(n-p)}$. Muestre que si las variables son centradas, $\hat{Y} = Y^*$.

6. Teorema de Gauss-Markov generalizado. Si $Var(Y) = \Gamma$, Γ invertible, entonces el estimador $\hat{\beta}$ insesgado de mínima varianza entre los estimadores lineales insesgados de β es aquel que minimiza $\|Y - X\beta\|^2_{\Gamma^{-1}}$.

a) Encuentre el estimador de máxima verosimilitud de β y Γ.
b) Demuestre el teorema.
c) Si el rango de X es igual a r, muestre que la norma del vector de residuos de un modelo lineal

$$\|Y - \hat{Y}\|^2_{\Gamma^{-1}} \sim \chi^2_{n-r}$$

en donde $\hat{Y}$ la proyección Γ^{-1}-ortogonal de Y sobre $Im(X)$.

7. Sea el modelo lineal: $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$, $i = 1, 2, \ldots, n$. Matricialmente $Y = X\beta + \varepsilon$, con $rango(X) = p + 1$, $E(\varepsilon) = 0$, $Var(\varepsilon)\sigma^2 I_n$.

a) Se escribe $X^t X = \begin{bmatrix} n & a^t \\ a & V \end{bmatrix}$. Dé las expresiones de a y V. Muestre que V es definida positiva. Muestre que a es un vector nulo cuando las variables explicativas están centradas $\left(\forall j : \sum_{i=1}^n x_{i,j} = 0\right)$. Relacione los valores propios de V con los de V^{-1}.

b) Muestre que $\sum_j Var(\hat{\beta})$ sujeto a $\forall j : \sum_{i=1}^n x_{i,j} = 0$ y $\forall j : \sum_{i=1}^n x_{i,j}^2 = c$ (c es una constante positiva) alcanza su mínimo cuando $X^t X$ es diagonal. c) En qué difieren de las propiedades optimales obtenidas en el teorema de Gauss-Markov?

d) Se supone que X^tX es diagonal con $\forall j : \sum_{i=1}^{n} x_{i,j} = 0$ y $\forall j : \sum_{i=1}^{n} x_{i,j}^2 = c$. Deducir las expresiones de $\hat{\beta}$, $Var(\hat{\beta})$, $\hat{Y}$. Exprese el coeficiente de correlación múltiple R^2 en función de los coeficientes de correlación lineal de Y con las variables explicativas X.

8. Sea el modelo lineal $Y = X\beta + \varepsilon$, con X de rango completo pero X^tX no diagonal.

a) Dé la expresión de una predicción de la variable respuesta Y y un intervalo de confianza asociado.
b) Se hace un cambio de base de las columnas de X, sea Z la matriz de las nuevas columnas, de manera que $Im(X) = Im(Z)$ y que Z^tZ sea diagonal. Muestre que el cambio de variables explicativas no cambia las predicciones de Y. Deduzca la expresión del intervalo de confianza en función de Z.

Apéndice A

BIBLIOGRAFÍA

1| BENJAMIN, *Probabilidad y Estadística en Ingeniería Civil*, Mc Graw Hill LatinoAmericana, 1981.

2| CRAMER H., *Mathematical Methods of Statistics*, Princeton University Press, 1961.

3 |DEGROOT M., *Probabilidad y Estadística*, Addison-Wesley, 1987.

4| DAVID, *Order Statistics*, Wiley, 1970.

5| GILBERT (1981), Estadística, Interamericana.

6| KENDALL M.G. , STUART A. (1966), The Advanced Theory of Statistics, Lossey-Bass.

7| LACOURLY N. (2000), Apuntes de Estadística, curso a profesores de matemática de enseñanza media. http://ideamas.cl

8| LEBART L. Et al. (1979) , Traitement des Données Statistiques, Dunod.

9| MILLER G. (1986), Beyond ANOVA, Basics and Applied Statistics, Wiley.

10| MOOD A., GRAYBILL F., Introducción a la Estadstica Matemática. Aguilar.

[11| MOSTELLER, TUKEY, Data Analysis and Regression, Addison-Wesley.

12| RICHARD J. (1997), Probabilidad y Estadstica para Ingenieros de Miller Freund, Printice Hall.

13| SAPORTA G. (1990), Probabilité, Analyse des Données et Statistique, Technip.

Apéndice B

CORRECCIÓN DE LOS EJERCICIOS

Capítulo 3

1. La función de densidad conjunta de la muestra es

$$L(x_1, ..., x_n/\alpha, \beta) = \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} e^{-\beta \sum x_i} \prod_{i=1}^{n} x_i^{\alpha-1}$$

$$\log(L(x_1, .., x_n/\alpha, \beta)) = n\alpha \log(\beta) - n \log(\Gamma(\alpha)) - n\bar{(x)}\beta + (\alpha - 1) \sum_{i=1}^{N} \log(x_i)$$

Las condiciones necesarias para que el par $(\hat{\alpha}, \hat{\beta})$ sea un punto de máximo son

$$\frac{\partial L(.)}{\partial \alpha} = 0 \iff \log(\hat{\beta}) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = \frac{1}{n} \sum_{i=1}^{n} \log(x_i)$$

y

$$\frac{\partial L(.)}{\partial \beta} = 0 \iff \frac{\hat{\alpha}}{\hat{\beta}} = \bar{x}$$

Recordemos ahora el Principio de Invarianza. Si $T : \Omega_1 \to \Omega_2$ es una transformación **biyectiva** y $\hat{\theta}$ el estimador de M.V. de θ, entonces el estimador de M.V. de $T(\theta)$ es $T(\hat{\theta})$.
Para aplicar esto aquí hacemos $\Omega_1 = \Omega_2 = \{(\alpha, \beta)\alpha > 0, \beta > 0\}$ y definimos la transformación T como $T(\alpha, \beta) = (\alpha, \frac{\alpha}{\beta})$
Notando que el Jacobiano de T es no nulo en Ω_1 se deduce que T es biyectiva [1]. Con esto

$$\frac{\hat{\alpha}}{\hat{\beta}} = \bar{x}$$

[1] Teorema de la Función Inversa

Como E(X)=$\frac{\alpha}{\beta}$ y E($\bar{x}$) = $\frac{\alpha}{\beta}$, el estimador es insesgado.
Además

$$Var(\bar{x}) = \frac{1}{n}Var(X) = \frac{1}{n}\frac{\alpha}{\beta^2}$$

Entonces $\lim_{n\to+\infty} Var(\bar{x}) = 0$. Con esto y la desigualdad de Chebyshev

$$I\!P(|\bar{x} - \frac{\alpha}{\beta}| > \epsilon) \leq \frac{Var(\bar{x})}{\epsilon^2} \quad \forall \epsilon > 0$$

se tiene que $\bar{x}$ es consistente.
Como $E(|\bar{x} - \frac{\alpha}{\beta}|^2) = Var(\bar{x})$, $\bar{x}$ converge en media cuadrática.

2. $L(x_1, ...x_n/\theta) = \theta^n \prod_{j=1}^n x_j^{\theta-1}$ entonces, $\log(L(x_1, ...x_n/\theta)) = n\log(\theta) - n + \sum_{j=1}^n \log(x_j)$ de donde $\hat{\theta} = \frac{n}{\sum_{j=1}^n -\log(x_j)}$
Notemos ahora que $z_j = -\log(x_j)$ tiene una densidad Exponencial [2] de parámetro θ, y que calcular la esperanza de $\hat{\theta}$ se reduce a calcular $E(\frac{1}{Z})$ con Z la suma de n v.a. Exponenciales de parámetro θ. Es fácil mostrar, usando funciones características, que Z tiene una densidad Gamma(n,θ). Así

$$E(\frac{1}{Z}) = \frac{1}{\Gamma(n)} \int_0^\infty Z^{-1}\theta^n Z^{n-1} e^{-\theta} dZ$$

usando la propiedad $\Gamma(n) = n\Gamma(n-1)$, se obtiene

$$\frac{\theta^2}{n(n-1)\Gamma(n-2)} \int_0^\infty \theta^{n-2} z^{(n-2)-1} e^{-\theta z} dz$$

de donde $E(Z^{-1}) = \frac{n-2\theta}{n(n-2)}$, ya que la integral anterior corresponde a la Esperanza de una densidad Gamma(n-2,θ) y por lo tanto vale $\frac{n-2}{\theta}$. Con esto, E($\hat{\theta}$) = $\frac{n-2}{n-1}\theta$ i.e. $\hat{\theta}$ es sesgado. (Subestima el valor de θ). Notando que $\lim_{n\to\infty} \frac{n-2}{n-1} = 1$, se deduce que el estimador es asintoóticamente insesgado.
Además, $Var(\hat{\theta}) = E(\hat{\theta}^2 - E(\hat{\theta}))^2 = E(\hat{\theta}^2 - \theta)^2$ y, haciendo un cálculo análogo al anterior, resulta que $E(\hat{\theta}^2) = \frac{n\theta}{n(n-1)}$. De esto se deduce que $Var(\hat{\theta}) = \theta(\frac{n^2}{n(n-1)} - 1)$. Con esto, $\lim_{n\to+\infty} Var(\hat{\theta}) = 0$ y , por lo tanto, convergente en media cuadrática. Esto implica que $\hat{\theta}$ es consistente.

3. a) $I\!P(SI) = I\!P(Q)I\!P(SI/Q) + I\!P(Q')I\!P(SI/Q') \Longrightarrow p = \theta\pi + (1-\theta)(1-\pi)$
b) Sea Y el número de personas que contestan SI; $Y \sim Bin(n, p)$.
El EMV de p es $\hat{p} = y/n \Longrightarrow \hat{\pi} = \frac{1-\hat{p}-\theta}{1-2\theta}$
$E(\hat{\pi}) = \frac{1-p-\theta}{1-2\theta} = \pi$ y $Var(\hat{\pi}) = \frac{p(1-p)}{n(1-2\theta)^2} = \frac{\pi(1-\pi)}{n} + \frac{\theta(1-\theta)}{n(1-2\theta)^2}$
c) El estimador $\hat{\pi}$ es insesgado y su varianza converge a 0, luego es consistente.
La varianza de $\hat{\pi}$ es máxima cuando $\theta = 0.5$; de toda manera $\hat{\pi}$ no esta definido en esta caso.

[2] Usar el Teorema del Cambio de variables o calcular directamente $I\!P(-\log(x_i) < C)$

4. a) $I\!P(X=x)=1 \Longrightarrow \sum_{x=0}^{\infty}\frac{a_x\theta^x}{h(\theta)}=1 \Longrightarrow h(\theta)=\sum_{x=0}^{\infty}a_x\theta^x$

$\Longrightarrow h'(\theta)=\sum_{x=0}^{\infty}xa_x\theta^{x-1}=\frac{1}{\theta}\sum_{x=0}^{\infty}xa_x\theta^x$

b) $f_n(\underline{x}/\theta)=\prod_i^n a_{x_i}\frac{\theta^{\sum x_i}}{(h(\theta))^n}$, luego $\lambda_n(\underline{x}/\theta)=logf_n(\underline{x}/\theta)=\sum loga_{x_i}+\sum x_i log\theta - nlogh(\theta)$

$\lambda_n'(\underline{x}/\theta)=\frac{\sum x_i}{\theta}-n\frac{h'(\theta)}{h(\theta)}=0$. El EMV $\hat{\theta}$ de θ cumple: $\bar{x}_n=\hat{\theta}\frac{h'(\hat{\theta})}{h(\hat{\theta}}$

c) $m_1(\theta)=E(X)=\sum xa_x\theta^x/h(\theta)$, luego $E(X)=\theta\frac{h'(\theta)}{h(\theta)}$ y $\bar{x}_n=\hat{\theta}\frac{h'(\hat{\theta})}{h(\hat{\theta})}$

d) Si $X\sim Bin(N,p)$, $I\!P(X=x)=a_x\theta^x/h(\theta)$, con $a_x=\left(\begin{array}{c}N\\x\end{array}\right)$ para $x=0,1,...N$; $\theta=\frac{p}{1-p}$; $h(\theta)=(1+\theta)^N$

El EMV de θ cumple: $\bar{x}=\frac{N\hat{\theta}}{1+\hat{\theta}}=\hat{p}$

Si $X\sim\mathcal{P}(\lambda)$, $I\!P(X=x)=\lambda^xe^{-\lambda}/x!$, luego tomando $a_x=1/x!$ y $\theta=\lambda$, se obtiene $\hat{\lambda}=\bar{x}$.

5. a) $E(T)=\theta \iff \sum_{i=1}^{N}\lambda_i(b_i+\theta)=\theta \iff \sum_{i=1}^{N}\lambda_i=1$ y $\sum_{i=1}^{N}\lambda_ib_i=0$ i.e. T debe ser una combinación convexa de los T_i.

b) Resulta cómodo usar notación vectorial. Sean $\underline{T}=(T_1\ldots T_N)^t$ y $\underline{\lambda}=(\lambda_1\ldots\lambda_N)^t$. Así $T=\underline{\lambda}^t\underline{T}$. entonces $Var(T)=\underline{\lambda}^t\Gamma\underline{\lambda}$, en que $\Gamma_{ij}=Cov(T_i,T_j)$; Γ es la matriz de varianza-covarianza de los T_i. El problema se puede plantear como
$\min_\lambda \underline{\lambda}^t\Gamma\underline{\lambda}$, con la restricción $\sum_{i=1}^{N}\lambda_i=1$. Así es posible calcular los coeficientes óptimos si la matriz de varianza covarianza es conocida.

c) Si los T_i son no correlacionados, la matriz de varianzas-covarianzas es diagonal. Llamemos σ_i^2 al iésimo elemento de la diagonal de Γ ($\sigma_i^2=Var(T_i)$). El problema de minimización anterior se escribe como $\min_{\underline{\lambda}}\sum_{i=1}^{N}\lambda_i^2\sigma_i^2$ sujeto a $\sum_{i=1}^{N}\lambda_i=1$. Si se usan Multiplicadores de Lagrange, el problema se transforma en

$$\min_{(\underline{\lambda},\alpha)} L(\underline{\lambda},\alpha)=\sum_{i=1}^{N}\lambda_i^2\sigma_i^2-\alpha(\sum_{i=1}^{N}\lambda_i-1)$$

La condición de primer orden es : $\frac{\partial L(.)}{\partial\lambda_k}=0 \iff 2\lambda_k=\frac{\alpha}{\sigma_k^2}\ \forall k$

Sumando esta igualdad de k=1 hasta N se obtiene el valor del multiplicador $\alpha=\frac{2}{\sum_{k=1}^{N}\frac{1}{\sigma_k^2}}$

y usando esto se obtiene el valor de λ_k: $\lambda_k=\frac{\alpha}{2\sigma_k^2}=\frac{1}{\sum_{i=1}^{N}\frac{\sigma_k^2}{\sigma_i^2}}$

d) Usaremos lo anterior. Es claro que los s_i^2 son estimadores insesgados de σ^2. Además, como $\frac{(n_i-1)s_i^2}{\sigma^2}$ tiene distribución χ_{n_i-1}, se deduce que $var(s_i^2)=\frac{2\sigma^4}{n_i-1}$

Con esto $\frac{var(s_k^2)}{var(s_i^2)}=\frac{n_i-1}{n_k-1}$

$\sum_{i=1}^{N}\frac{var(s_k^2)}{var(s_i^2)}=\frac{\sum_{i=1}^{N}n_i-N}{n_k-1}$ $\lambda_i=\frac{(n_k-1)}{\sum_{i=1}^{N}n_i-N}$, que son exactamente los coeficientes del estimador S^2.

Capítulo 4

1. a) El intervalo para θ es $I = |\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n}|$ y tiene un largo igual a $3.96\sigma/\sqrt{n} \Longrightarrow \sqrt{n} > 392 \Longrightarrow n > 153664$
b) $I\!P(a \leq \theta \leq b) = 1 - \alpha$, con $L = b - a = \sigma/5$
$I\!P(\frac{\sqrt{n}(\bar{x}-b)}{\sigma} \leq \frac{\sqrt{n}(\bar{x}-\theta)}{\sigma} \leq \frac{\sqrt{n}(\bar{x}-a)}{\sigma} = 1 - \alpha \Longrightarrow \frac{\sqrt{n}(b-a)}{\sigma} = \frac{\sqrt{n}}{5}$

n	10	20	30	100
I	[-0.316, 0.316]	[-0.447, 0.447]	[-0.548, 0.548]	[-1.0, 1.0]
$1-\alpha$	0.43	0.52	0.55	0.76

c) Si σ^2 es desconocido, se toma $L = s/5$ y el estadistico t_{n-1}.

n	10	20	30	100
I	[-0.316, 0.316]	[-0.447, 0.447]	[-0.548, 0.548]	[-1.0, 1.0]
$1-\alpha$	0.25	0.35	0.40	0.70

Cuando n aumenta, el nivel de confianza aumenta también en ambas casos. Pero e el segundo caso el nivel es siempre más pequeño para un n dado.

2. $|36.08, 43.92|$

3. $\bar{x} = 247.3$, $s^2 = (1/n)\sum(x_i - \bar{x})^2 = 2.01$.
a) $\bar{x} \sim \mathcal{N}(\mu, 0.15) \Longrightarrow I\!P(|\frac{(\bar{x}-\mu)}{\sqrt{0.15}}| \leq 1.96) = 0.95$; el intervalo de confianza para μ es igual a $[\bar{x} - 1.96\sqrt{0.15}, \bar{x} + 1.96\sqrt{0.15}] = [246.54, 248.06]$
b) $\frac{\bar{x}-\mu}{s_n/3} \sim t_9$; luego $I\!P(|\frac{(\bar{x}-\mu)}{s_n/3}| \leq 2.26) = 0.95$; el intervalo de confianza para μ es igual a $|\bar{x} - 2.26s_n/3, \bar{x} + 2.26s_n/3| = |246.23, 248.36|$
Se observara que hacer la aproximación normal daría un intervalo más pequeño.
c) $ns^2/\sigma^2 \sim \chi^2_9$.
$I\!P(2.7 \leq ns^2/\sigma^2 \leq 19.02) = 0.95$; luego se obtiene el intervalo de nivel de confianza a 0.95%: $[ns^2/19.02, ns^2/2.7] = [1.057, 7.44]$.

4. $I\!P(a \leq \theta \leq b) = I\!P(\frac{n\bar{x}-nb}{\sqrt{n\theta(1-\theta)}} \leq \frac{n\bar{x}-n\theta}{\sqrt{n\theta(1-\theta)}} \leq \frac{n\bar{x}-na}{\sqrt{n\theta(1-\theta)}})$
$= I\!P(-1.96 \leq \frac{n\bar{x}-n\theta}{\sqrt{n\theta(1-\theta)}} \leq 1.96) = 0.95$
$\Longrightarrow \frac{(n\bar{x}-n\theta)^2}{n\theta(1-\theta)} \leq (1.96)^2$
La inecuación que define el intervalo de confianza es entonces:

$$(3.84n + n^2)\theta^2 - (3.84n + 2n^2\bar{x})\theta + (n\bar{x})^2 \leq 0$$

5. Si t_o es tal que $I\!P(-t_o < t < t_o) = 0.95$, como n se supone grande, $t_o = 1.96$ y el intervalo es $|Y^* - 1.96S^*, Y^* + 1.96S^*|$.

6. a) $\sigma_2 = k\sigma_1$, entonces se puede construir un estadístico que no depende de σ_1^2 y σ_2^2:

$$\frac{\bar{x}_1 - \bar{x}_2 - \mu_1 - \mu_2}{\sqrt{(\frac{n_1\hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2/k^2}{n_1+n_2-2})(\frac{k^2n_1+n_2}{n_1n_2})}} \sim t_{n_1+n_2-2}$$

Si $v = \sqrt{(\frac{n_1\hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2/k^2}{n_1+n_2-2})(\frac{k^2n_1+n_2}{n_1n_2})}$, el intervalo es entonces igual a:

$$[\bar{x}_1 - \bar{x}_2 - t^{\alpha}_{n_1+n_2-2}v, \bar{x}_1 - \bar{x}_2 + t^{\alpha}_{n_1+n_2-2}v]$$

b) Las cotas del intervalo convergen en probabilidad hacia $\mu_1 - \mu_2$.
c) El método de los momentos produce un estimador de k igual a $\hat{k} = \frac{\sigma_2}{\sigma_1}$.

7. a) $C_a(x) = \begin{cases} [\text{x - a, 0}] & \text{si } x < -a \\ [\text{x-a, x+a}] & \text{si } -a < x < a \\ [0, \text{x+a}] & \text{si } x > a \end{cases}$

$I\!P(\mu \in C_a(x)/\mu) = I\!P(-a < \mu < 0/\mu = 0 \wedge x < -a)I\!P(x < -a) + I\!P(x - a < \mu < x + a/\mu = 0 \wedge -a < x < a)I\!P(-a < x < a) + I\!P(0 < \mu < x + a/\mu = 0 \wedge x > a)I\!P(x > a) =$
$I\!P(x < -a) + I\!P(-a < x < a) + I\!P(x > a) = 1$

b) $I\!P(\mu \in C_a(x))/x) = \begin{cases} I\!P(\mu \in [x - a, 0]) & \text{si } x < -a \\ I\!P(\mu \in [x - a, x + a]) & \text{si } -a < x < a \\ I\!P(\mu \in [0, x + a]) & \text{si } x > a \end{cases}$

$$x \sim \mathcal{N}(\mu, 1) \Longrightarrow \begin{cases} I\!P(\mu \in [x - a, 0]) = I\!P(x - \mu < a) = 0.95 & \text{si } x < -a \\ I\!P(\mu \in [x - a, x + a]) = I\!P(\mid x - \mu \mid < a) = 0.95 & \text{si } -a < x < a \\ I\!P(\mu \in [0, x + a]) = I\!P(x - \mu > -a) = 0.95 & \text{si } x > a \end{cases}$$

c) Si la distribución a priori de $\mu = 1$, $\forall\mu$, la distribución a posteriori de μ es $\xi(\mu/x) = \frac{1}{\sqrt{2\pi}}exp\{-\frac{1}{2}(x - \mu)^2\} \Longrightarrow \mu \sim \mathcal{N}(x, 1)$.
d) Como $\mu \sim \mathcal{N}(x, 1) \Longrightarrow \mu - x \sim \mathcal{N}(0, 1)$

$$I\!P(\mu \in C_a(x))/x) = \begin{cases} I\!P(\mu \in [x - a, 0]) & \text{si } x < -a \\ I\!P(\mu \in [x - a, x + a]) & \text{si } -a < x < a \\ I\!P(\mu \in [0, x + a]) & \text{si } x > a \end{cases}$$

$$I\!P(\mu \in C_a(x))/x) = \begin{cases} I\!P(\mu - x \in [-a, -x]) & \text{si } x < -a \\ I\!P(\mu - x \in [-a, +a]) & \text{si } -a < x < a \\ I\!P(\mu - x \in [-x, +a]) & \text{si } x > a \end{cases}$$

$$I\!P(\mu \in C_a(x))/x) = \begin{cases} \Phi(-x) - \Phi(-a) & \text{si } x < -a \\ \Phi(a) - \Phi(-a) & \text{si } -a < x < a \\ \Phi(a) - \Phi(-x) & \text{si } x > a \end{cases}$$

e) Si $a = 1.65$, $\Phi(a) = 0.95$, $\Phi(-a) = 0.05$ y $\Phi(-x) > 0.95$ si $x < -a$, luego $I\!P(\mu \in C_a(x))/x) \begin{cases} > & 0.90 & \text{si } x < -a \text{o } x > a \\ = & 0.90 & \text{si } -a < x < a \end{cases}$

$\lim_{a\to\infty} \Phi(-a) = 0$, $\lim_{a\to\infty} \Phi(a) = 1$; $\Phi(-x) = 1$, $six < -a$ y $\Phi(x) = 1$, $six > a$. Luego $\lim_{a\to\infty} I\!P(\mu \in C_a(x)/x) = 1$.

Capítulo 5

1. a) $X \to P(\lambda)$, donde X es el número de pasas en una empanada.
La distribución de probabilidad de una Poisson es $f_{x_i} = e^{-\lambda}\lambda^{x_i}/x_i!$.
Luego, la verosimiliutd está dada por $L(\underline{x}, \lambda) = e^{-n\lambda}\lambda^{\Sigma x_i}\prod(1/x_i!)$
Tomando el logaritmo de L y derivando con respecto al parámetro tenemos:

$$\frac{\partial L}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0 \Longrightarrow \hat{\lambda} = \bar{x}$$

b) (i) Sea $p = I\!P$(una cierta pasa esté en una empanada). La probabilidad de tener x pasas en una empanada está dada por una distribución binomial pues definimos una variable aleatoria Y_i como valiendo 1 si hay una pasa, y cerro si no la hay; luego, en n empanadas hay $\sum_{i=1}^{n} Y_i$ empanadas, v.a. que sigue una $Binomial(n, p)$.
Esta distribución se puede aproximar por una Poisson (ver curso de Probabilidades).
(ii) Se puede probar la hipótesis nula: $H_0 : X \sim \mathcal{P}(\lambda)$ con un test χ^2.
El estadístico utilizado es, por construcción: $Q = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}$, donde $n = 20$ (el tamaño de la muestra) y $k = 7$ (el número de clases en que están repartidos los datos).
Los grados de libertad de la χ^2 están dados por $k - 1 - 1$ pues hay un parámetro estimado.
Haciendo los cálculos se obtiene $Q = 0.449$. Con un nivel de significación del 5%, obtenemos la región crítica $Q \geq 11.070$. Como $0.449 < 11.070$, no se rechaza la hipótesis nula.
c) Para $H_0 = 3.5$ vs. $H_1 = 2.5$, se favorece al cocinero, pero para H_0' vs. H_1', se favorece a los alumnos. d) Para el test H_0, tenemos: $P(rechazar H_0/H_0 cierto) = 0.05$.
Del Lema de Neymann-Pearson tenemos que la región crítica está caracterizada por $\bar{x} < a$.
Por otro lado, $\sum x_i \sim \mathcal{P}(n\lambda)$ donde $n\lambda = 70$ bajo H_0.
Luego, $z = \dfrac{\sum x_i - E(\sum x_i)}{Var(\sum x_i)} \sim \mathcal{N}(0, 1)$.
y $I\!P(recharzar H_0/H_0 cierta) = I\!P(z < \dfrac{a - n\lambda}{\sqrt{n\lambda}}/H_0) = 0.05$.
Luego, utilizando la tabla de la distribución Normal, se obtiene que $a = 56.195$ y la región crítica está dada por $\sum x_i < 56.195$. En la muestra $\sum x_i = 60$, por lo que no se rechaza H_0.
Calculemos $\beta = I\!P(recharzar H_0/H_1 cierta) = I\!P(\sum x_i < 56.195/\lambda = 2.5) =$
$I\!P(z < \dfrac{56.195 - n\lambda}{\sqrt{n\lambda}}/\lambda) = 0.5)$
Bajo H_1, $\sum x_i \sim \mathcal{P}(50)$; luego, $\beta = 0.81$.
e) Para H_0' procediendo de manera análoga se obtiene:
$0.05 = P(\sum x_i > a/\lambda = 2.5) = P(z > \dfrac{a - 50}{\sqrt{50}}) \Longrightarrow a = 61.667)$
y la región crítica $\{\sum x_i > 61.667\}$.
Considerando la muestra concluimos que no se rechaza $H_0{}'$.
f) Se resumen los resultados anteriores en la tabla siguiente:

$\sum x_i$	$]-\infty, 56.195]$	$]56.195, 61.667]$	$]61.667, +\infty]$
$H_0 : \lambda = 3.5$	no son aceptables	son aceptables	son aceptables
$H_0' : \lambda = 2.5$	no son aceptables	no son aceptables	son aceeptables

Estamos ante una zona de contradicción: $|56.195,\ 61.667|$. Para suprimirla se propone dos maneras que se dejan propuestas:
(i) Fijar n y escoger un α que elimine la zona de contradicción.
(ii) Fijar un α y escoger un n

2. Se puede resolver de dos maneras:
a) Un test de Student:
Sea $D_i = p'_i - p_i$ las diferencias de pesos después y antes del matrimonio. Para responder, resolvemos la siguiente hipótesis: $H_0 : E(D) = 0$ vs. $H_1 : E(D) > 0$.
Suponiendo la normalidad tenemos: $\dfrac{\bar{D} - E(D)}{s/\sqrt{n-1}} \sim t_{n-1}$.
Luego, bajo H_0, se tiene $\dfrac{\bar{D}}{s/\sqrt{n-1}} \sim t_{n-1}$ con s^2 la varianza empírica.
$I\!P(t_9 > 1.83) = 0.05$. Luego, la región crítica estará dada por $\bar{D} > \dfrac{1.83s}{3}$.
Luego, de los datos se concluye:

- Para los hombres, $\bar{D} = 2.58$ y $s = 2.65$; luego, se rechaza H_0.
- Para las mujeres, $\bar{D} = -0.8$ y $s = 1.99$; luego, no se rechaza H_0.

Por tanto, hay evidencia para afirmar que los hombres suben de peso, las mujeres no.
b) Un test de Wilcoxon:
Sobre el rango de las diferencias positivas:

Hombres	$\lvert D_i\rvert$	5	0.5	4	3.5	2.5	2	3	1	1.5	6
	rango de $\lvert D_i\rvert$	9	1	8	7	5	4	6	2	3	19
	signo	+	−	+	+	+	−	+	−	+	+
Mujeres	$\lvert D_i\rvert$	2	1	3	1	2	2	2	3	3	1
	rango de $\lvert D_i\rvert$	5.5	2	9	2	5.5	5.5	5.5	9	9	3
	signo	+	−	−	−	+	−	−	−	+	−

Cuando hay empates, se da el rango promedio.
Luego:

$$W^+_{Hombres} = 48 \quad y \quad W^-_{Hombres} = 7 \qquad W^+_{Mujeres} = 20 \quad y \quad W^-_{Mujeres} = 35$$

donde $W^+ = \sum_{i=1}^{n} i\,\epsilon_i$, donde ϵ_i vale 1 si la diferencia es positiva, y 0 si no.
Para H_0, $I\!P(\epsilon_i^+) = \frac{1}{2}$ y ϵ_i^+ son independientes.
Luego, $E(W^+) = n(n+1)/4$ y $Var(W^+) = n(n+1)(2n+1)/24$. Por lo cual: $I\!P(\dfrac{W^+ - 27.5}{9.8} \geq 1.65) = 0.05$, de donde se obtiene como región crítica $W^+ \geq 43.7$.
Como $W^+_{Hombres} = 48$, se rechaza H_0. Para las mujeres no se puede usar esta aproximación ya que hay empates.

3. a). Sean $n_{i,j}$ las frecuencias de la tabla. Entonces: $n_{\bullet 1} = \sum_{i=1}^{2} n_{i1} \sim B(n, p'_1)$ y $n_{2\bullet} = \sum_{i=1}^{2} n_{2j} \sim B(n, p_2)$.
Entonces se tiene la aproximación: $d = \dfrac{n_{2\bullet} - n_{\bullet 1}}{n} \sim \mathcal{N}(p_2 - p'_1, \dfrac{p_2(1-p_2) + p'_1(1-p'_1)}{n})$.

Estimando la varianza con $\hat{p_2} = \frac{n_{2\bullet}}{n}$ y $\hat{p_1'} = \frac{n_{\bullet 1}}{n}$, se obtiene bajo H_0 que $d \sim \mathcal{N}(0, 0.00244)$.
Luego, $I\!P(\frac{|d|}{0.049} \geq 1.96) = 0.05$ implica una región crítica de la forma $|d| \geq 0.096$.
En la muestra, $d = -0.05$, por lo que no se rechaza H_0.
Para H_0', se deja propuesto.

b) Se calcula el χ^2 de contingencia: $Q = \sum \frac{(n_{ij} - n_{i\bullet}n_{\bullet j}/n)^2}{n_{i\bullet}n_{\bullet j}/n} = 97.306$
Bajo la hipótesis de independencia, Q sigue una χ^2_1. Como $P(Q > 3.84) = 0.05$, se rechaza la independencia, lo que concuerda con los resultados de la parte anterior.

4. a) Si $0 < y < \phi$, entonces $P(Y_n) \leq y = (\frac{y}{\phi})^n$ pues la muestra es aleatoria, los X_i siguen una $U(0, \phi)$ y se calcula la distribución del máximo.
Análogamente, si $y \geq \phi$, $P(Y_n \leq y) = 1$ pues $Y_n \leq \phi$ por definición de los X_i.
Por lo tanto, $\pi(\phi) = \begin{cases} I\!P(Y_n \leq 1.5) = 1 & \text{si } \phi \leq 1.5 \\ I\!P(Y_n \leq 11.5) = (\frac{1.5}{\phi})^n & \text{si } \phi > 1.5 \end{cases}$
b) El tamaño del test es $\alpha = sup_{\phi \geq 2}(\frac{1.5}{\phi})^n = (\frac{1.5}{2})^n = (\frac{3}{4})^n$.

5. a) Para cualquier valor de p, $\pi(p) = P(Y \geq 7) + P(Y \leq 1)$ por propiedad de la probabilidad.
Para saber la distribución de Y hacemos lo siguiente:
Sea Y_i una variable que vale 1 si la pieza i es defectuosa, y 0 sino. Esta v.a. sigue una distribución de Bernoulli de parámetro $P(Y_i = 1) = I\!P$(la pieza i sea defectuosa)$=p$.
Luego, $Y = \sum_{i=1}^{20} Y_i \sim B(20, p)$.
Para $p = 0$, $I\!P(Y \geq 7) = 0$ y $I\!P(Y \leq 1) = 1$. Por lo tanto, $\pi(0) = 1$.
Para $p = 0.1$, de una tabla de distribución binomial, concluimos:
$P(Y \geq 7) = .0020 + .003 + .001 + .000 = .00214$ y $P(Y \leq 1) = .1216 + .2701 = .3917$.
Luego, $\pi(0.1) = .3941$
Similarmente, $\pi(0.2) = 0.1558$
Nota: Si la muestra fuese de mayor tamaño, se puede utilizar la aproximación normal: $\mathcal{N}(p, p(1-p))$. Se puede escoger la varianza máxima, esto es. cuando $p = \frac{1}{2}$.
b) Puesto que H_0 es una hipótesis simple, el tamaño del test α se obtiene evaluando la función de potencia en el punto especificado por H_1. Luego, $\alpha = \pi(0.2) = 0.1558$.

6. Como la hipótesis nula es simple, el tamaño del test está dado por $\alpha = P(rechazar H_0/\mu = \mu_0)$. Bajo H_0 la v.a. $Z = \sqrt{n}(\bar{X}_n - \mu_0) \sim \mathcal{N}(0, 1)$. Luego, como el tamaño de la muestra es 25, $\alpha = I\!P(|\bar{X}_n - \mu_0| \geq 0) = P(|Z| \geq 5c) = 2|1 - \Phi(5c)| = 0.05$
en donde $\Phi(5c) = 0.975$, lo que implica que $c = 0.392$.

7. a) Las condiciones aquí son diferentes de las del Lema de Neyman-Pearson: en vez de fijar el valor de $\alpha(\delta)$ y minimizar $\beta(\delta)$, debemos fijar aquí el valor de $\beta(\delta)$ y minimizar $\alpha(\delta)$. Sin embargo, la misma demostración dada para el Lema de Neyman-Pearson muestra que el procedimiento óptimo es rechazar H_0 si $\frac{f_1(x)}{f_0(x)} > k$, donde k se escoge tal que $\beta(\delta) = I\!P$(No rechazar $H_0/H_1) = I\!P(\frac{f_1(x)}{f_0(x)} < k/H_1) = 0.05$.

Haciendo los cálculos, que son simples, se obtiene $\log \frac{f_1(x)}{f_0(x)} = \frac{3}{2}n\bar{x_n} - (constantes)$.
Luego, la razón de verosimilitud será más grande que k si y solo si $\bar{x}_n > k'$, con lo cual

se ha hallado un procedimiento óptimo para rechazar H_0, escogiendo k' tal que: $I\!P(\bar{X}_n < k'/H_1) = 0.05$.
Para calcular k', notemos que si H_1 es cierta, entonces $Z = \sqrt{n}(\bar{x}_n - 5.0) \sim \mathcal{N}(0,1)$. Luego, $I\!P(\bar{x}_n < k'/H_1) = \Phi|\sqrt{n}(k' - 5.0)| = 0.05$
en donde $k' = 5.0 - \dfrac{1.645}{\sqrt{n}}$.
b) Para $n = 4$, se rechaza H_0 si y solo si $\bar{x}_n > 4.1775$ (basta reemplazar $n = 4$ arriba).
Por lo tanto, $\alpha(\delta) = I\!P(\text{rechazar } H_0/H_0 \text{ cierta}) = I\!P(\bar{x}_n > 4.1775/H_0)$.
Bajo H_0, $\bar{x}_n \sim \mathcal{N}(3.5, 0.25)$. Por lo tanto, $Z = 2(\bar{x}_n - 3.5) \sim \mathcal{N}(0,1)$. Luego, $\alpha(\delta) = P(Z > 2(4.1775 - 3.5)) = 0.0877$.

8. a) Bajo H_0, $X \sim U(0,1)$. Luego es imposible obtener un valor de X más grande que 1 bajo H_0; sin embargo, esto es posible bajo H_1. Por lo tanto, si una solución del test rechaza H_0 sólo cuando $x > 1$, $\alpha(\delta) = 0$ y $\beta(\delta) = P(X < 1/H_1) = \frac{1}{2}$.
b) Para tener $\alpha(\delta) = 0$, en la región crítica sólo podemos incluir los puntos tales que tienen probabilidad nula bajo H_0. Por tanto, sólo los puntos $x > 1$ deben ser considerados. Para minimizar $\beta(\delta)$ podríamos escoger un conjunto que maximiza la probabilidad bajo H_1. Los puntos $x > 1$ pueden ser usados en esta región crítica.

9. Aplicamos el teorema que minimiza la combinación lineal de los errores con $a = b = 1$. El procedimiento óptimo que rechaza H_0 cuando $\dfrac{f_1(x)}{f_0(x)} > 1$.
En este caso, $\log \dfrac{f_1}{f_0} = y \ \log(\dfrac{\lambda_1}{\lambda_0}) - n(\lambda_1 - \lambda_0)$.
Puesto que $\lambda_1 > \lambda_0$, $\dfrac{f_1}{f_0} > 1$ si y solo si $\bar{x}_n > \dfrac{\lambda_1 - \lambda_0}{\log \lambda_1 - \log \lambda_0}$.

10. La probabilidad de rechazar H_0 es 0.05. Por lo tanto, el valor de la función de potencia en todo valor de ϕ es 0.05.

11. Cambiamos el parámetro de ϕ a $\zeta = -\phi$. En términos de este nuevo parámetro, las hipótesis se reescriben: $H_0 : \zeta \leq -\phi_0$ vs $H_1 : \zeta > -\phi_0$.
Sea $g_n(x/\zeta) = f_n(x/-\zeta)$. Si $\zeta_1 < \zeta_2$, entonces $\phi_1 = -\zeta_1 > \phi_2 = -\zeta_2$. Por lo tanto, la razón $\dfrac{g_n(x/\zeta_2)}{g_n(x/\zeta_1)}$ será una función decreciente de $r(X)$. Se sigue que esta razón será una función creciente de $s(X) = -r(X)$. Así, en términos de ζ podemos aplicar el teorema correspondiente, y por lo tanto se rechaza H_0 cuando $s(X) \geq c'$ y éste procedimiento es UMP. Pero, $s(X) \geq c'$ si y solo si $r(X) \leq c$, donde $c = -c'$. Por lo tanto, el test que rechaza H_0 cuando $T \leq c$ es un test UMP. Si c es escogido como se afirma en el enunciado, entonces del mismo teorema se sigue que el tamaño del test es α_0.

12. Es fácil verificar que un test que rechaza H_0 cuando $\sum_{i=1}^{n} X_i \leq c$ será un test UMP. Cuando $n = 10$ y $\lambda = 1$, $\sum_{i=1}^{n} X_i \sim Poisson(10)$ y $\alpha_0 = I\!P(\sum_{i=1}^{n} X_i \leq c/\lambda = 1)$. De una tabla de Poisson se tiene:

c	0	1	2	3	4
α	0.0000	0.0005	0.0028	0.0104	0.0293

Para valores más grandes de c, $\alpha_0 > 0.03$.

13. H_0 es una hipótesis simple. Por el Lema de Neyman-Pearson, el test que maximiza la función de potencia en un valor particular $\phi_1 > 0$ será rechazado si $\frac{f(x/\phi=\phi_1)}{f(x/\phi=0)} > c$, donde c es escogido tal que la probabilidad que esta desigualdad será satisfecha cuando $\phi = 0$ es α_0. Aquí, $\frac{f(x/\phi=\phi_1)}{f(x/\phi=0)} > c$ siy solo si $(1-c)x^2 + 2c\phi_1 x > c\phi_1^2 - (1-c)$. Para cada valor de ϕ_1, el valor de c se escoge tal que el conjunto de los puntos que satisfacen esta desigualdad tienen probabilidad α_0 cuando $\phi = 0$. Para dos valores diferentes de ϕ_1, estos dos conjuntos serán diferentes. Por lo tanto, diferentes procedimientos de test maximizarán la potencia con dos valores diferentes de ϕ_1. Se conclue que no existe un test UMP.

14. El test UMP rechazará H_0 cuando $\bar{X}_n \geq c$, donde $I\!P(\bar{X}_n \geq c/\mu = 0) = I\!P(\sqrt{n}\bar{X}_n \geq \sqrt{n}c/\mu = 0) = 0.025$. Sin embargo, cuando $\mu = 0$, $Z = \sqrt{n}\bar{X}_n \sim \mathcal{N}(0,1)$. Por tanto, $I\!P(Z \geq 1.96/\mu = 0) = 0.025$, de donde $c = \frac{1.96}{\sqrt{n}}$.

Cuando $\mu = 0.5$, la v.a. $Z = \sqrt{n}(\bar{X}_n - 0.5) \sim \mathcal{N}(0.1)$. Por lo tanto, $\pi(0.5/\delta^*) = I\!P(\bar{X}_n \geq 1.96n^{-1/2}/\mu = 0.5) = I\!P(Z \geq 196 - 0.5n^{-1/2}) = I\!P(z \leq 0.5n^{-1/2} - 1.96)$ por la simetría de la normal.

Pero, $\Phi(1.282) = 0.9$. Por lo tanto, $\pi(0.5/\delta^*) \geq 0.9$ si y solo si $n \geq 42.042$. Así, una muestra de tamaño $n = 43$ es requerida. Puesto que la función de potencia es una función estrictamente creciente de μ, entonces se tendra $\pi(0.5/\delta^*) \geq 0.9$ para $\mu > 0.5$.

Cuando $\mu = -0.1$, la v.a. $Z = \sqrt{n}(\bar{X}_n + 0.1) \sim \mathcal{N}(0,1)$. Por lo tanto, $\pi(-0.1/\delta^*) = 1 - Phi(1.96 + 0.1n^{1/2}$. Pero como $\Phi(3.10) = 0.999$, $\pi(-0.1/\delta^*) \leq 0.001$ si y solo si $n \geq 129.96$. Así, una muestra de tamaño $n = 130$ se requiere para que $\pi(-0.1/\delta^*) \leq 0.001$. Como la función de potencia es estrictamente creciente como función de μ, entonces $\pi(\mu/\delta^*) \leq 0.001$ para $\mu < -0.1$.

15. Sea $\alpha_0 = 0.05$ y sea $c_1 = 3\alpha_0^{1/n}$ (Justifique la elección de esta constante). Sea $c_2 = 3$. Entonces $\pi(\phi/\delta) = I\!P(T \leq 3\alpha_0^{1/n}/\phi) + I\!P(T \geq 3/\phi)$.

Puesto que $I\!P(T \geq 3/\phi) = 0$ para $\phi \geq 3$, entonces: $\pi(\phi/\delta) = [\frac{3\alpha_0^{1/n}}{\phi}]^n + 1 - (\frac{3}{\phi})^n > \alpha_0$.

Capítulo 7

1. a) Se observa un R^2 relativamente alto, R=cor(y,$\hat{y}$)=0.85. El F de Fisher a 3 y 196 g.l. muestra que se rechaza $H_o : E(y) = CONSTANTE$. Sin embargo, sólo el Atraso Medico tienen influencia significativa sobre el tiempo de espera, y el número de medicos en menor grado, como lo muestran las probabilidades de las T de Student que tienen 196 g.l.

b)

$$F = \frac{R^2/3}{(1-R^2)/196}$$

$$\hat{\sigma}^2 = \sum \hat{\epsilon}_i^2/196$$

$$\hat{\sigma}^2 = \sum y_i^2(1-R^2)/196 = 64.286$$

Si se agrega una variable explicativa, el coeficiente de correlación multiple R^2 no puede disminuir, dado que el subespacio W se amplia.

c) $\frac{\hat{\beta}_j - \beta_j}{\sigma_{\beta_j}} \sim t_{196} \Longrightarrow I\!P(|t_{196}| < 1.96) = 0.95$
Luego, el intervalo de confianza simétrico a 95% para el coeficiente del ATRASO MEDICO es: $[0.09 - 1.96 \times 0.01, 0.09 + 1.96 \times 0.01] = [0.0704, 0.1096]$
d) La predicción del tiempo de espera para este nuevo paciente es:

$$\hat{y}_o = 22.00 + 0.09 \times 100 - 0.02 \times 0 - 1.61 \times 4 = 24.56$$

El intervalo de confianza a 95% es igual a: $[\hat{y}_o - 1.96\hat{\sigma}_o, \hat{y}_o + 1.96\hat{\sigma}_o]$,
con la varianza del error de predicción igual a: $\hat{\sigma}_o^2 = \hat{\sigma}^2(1 + x_o^t(X^tX)^{-1}x_o)$

2. a) El estimador de M.V. de β es: $\hat{\beta} = (X^tX)^{-1}X^tY$
Como aquí X^tX es de la forma diagonal por bloque:

$$\hat{\beta} = (X_1^tX_1)^{-1}X_1^tY + (X_2^tX_2)^{-1}X_2^tY$$

Pero $(X_1^tX_1)^{-1}X_1^tY$ es el estimador de M.V. del modelo $Y = X_1\alpha_1 + \epsilon$ y $(X_2^tX_2)^{-1}X_2^tY$ es el estimador de M.V. del modelo $Y = X_2\alpha_2 + \epsilon$.
b) En el caso de $X_1^tX_2 \neq 0$, el estimador $\hat{\beta}$ para β es sesgado.

3. a) El coeficiente de correlación multiple no es muy elevado, sin embargo los coeficientes son significativos individual y globalmente (se rechaza las hipótesis $H_o : \alpha = 0$ y $H_o : \beta = 0$ y $H_o : \alpha = \beta = 0$).
b) $R^2 = \frac{Var(\hat{y})}{Var(y)}$; luego $\sum_i e_i^2 = nVar(y)(1 - R^2)$.
$\hat{\sigma}^2 = \frac{\sum_i e_i^2}{n-2} = \frac{nVar(y)(1-R^2)}{n-2} = 7.687$
c) Se denota una tendencia creciente de los residuos, lo que indica un sesgo en la estimación: falta algunas variables explicativas en el modelo.
d) La predicción para y_o se hace a través del modelo:
$\hat{y}_o = 7.05 + 0.79 * 6.50 = 12.185$
$Var(\hat{y}_o) = \frac{\sum x_i^2 - 2x_o\sum x_i + nx_o^2}{n\sum x_i^2 - (\sum x_i)^2} = 0.0264$.
e) Como $\sum_i x_i z_i = 0$ y $\sum z_i = 0$, se tiene que $X_1^tX_2 = 0$. Luego del problema 1 se deduce que $\hat{\beta}_0 = 7.05$ $\hat{\beta}_1 = 0.79$ y $\hat{\beta}_2 = 1.00$.

4. El modelo a ajustar es tal que: $Y_i = \beta_j + \epsilon_i \quad$ si $t_i \in I_j$.
Sean las variables indicadoras z_j definidas sobre los intervalos I_j por:

$$z_{ji} = \begin{cases} 1 & \text{si } t_i \in I_j \\ 0 & \text{si } t_i \notin I_j \end{cases}$$

El modelo lineal se escribe entonces: $Y_i = \sum_{j=1}^k \beta_j z_{ji} + e_i$.
b) El criterio de los minimos cuadrados se escribe:
$Q = \sum_i (Y_i - f(t_i))^2 = \sum_j \sum_{i \in I_j} (Y_i - \beta_j)^2 \Longrightarrow \frac{dQ}{d\beta_j} = 2\sum_{i \in I_j}(Y_i - \beta_j) = 0$
Se obtiene $\hat{f(t)} = \hat{\beta}_j = \sum_{i \in I_j} Y_i/n = Y_j$ si $t \in I_j$.
c) Como $Y_j \sim \mathcal{N}(\beta_j, \sigma^2/n_j)$, se obtiene un intervalo de confianza para β_j a 95% con: $[Y_j - 1.96\sigma/\sqrt{n_j}, \bar{Y}_j + 1.96\sigma/\sqrt{n_j}]$.
Como $\int_{a_0}^{a_K} \hat{f(t)}dt = \sum_1^K (a_j - a_{j-1})\bar{Y}_j$ y los $\bar{Y}_j$ son independientes, $\sum_1^K (a_j - a_{j-1})\bar{Y}_j \sim$

$\mathcal{N}(\int_{a_0}^{a_K} f(t)dt, \sigma^2 \sum_1^K \frac{(a_j - a_{j-1})^2}{n_j})$; se obtiene entonces un intervalo de confianza a 95% para $\int_{a_0}^{a_K} f(t)dt$ con:

$$[\sum_1^K (a_j - a_{j-1})\bar{Y}_j - 1.96\sigma\sqrt{\sum \frac{(a_j - a_{j-1})^2}{n_j}}, \sum_1^K (a_j - a_{j-1})\bar{Y}_j + 1.96\sigma\sqrt{\sum \frac{(a_j - a_{j-1})^2}{n_j}}]$$

5. Como $W = H_a \oplus \Delta_a$ con $H_a \perp \Delta_a$, $P_W(Y) = P_{H_a}(Y) + P_{\Delta_a}(Y)$. Además H_a pertenece a W, luego $P_{\Delta_a}(Y) = P_{\Delta_a}(\hat{Y}) = (\frac{a^t\hat{Y}}{a^t a})a$. Luego, $\hat{Y} = Y^* - (\frac{a^t\hat{Y}}{a^t a})a$.

b) Como $H_a \perp \Delta_a$, $Var(b^t\hat{y}) = Var(b^t y^*) + Var((\frac{a^t\hat{y}}{a^t a})b^t a)$.

Pero $Var((\frac{a^t\hat{y}}{a^t a})b^t a) = (\frac{b^t a}{a^t a})^2 Var(a^t\hat{y})$.

Por otro lado $Var(\hat{y}) = \sigma^2 P_W$, y como $a \in W$, $P_W(a) = a$.

Luego, $Var(b^t\hat{y}) = Var(b^t y^*) + \sigma^2\frac{(b^t a)^2}{a^t a}$.

c) $\sqrt{\sum \epsilon_i^{*2}}$ es la norma de la proyección de un vector normal $\mathcal{N}_n(0, \sigma^2 I_n)$ sobre el subespacio ortogonal a H_a en $\mathbb{R}^n$, luego $\frac{\sum_{i=1}^n \epsilon_1^{*2}}{\sigma^2} \sim \chi^2_{n-p}$, siendo n-p la dimensión de $H_a^\perp$.

d) La distribución de $\frac{\sum y_i^{*2}/p}{\sum \epsilon_i^{*2}/(n-p)}$ es una F de Fisher a p y n-p g.l., dado que Y^* pertenece a $H_{\mathbf{I}_n}$ que tiene dimensión p y los residuos al ortogonal de $H_{\mathbf{I}_n}$, que tiene dimensión n-p.
Si las variables son centradas, W es ortogonal al vector $\mathbf{1}_n$ luego $\hat{Y}$ lo es también; si $a = \mathbf{1}_n$ entonces $W = H_{\mathbf{I}_n}$, e $\hat{Y} = Y^*$.

Apéndice C

RESUMEN DE DISTRIBUCIONES

Apéndice D

TABLAS ESTADÍSTICAS

- TABLA 1: DISTRIBUCIÓN NORMAL
- TABLA 2: DISTRIBUCIÓN t DE STUDENT
- TABLA 4: DISTRIBUCIÓN χ^2
- TABLA 4: DISTRIBUCIÓN F DE FISHER
- TABLA 5: PROBABILIDADES BINOMIALES
- TABLA 6: PROBABILIDADES DE POISSON
- TABLA 7: TABLA DE NÚMEROS AL AZAR