Universidad de Chile
Fac. Cs. Fís. y Matemáticas
Departamento de Ingeniería Matemática

Proyecto FONDEF
D99I1049
IDEA+

Estadística

Nancy Lacourly

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

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

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 (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.

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.

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 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:

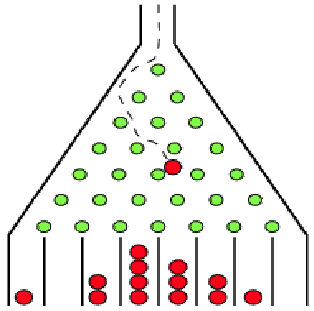
$$\min_z \sum_{i=1}^{n} (x_i - z)^2$$

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

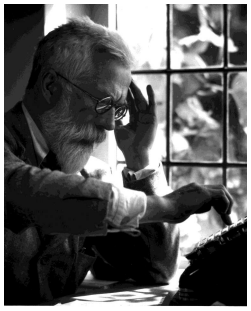
Es con la introducción de nuevas aplicaciones que la teoría de las probabilidades del siglo XVIII funda la estadística matemática. El término de *estadística* se debe posiblemente a G. Achenwall (1719-1772), profesor de la Universidad de Göttingen, tomando del latín la palabra *status*.[2] Achenwall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

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

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.

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.

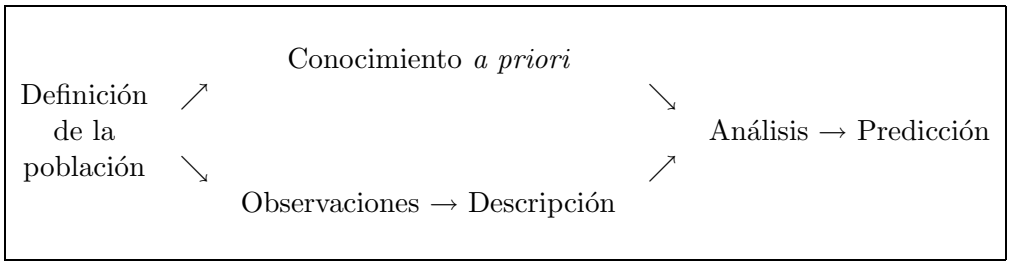
El estadístico trata entonces de inferir informaciones sobre la población a partir de los valores observados en la muestra. La muestra podrá no ser **representativa** de la población en el sentido que algunas características de interés podrán ser sobreestimadas o subestimadas.

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 **??**).

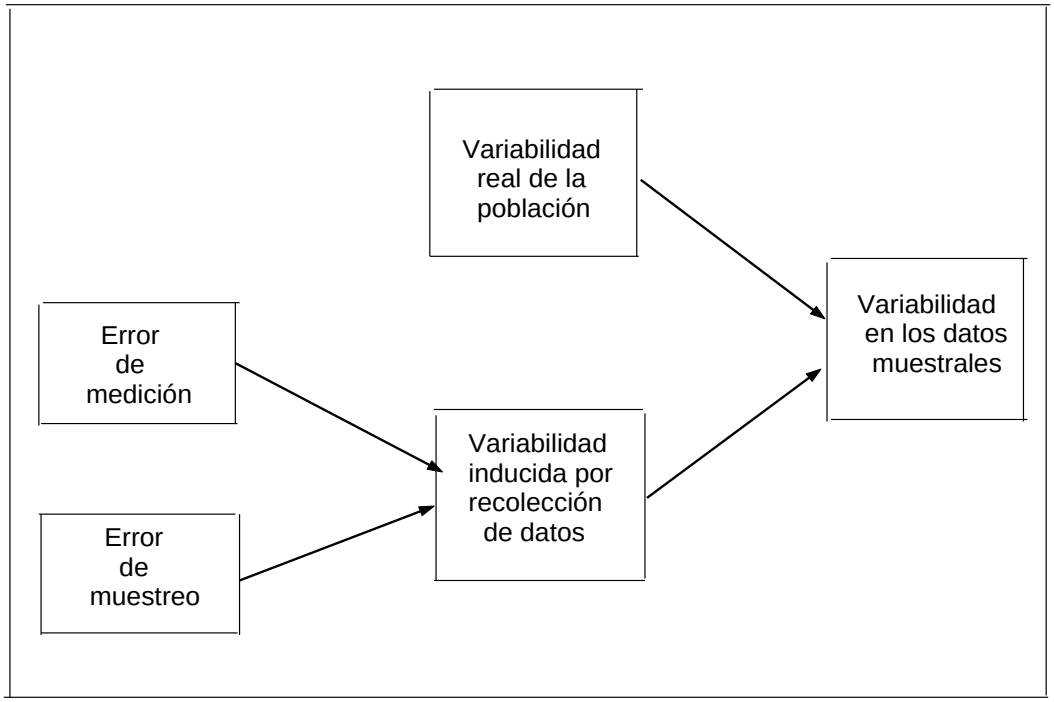


Figura 1.1: Esquema de las variabilidades

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 $\begin{pmatrix} N \\ n \end{pmatrix}$. Luego cada muestra tiene una probabilidad igual a:

$$\frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

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

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 $f(2, n) = \begin{pmatrix} 1+n \\ 1 \end{pmatrix}$.

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 **??**) .

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 **??** 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 **??**).

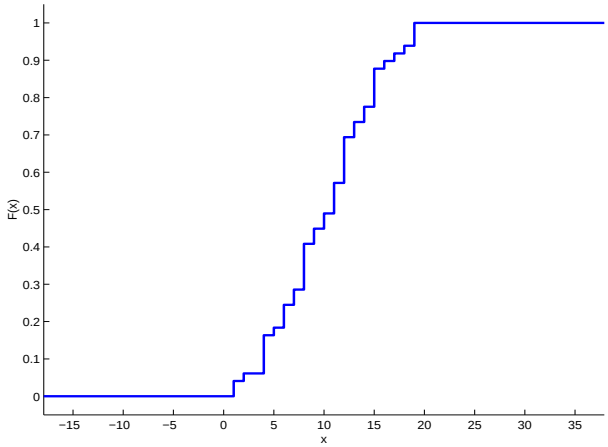


Figura 2.1: Una función de distribución empírica

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 **??** 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:

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

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 **??**). 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

Cuadro 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

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

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

■

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 **??** y del corolario **??**.

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

$$I\!P(X < x_\alpha) \leq \alpha \leq I\!P(X \leq x_\alpha)$$

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 **??**), 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 **??** 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:

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

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 parámetrica** 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 **??** de las ampolletas, hicimos el supuesto que $F(x)$ pertenece a la familia de las distribuciones $Gamma(\alpha, \beta)$, en los ejemplos **??** de la talla de las chilenas y **??** de las bolsas de azúcar, la distribución normal $\mathcal{N}(\mu, \sigma^2)$ y los ejemplos **??** del candidato a la eleción y **??** 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 **??** 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 **??**, 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* $X : \mathcal{P} \longrightarrow Q$
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 $\delta : Q^n \longrightarrow \Omega$, 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 **??**).

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*

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

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 $E(\hat{\theta}) \overset{n\to\infty}{\longrightarrow} \theta$.*

Se puede construir un estimador insesgado a partir de S_n^2: $\tilde{\sigma}^2 = \frac{1}{n-1}\sum(x_i - \bar{X}_n)^2$. 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
($\hat{\theta}_n \overset{m.c.}{\longrightarrow} \theta$).

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.*

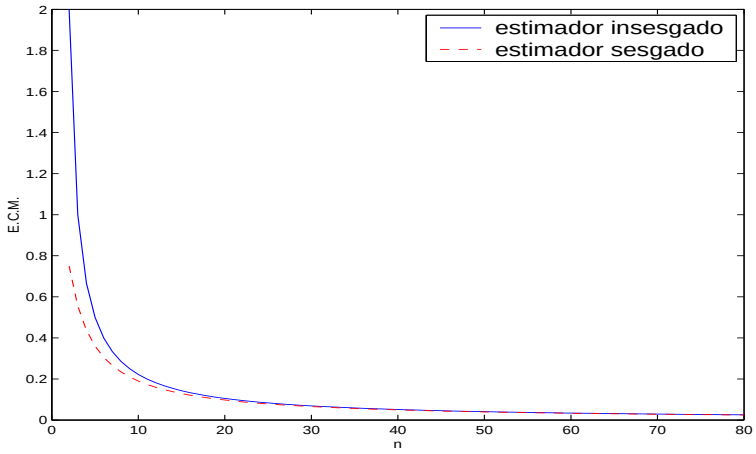


Figura 3.1: Error cuadrático medio en función de n

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 **??** la variación del error cuadrático medio en función de la tamaño del la muestra para los dos estimadores cuando $\sigma^2 = 1$.

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 **??** donde el centro del blanco representa el parámetro a estimar y los otros puntos son diferentes estimaciones obtenidos de un estimador).

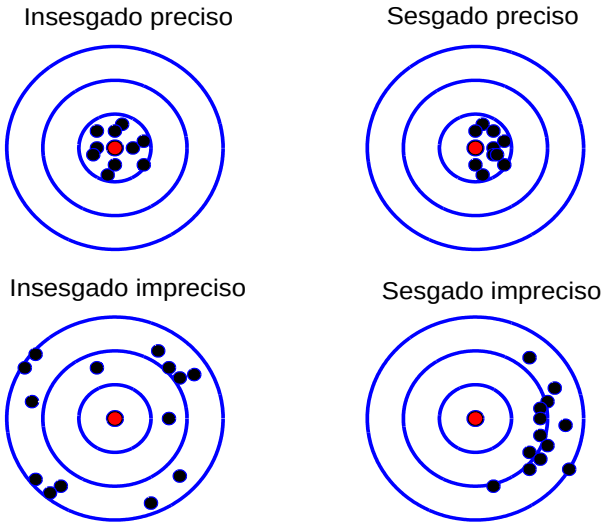


Figura 3.2: Sesgo y varianza

Otra manera de ilustrar el problema entre sesgo y precisión esta dada en las figuras **??** 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.

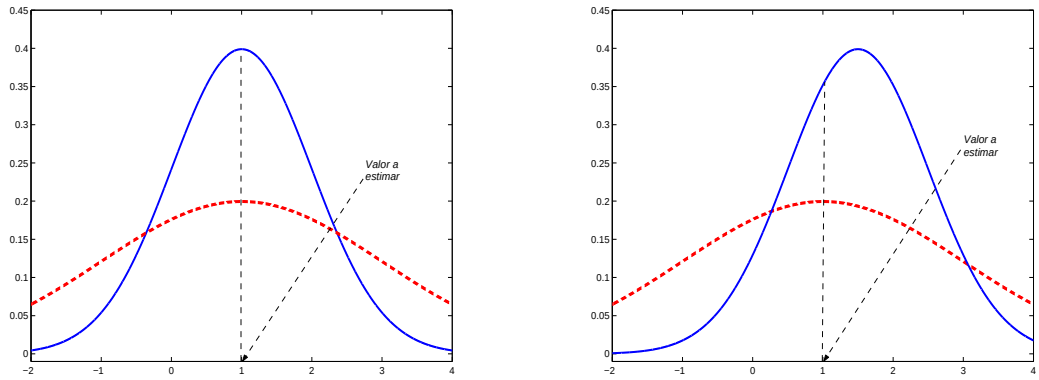


Figura 3.3: Sesgo y varianza

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 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 **??**, 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 **??**, 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 **??** y **??**, 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 **??**, $x_1, x_2, ..., x_n$ 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 **??**.

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

Cuando $\theta \ge x_i$ 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:

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

es decir: $f_\theta(\underline{x}) = 1 \quad si \quad max\{x_1, ..., x_n\} - 1 \le \theta \le min\{x_1, ..., x_n\}$. 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.