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

2005

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 múltiles 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

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

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

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.

correcta. De Méré siguió planteando problemas que no pudieron resolver los matemáticos de su época. Sin embargo, Jacques de Bernoulli (1654-1705), el primero de una famosa familia de matemáticos suizos, dio una demostración de la ley de los Grandes Números y Abraham de Moivre enunció el teorema de la regla de multiplicación de la teoría de la probabilidad.

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

La demografía

Las reglas de cálculo desarrolladas hasta entonces para los juegos de azar vieron sus aplicaciones en otras disciplinas. Los censos demográficos, que se hacían desde la antigüedad, requieren recolectar muchos datos. En Inglaterra, a pesar que John Grant tenia la noción de las tablas de mortalidad, es Edmund Halley (1656-1742) que construye por primera vez una tabla de mortalidad utilizando observaciones.

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

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

riesgo de morir de la enfermedad era el mismo en todas las edades. Lo que evidentemente es muy discutible.

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

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

La teoría de los errores y la distribución normal

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

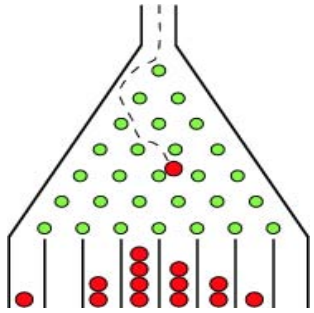
$$\min_z \sum_{i=1}^{n} (x_i - z)^2$$

La solución es el promedio de las mediciones.

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

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

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



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

Nacimiento de la estadística Moderna

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

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

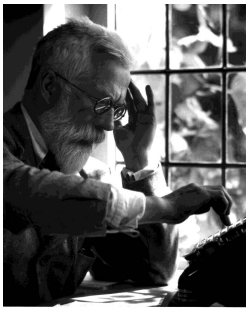
La estadística se empezó a usar de una manera u otra en todas las disciplinas, a pesar de un estancamiento de la teoría de las probabilidades. En particular, muchos vieron la dificultad de aplicar el concepto de simetría, o de casos igualmente posibles, en todas las aplicaciones. Hubo que esperar a que 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.

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

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



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.

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 veri-ficar 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.

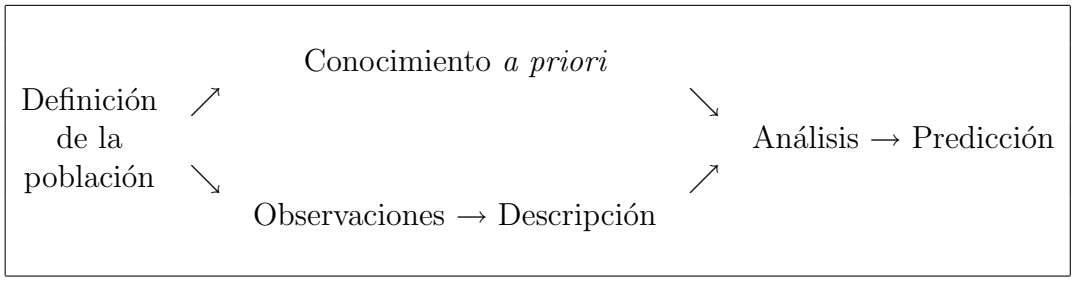
El estadístico trata entonces de inferir informaciones sobre la población a partir de los valores observados en la muestra. La muestra podrá no ser **representativa** de la población en el sentido que algunas características de interés podrán ser sobreestimadas o subestimadas.

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

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

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

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

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

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

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

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

Los valores de las variables obtenidos sobre los elementos de la muestra se llaman **valores muestrales**. Si la muestra se obtiene de un muestreo aleatorio, los valores muestrales son variables aleatorias cuya distribución depende de la población. Las características calculadas a partir de los valores muestrales son aleatorias también.

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

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

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

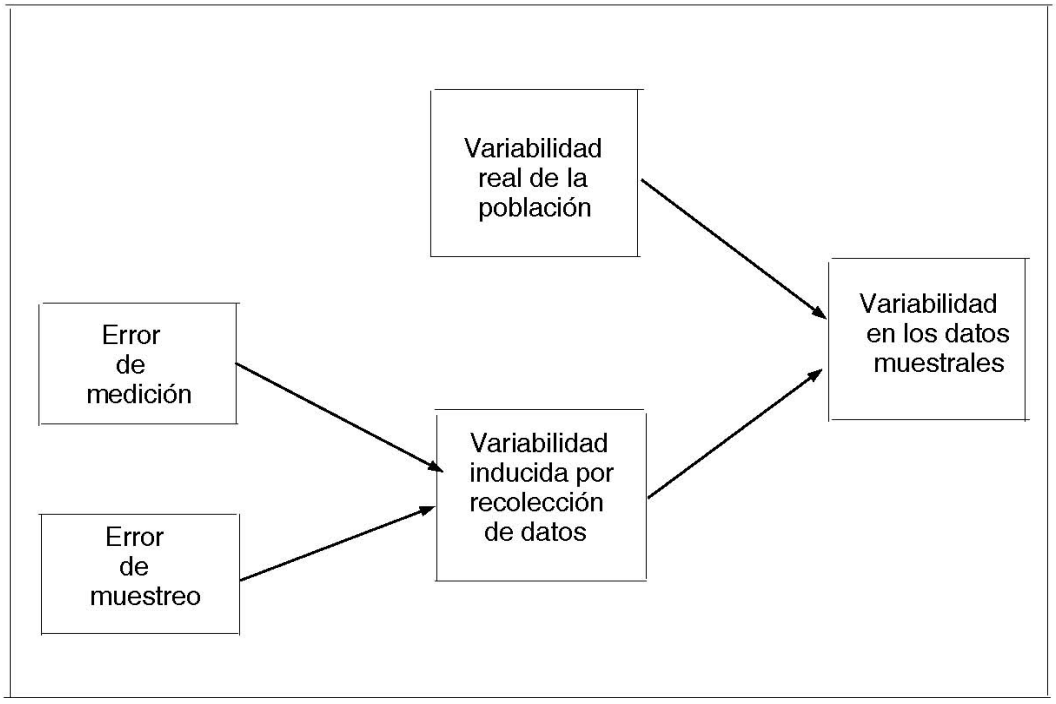


Figura 1.1: Esquema de las variabilidades

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

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

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

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

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

En lenguaje probabilista:

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

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

En una población finita de tamaño N con todos sus elementos equiprobables, el número total de muestras posibles sin reemplazo de tamaño n es igual a $\begin{pmatrix} N \\ n \end{pmatrix}$. 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 estra-tificado 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.