APUNTES DE ESTADISTICA

Nancy Lacourly

2001

PREFACIO

Este curso de estadística hace parte del plan común de ingeniería [1] Como para algunas carreras es el único curso de estadística que tendrá el alumno, se ha trata aquí dar una visión de la metodología basica de la Inferencia Estadística y una introducción a los modelos lineales y métodos multidimensionales. 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 grandes volumenes de datos que en la actualidad pueden ser estudiados fácilmente.

Si bien el cálculo de las probabilidades es una teoría matemática abstracta, que deduce consecuencias de un conjunto de axiomas, al contrario la estadistica trata encontrar un modelo que refleja mejor los datos obtenidos a partir de experimentos y necesita, entonces, dar una interpretación concreta a la noción de probabilidad. Varias interpretaciones fueron propuestas por los estadisticos, 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 a lo incierto, para definir un grado de creencia.

[1] Proyecto FONDEF D99I1049, Departamento de Ingeniera Matemtica, Universidad de Chile.

Índice General

1 INTRODUCCION A LA ESTADISTICA

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, estudiaron el experimento relativo a tirar un dado y hicieron el supuesto que el dado no está cargado (sucesos elementales equiprobables), lo que permite deducir que la probabilidad de sacar ün número par"es igual a $1/3$**. A partir de un modelo probabilístico adecuado, se deduce nuevos modelos o propiedades. En Estadística tratamos responder a la pregunta** *¿el dado no 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 ajusta bien los datos y no lo contrario.**

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 lanzo 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.

La palabra Estadísticas (al plural) designa un conjunto de datos observados y la palabra Estadística (al singular) designa la rama del método científico que trata estos datos observados y consiste de métodos para la recolección de los datos, y para el tratamiento, interpretación y análisis de estos datos.

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 se usa para resolver tales problemas. Terminamos esta introducción con la presentación de 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 ingls Benjamin Disraeli (1804-1881)

1.1 Historia del azar y del desarrollo de la Estadística

El desarrollo de la computación trastornó los progresos de la Estadística y su enseñanza. Vamos a ver aquí como y por quién se desarrollo la Estadística, desde la prehistoria hasta la actualidad. Es difícil separar la evolución de la Estadística sin considerar la de las Probabilidades. El progreso de ambas disciplinas puede verse como la historia de una única ciencia: la ciencia del azar.

1.1.1 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[2] 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) se encuentran tres tipos de nociones de probabilidad, que definen más bien actitudes frente al azar y la fortuna, que siguen vigentes hasta nuestros días: (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.

1.1.2 La edad Media

Durante la edad media hubo una gran actividad científica y artística en Oriente y el nombre de *azar* parece haber venido desde Siria a Europa. La flor de zahar, que aparecía en los dados de la época podría ser el origen de la palabra. Las compañías aseguradoras iniciaron investigaciones matemáticas desde tiempos muy antiguos, y en siglo XVII aparecieron los primeros famosos problemas de juegos de azar. En la sociedad francesa, el juego era uno de los entretenimientos más frecuentes. Los juegos cada vez más complicados y las apuestas muy elevadas hicieron sentir la necesidad de calcular las probabilidades de los juegos de manera racional. El caballero de Méré, un jugador apasionado, escribiendo a Blas Pascal

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

(1623-1662) sobre ciertos juegos de azar, dio origen a una correspondencia entre algunos matemáticos de la época. Las preguntas de De Méré permitieron, en particular, iniciar una discusión entre 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", 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 cualquier sea lo 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 su base de 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:

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

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 X y a Bertrán le falta Y para ganar. ¿Cómo tienen que repartirse las apuestas? Es uno de los famosos problemas propuestos por De Méré y que fue resuelto por Pascal y Fermat**

1.1.3 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 antigedad, requieren recolectar muchos datos. La demografía y los seguros de vida aprovecharon del desarrollo de la teoría de las probabilidades. Consideramos, 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 con Pierre Simon, Marqués de Laplace (1749-1827), estos problemas fueron reconocidos como similares a los de un juego, y se encontraron 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 a toda edad. Lo que evidentemente es muy discutible.

1.1.4 La teoría de los errores

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 hubiera errores. En su Anexo sobre el método de los mínimos cuadrados, de "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) 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 (ver anexo). Galton trabajo en meteorología y en herencia. Era el primo de Charles Darwin.

> **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**

1.1.5 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 Gttingen, tomando del latín la palabra *status*.

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 campo de la genética y la agricultura.

1.1.6 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. En Francia (J. P. Benzécri) y en los Estados unidos (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**

1.1.7 Cálculo de Probabilidades y Estadística

Algunas palabras para concluir. Si bien la historia de la Estadística no se puede separar de la historia del Clculo de las Probabilidades, la Estadística no puede considerarse como una simple aplicacin del Cálculo de las Probabilidades. Podemos comparar esta situacin a la de la Geometra 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 teora matemtica 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 lanz 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 Estadstica la que permite hacerlo.

> **El geómetra no se interesa por saber si existen en la prctica 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 al problema, como matem'atico, 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 EJEMPLOS DE PROBLEMAS ESTADISTICOS

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.

subsubsectionEstadísticas y el Estado

Un estado necesita conocer su poblacin: Los censos permiten obtener estadísticas demográficas y los métodos estadísticas hacer predicciones demográficas. Para poder elaborar una planificación de la salud, el gobierno tiene que tener informaciones sobre las necesidades (datos demográficos, enfermedades según las estaciones, etc.). En función de estas informaciones, se crean nuevos hospitales, consultorios se amplían antiguos. Para radicar la pobreza o definir una política del empleo, hay que saber cual es el problema. En el campo de la agricultura, se requiere hacer buenas predicciones de la producción (de trigo, por ejemplo) y decidir si estas van a satisfacer la demanda. En la explotación de los bosques es importante estimar los volúmenes y calidad de madera esperada en una zona dada para la planificación de las cosechas y los requerimientos de la demanda.

subsubsectionEstadísticas y empresas

Una fabrica o una empresa de servicios requiere saber de sus recursos y productos manufacturados, de la demanda por sus productos, pero de la competencia también. 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.

subsubsectionEstadísticas 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 experimentaciones 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 en métodos estadísticos. Para contar plantas o animales se usa un procedimiento de muestreo aleatorio. Las famosas leyes de Mendel, a pesar de referirse a caracteres genéticos cualitativas, pueden considerarse como leyes estadísticas.

subsubsectionEstadísticas 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, el psicometrista mide caractersticas psicológicas relativas al comportamiento, aprendizaje y el rendimiento de los estudiantes. Son tests llevados a escalas numéricas, a partir de los cuales se pueden hacer estudios estadísticos.

1.3 EL RAZONAMIENTO ESTADISTICO

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 el cual se obtienen informaciones para sacar conclusiones sobre el grupo. Cuando obtener mediciones sobre cada elemento de la población (es 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. Por ejemplo, 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 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.

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 interesa separadamente pero solamente por que hace parte de la población. La teoría de muestreo nos ofrece métodos para definir muestras. Distinguiremos entonces la Estadística Descriptiva, que es una actividad que consiste en resumir y representar informaciones, de la Inferencia Estadística, que es 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 anterioramente 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.

1.3.1 Población y muestras

Los datos experimentales son obtenidos sobre conjunto de individuos u objetos, que llamaremos unidad de observación, sobre el cual se quiere conocer algunas características. La totalidad de estas unidades de observación que concierna el estadístico se llama población. La población puede ser finita: en una encuesta de opinión, es la población de un país; el conjunto de ampolletas fabricadas por una maquina; los arboles de un bosque.

La población puede ser considerada también comos 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 a un paradero de bus. En estos casos la población es definida por un conjunto de cardinal infinito y generalmente esta definido por una variable aleatoria y su distribución de probabilidad.

Generalmente la población a estudiar, aún si es finita, es demasiado vasta, se extrae entonces solamente un subconjunto de la población, llamada subpoblación o muestra sobre lo cual se observan características llamadas variables. Los elementos de la muestra podran ser repetidos o no y el orden de extracción podra 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** 100 **ampolletas.**

El estadístico trata entonces concluir informaciones sobre la población a partir de los valores observados en la muestra. La muestra podra no ser representativa de la población en el sentido que algunas características podran ser sobreestimadas o subestimadas.

Definición 1.1 ***Se dice que una muestra es representativa de una población si toda unidad de observación puede aparecer en la muestra y esto con una probabilidad conocida.***

1.3.2 Etapas del razonamiento estadístico

El razonamiento estadístico se descompone generalmente en varias etapas:

- **Definición del problema: objetivos y definición de la población**

- **Determinacin del muestreo.**
- **Recolección de los datos.**
- **Descripción estadística de los datos.**
- **Análisis de los datos.**
- **Decisión o predicción.**

Conocimiento *a priori* $\searrow$

Análisis $\rightarrow$ **Predicción**

Observaciones $\rightarrow$ **Descripción** $\nearrow$

1.3.3 Recolección de los datos

Se distingue los censos -en que los datos estan recolectados sobre la integralidad de las unidades de observación de la población considerada- de los muestreos -en los cuales se recoge informaciones sobre sólo 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 proviene?* **La forma de elegir la muestra depende del problema (teorías del diseño de muestreo y del diseño de experimentos) y puede ser muy compleja, pero generalmente la muestra está obtenida aleatoriamente y llama a usar la teoría de las probabilidades.**

1.3.4 Descripción estadística de los datos

La descripción estadística permite resumir, reducir y presentar 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.

1.3.5 Análisis de los datos

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

La inferencia estadística consiste en metodos que extenden características obtenidas sobre una muestra a la población. Se basa en un modelo que dependerá de los

datos y eventualmente del conocimiento *a priori* que se puede tener sobre el fenómeno estudiado (teoría bayesiana de la estadística). 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. 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 un 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 un 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?

La primera pregunta se responde con la teoría de muestreo y la segundo con la inferencia estadística. Se encuentran dos tipos de problemas:

- El problema de estimación: Se busca precisar alguna característica totalmente desconocida de la población a partir de los datos obtenidos sobre una muestra. Por ejemplo, se quiere conocer la duración promedio de las ampolletas ILUMINA DE 100W a partir de una muestra de 500 ampolletas.
- El problema de test de hipótesis cuando se quiere comprobar alguna información sobre la población. Por ejemplo, las ampolletas ILUMINA de 100W no duran más de 250 horas. Entonces se tiene que comprobar esta aseveración a partir de la información relativa a una muestra.

1.3.6 Decisión o predicción

El análisis esta condicionado por la finalidad del estudio, que consiste en general en tomar una decisión o proceder a alguna predicción. Por ejemplo, se tiene que decidir, a partir de algunos experimentos, si un tratamiento es eficaz para combatir una determidada enfermedad, o bien predecir el IPC del próximo mes, o las temperaturas mínima y máxima de mañana en Santiago.

Un candidato a una elección presidencial encarga a un centro de estudio de opiniones un estudio 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 45contra 55nivel de confianza de 95tiene muy poca posibilidad de ser elegido, salvo si cambia su campaña electoral.

1.4 TEORIA DE MUESTREO

Una base importante de la estadística está contenida en la teoría de muestreo. Pero la elección de un método de muestreo depende de la población y de las mediciones que se recolecta sobre las unidades de observación.

En los problema citados anterioramente de 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 eleción, nos preguntamos

¿Qué esperamos de una muestra para responder a estos problemas?

- **Que no tenga sesgo: Para tener una muestra sin sesgo, es decir una muestra que no sobreestima o no subestima alguna característica de la población, todo elemento de la población debería tener la posibilidad de ser elegido en la muestra. Además la selección debería ser objetiva, sin que ningún factor personal intervenga. De aquí que se da un carácter aleatorio al muestreo, es decir asignar a cada elemento de la población una probabilidad de selección para la muestra y que cada probabilidad sea 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 de distribución que depende de la población

y las caracteísticas calculadas a partir de los valores muestrales son aleatorias también.

Ahora bien, cuando se emiten conclusiones sobre una población a partir sólo de valores obtenidos sobre una muestra aleatoria, están afectados de errores debidos al muestreo. Pero se tiene generalmente errores de medición que pueden influir sobre la precisión de las conclusiones también. Estos errores de medición pueden tener un caracter aleatorio y entonces pueden compensarse o ser sistemáticos.

Es importante observar que los errores de muestreo decrecen cuando el tamaño de la muestra crece, pero que 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 dificil en la práctica evaluar los errores de medición.

Consideramos aquí 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 a producir una muestra representativa de la población de donde estan extraidas, es decir para 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 ambig$\ddot{u}$edad.

El muestreo aleatorio simple (m.a.s.) permite sacar muestras de tamaño dado todas equiprobables de una población finita o infinita, distinguiendo el m.a.s. con reemplazo del m.a.s. sin reemplazo.

Dado un experimento aleatorio $\mathcal{E}$ y una población (o espacio muestral) Ω de sucesos elementales, el conjunto de n realizaciones del experimento $\mathcal{E}$ es una muestra de tamaño n.

- **Una muestra aleatoria simple con reemplazo se obtiene realizando n repeticiones independientes del experimento $\mathcal{E}$, tomando sobre Ω los sucesos elementales equiprobables. Se obtiene entonces una n-tupla de Ω.**
- **Una muestra aleatoria simple sin reemplazo (o sin repetición) se obtiene de la población Ω realizando el experimento $\mathcal{E}$:**
  - **sobre Ω. Se obtiene un suceso ω_1 con equiprobabilidad;**
  - **sobre $\Omega \setminus \{\omega_1\}$. Se obtiene un suceso ω_2 con equiprobabilidad;**

– **sobre $\Omega \setminus \{\omega_1, \omega_2\}$. Se obtiene un suceso ω_3 con equiprobabilidad, etc... hasta completar la muestra de tamaño n.**

En el caso de 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 hay que dividir el número total de muestras sin reemplazo de tamaño n por el número de permutaciones de los n elementos de la muestra, es decir por $n!$. La probabilidad de cada muestra es igual entonces a:

$$\frac{(N-n)!}{N!}$$

En el muestreo aleatorio sin reemplazo se obtienen elementos de Ω, todos distintos. 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 ser seleccionadas. 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 cada 10. Este procedimiento tiene una ventaja práctica, pero obliga a controlar que estos pozos no tengan justamente algunas particularidades.

Puede asegurar una mejor representatividad que el muestreo aleatorio simple si las unidades de observación fueron clasificadas según un criterio, por ejemplo la profundidad, y que este criterio esta correlacionado con las variables de interés, entonces se tendra en la muestra pozos de todas las profundidades. Pero 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 puede seleccionarlas proporcionalmente al número de empleados; para la población de los campos agricolas, 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 -la comuna y el 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 podria 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 llevaria a demasiado estratos. Se puedria estratificar según la región, o bien proceder en dos etapas: seleccionar al azar algunas comunas (unidades de observación primarias) y enseguida etapas. En cada etapa sfe puede usar probabilidades iguales o desiguales.

El muestreo por conglomerados es un caso particular de muestreo por etapas, en lo cual 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 eligidas.

2 DISTRIBUCIONES EN EL MUESTREO

2.1 INTRODUCCION

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

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

Ejemplo 1: Se saca una muestra al azar de 500 ampolletas del mismo tipo en un proceso de producción ILUMINA 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 todas las observaciones provienen de una misma variable aleatoria X de distribución desconocida abstracta llamada **distribución de población**, el tiempo de vida de este tipo de ampolleta.

Ejemplo 2: Se saca una muestra al azar de 1000 chilenas mayores de 14 años y se mide sus tallas. En este caso la población no es abstracta ya que se podria medir la talla de todas las chilenas mayores de 14 años y entonces determinar la distribución de población, que es discreta, y por ejemplo calcular la talla media de la población. Sin embargo es muy dificil realizar en la práctica y la función de distribución de población se considera como abstracta y se tomara como continua (normal en general) dado que el tamaño de la población es muy elevado.

Si se tiene una sola variable aleatoria X cuya función de distribución F de población es generalmente desconocida, obteniendo observaciones de esta variable X, buscaremos conocer a la función de distribución F. Se le da en general una expresión **teórica**. Los valores $X_1, X_2, ..., X_n$ de una v.a. X obtenidos sobre una muestra de tamaño n son **los valores muestrales**.

Se quiere saber entonces de que forma estos valores muestrales procuren información sobre algunas características de la población. No es posible de responder directamente a esta pregunta. Hay que transformar en otra pregunta: **si tenemos tal población,¿cual seria la probabilidad de obtener la muestra que obtuvimos?**

Se busca entonces, por ejemplo, **estimar** la media de la distribución F a partir de los valores muestrales. Esto tendrá sentido si la muestra es **representativa** de la población.

2.2 TIPOS DE VARIABLES

La cantidad y la naturaleza de las cacterísticas que se puede medir sobre los elementos de una población Ω son de varios tipos. Supondremos aquí una sola variable que es una función

X: $\Omega \longrightarrow Q$. Se distingue la naturaleza de la variable X según el conjunto Q:

- variable cuantitativa (también llamada intervalar) si Q es un intervalo de $\mathbb{R}$ o todo $\mathbb{R}$. Por ejemplo, la ead, el peso o la talla de una persona. Estas variables se consideran como real continua aún si se miden de manera discontinua (en año, en kg o cm).
- variable discreta si Q es un subconjunto de $\mathbb{N}$. Por ejemplo, el número de hijos de una familia.
- variable cualitativa (o nominal) si Q es un conjunto finito de atributos (o modalidades) no numéricos. Por ejemplo, el estado civil, el sexo o la ocupación de una persona.
- variable ordinal si Q es un conjunto de atributos no numéricos que se pueden ordenar. Por ejemplo, el ranking de la críticca cinematográfica.

Los métodos estadísticos depende 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 de la variables edad requerida y los métodos utilizados se usara la edad como variable cuantitativa o nominal.

2.3 FUNCION DE DISTRIBUCION EMPIRICA

2.3.1 Caso de variables numericas (reales o enteras)

Sean $X_1, X_2, ..., X_n$, los valores muestrales obtenidos de un m.a.s..

se define 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 $F_n(x)$ tiene las propiedades de una función de distribución: $F_n(x)$ es monotona no decreciente; tiene limites a la derecha y a la izquierda; es continua a la derecha; $F(-\infty) = 0$; $F(+\infty) = 1$. Además sus puntos de discontinuidad son en número finito y son con salto;

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 $F(x)$, o sea $nF_n(x)$ sigue una distribución binomial: $nF_n(x)A \sim \mathcal{B}(n, F(x))$.

Teorema 2.1 *Para todo x, $F_n(x)$ converge casi-seguramente hacia el valor teórico $F(x)$.*

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

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

O sea que $F_n(x) \xrightarrow{c.s.} F(x)$

Se tiene dos otros resultados que no se demuestran.

Teorema 2.2 *(Glivenko-Cantelli)*

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

Teorema 2.3 *(Kolmogorov) La distribución asintótica de D_n es conocida y no depende de X:*

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

2.3.2 Caso de variables no son numéricas (nominal u ordinal)

Cuando las variables no son numéricas, Q es un conjunto finito $Q = \{q_1, q_2, ..., q_r\}$. Se define la población por las probabilidades $¶(X = q_j) = p_j \ (\forall j = 1, ..., r)$.

Dada una muestra aleatoria simple $X_1, X_2, ..., X_n$ de tamaño n, se define las proporciones en el muestreo $f_n(q_j) = \frac{Card\{X_i = q_j\}}{n}$, $j = 1, ..., r$.

$\forall q_j \in Q)$, $nf_n(q_j)$ $Binomial(n, p_j)$ o sea que $f_n(q_j) \xrightarrow{c.s.} p_j$. Las r v.a. binomiales $nf_n(q_j)$ no son independientes entre si: $\sum_j nf_n(q_j) = n$.

Veremos ms adelante que estas v.a. forman un vector aleatorio *multinomial*.

2.4 DISTRIBUCIONES EN EL MUESTREO Y EN LA POBLACION

Sea una v.a. X de distribución F. Sean $X_1, X_2, ..., X_n$ valores muestrales independientes obtenidos sobre una muestra de tamaño n. Si nos interesamos a la media μ de la población (esperanza de la distribución F), la muestra nos permitirá **estimar** a μ; pero si se saca otra muestra del mismo tamaño obtendremos posiblemente otra estimación. **El resultado de la estimación es aleatorio**. El caracter aleatorio del resultado proviene de la aleatoriedad del muestreo y depende del tamaño y del tipo de muestreo efectuado. Es decir que los valores muestrales y resumenes de estos que sirven a la estimación son variables aleatorias. La determinación de las distribuciones de las estimaciones permite de inferir a la población.

Definición 2.1 *Las funciones de los valores muestrales son variables aleatorias llamadas* **estadísticos** *y las distribuciones de los estadísticos se llaman* **distribuciones en el muestreo**.

Se supone que la distribución de población pertenece a una familia de distribuciones teóricas, por ejemplo la distribución normal, la distribución beta o la distribución de Poisson. Sólo

algunas características quedan desconocidas, en este caso, como por ejemplo la media y la varianza. Estas características desconocidas, son los **parámetros** de la distribución de población, que se buscan a estimar. Los estadísticos y sus distribuciones en el muestreo (o sus distribuciones asintóticas cuando el tamaño n tiende a $+\infty$) permiten **estimar** los parámetros desconocidos de la distribución de población.

Las fluctuaciones de un estadístico S en el muestreo se miden con respecto al parámetro θ asociado a la población. Se usa el **error cuadrático medio** $E(S-\theta)^2)$ o su raíz llamada **el error estándar**, que permite medir la precisión del estadístico con respecto al parámetro θ.

2.4.1 Proporción muestral

Suponemos $X_1, X_2, ..., X_n$ los valores muestrales independientess obtenidos de una población de Bernoulli de parámetro p desconocido.

Consideramos en primer lugar el caso de una población infinita o una población finita con un muestreo 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 de una moneda. El parámetro p esla probabilidad de sacar "cara", que vale $\frac{1}{2}$ en el caso de una monedad 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 desconocida es la probabilidad de que una pieza sea defuectuosa en este proceso y $1-p$ esla probabilidad que no sea defectuosa.

Se define la proporción muestral como $f_n = \sum_{i=1}^{n} \frac{X_i}{n}$ la proporción de caras (o piezas defectuosas) encontradas entre las n observadas. La distribución de nf_n esta totalmente definada: nf_n toma valores en $\{0, 1, ..., n\}$ y sigue una distribución $Binomial(n, p)$, con

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

$E(f_n) = p$ 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: $e(f_n - p) = \sqrt{\frac{p(1-p)}{n}}$

Observamos que se tiene la convergencia en media cuadrática: $e[(f_n - p))^2] \longrightarrow 0$

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 o por la ley debil de los grandes números: la diferencia $\|f_n - p\|$ es tal que para un ϵ dado:

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

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

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

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 hipergeometrica:

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

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

Si el tamaño N de la población es grande, se tienen los mismos que en el un 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 e dado es casi independiente del tamaño N de la población:

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

Se presenta a continuación los tamaños muestrales necesarios para obtener un error $e = 0.05$ cuando $p = 0.5$:

N	500	1000	5000	10000	50000	∞
n	83	91	98	99	100	100

2.4.2 Media muestral

Sean $X_1, X_2, ..., X_n$, los valores muestrales independientes e identicamente distribuidos (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 y varianza σ^2 ($E(X_i) = \mu$ y $Var(X_i) = \sigma^2$ para todo i), entonces $E(\bar{X}_n) = \mu$. Lo que significa que **en promedio** los valores dados por los $\bar{X}_n$ coinciden con la media μ de la población. Pero para una muestra dada, el valor $\bar{X}_n$ se encontrara en general un poco por debajo o encima de μ debido a las fluctaciones del

muestreo. La pregunta entonces es de evaluar esta fluctación. La repuesta esta dada por la varianza de $\bar{X}_n$, es decir la dispersión promedio de $\bar{X}_n$ alrededor de μ:

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

Observamos que la dispersión de los valores de $\bar{X}_n$ alrededor de μ disminuye cuando el tamaño n de la muestra crece.

Además con la desigualdad de Chebychev, para un ϵ dado, se tiene:

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

Nota: 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 con reemplazo.

Si además la distribución de población es normal entonces la distribución en el muestreo de $\bar{X}_n$ también lo es. 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 DEL LIMITE CENTRAL).

Teorema 2.4 *(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 priviegiado de la distribución normal. Pero 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. $\{X_1, X_2, ..., Xn\}$, con $\mathrm{E}(X_i) = \mu$ y $\mathrm{Var}(X_i) = \sigma^2$ para tod i. Se define la **varianza muestral** o **varianza empírica** a 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:

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

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 $\bar{X}_n^2 \xrightarrow{c.s.} [E(X)]^2$).
- $S_n^2 \xrightarrow{m.c.} \sigma^2$ $\quad$ ($E((S_n^2 - \sigma^2)^2) \longrightarrow 0$).
- 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.
  $Var(S_n^2) \approx \frac{\mu_4 - \sigma^4}{n} \longrightarrow 0$.
- Cálculo de $Cov(\bar{X}_n, S_n^2)$
  $Cov(\bar{X}_n, S_n^2) = E((\bar{X}_n - \mu)(S_n^2 - \frac{n-1}{n}\sigma^2))$
  $Cov(\bar{X}_n, S_n^2) = E((\frac{1}{n}\sum X_i - \mu)(\frac{1}{n}\sum(X_j - \mu)^2 - (\bar{X}_n - \mu)^2 - \frac{n-1}{n}\sigma^2))$
  $Cov(\bar{X}_n, S_n^2) = E((\frac{1}{n}\sum(X_i - \mu))(\frac{1}{n}\sum(X_j - \mu)^2 - (\bar{X}_n - \mu)^2 - \frac{n-1}{n}\sigma^2))$
  $E(X_i - \mu) = 0 \ \ \forall i$ y $E(X_i - \mu)(X_j - \mu) = 0 \ \ \forall (i,j)$
  $Cov(\bar{X}_n, S_n^2) = \frac{1}{n^2}E(\sum(X_i - \mu)^3) - E((\bar{X}_n - \mu)^3)$
  $Cov(\bar{X}_n, S_n^2) = \frac{1}{n^2}E(\sum(X_i - \mu)^3) - \frac{1}{n^3}E(\sum X_i^3)$
  $Cov(\bar{X}_n, S_n^2) = \frac{\mu_3}{n} - \frac{\mu_3}{n^2} = \frac{n-1}{n^2}\mu_3$
  Si n $\to +\infty$, $Cov(\bar{X}_n, S_n^2) \to 0$ (lo que no significa que hay independencia).
  En particular 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

$X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. $\Longrightarrow \bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$

$S_n^2 = \frac{1}{n}\sum (X_i - \mu)^2 - (\bar{X}_n - \mu)^2$

$\frac{nS_n^2}{\sigma^2} = \sum (\frac{X_i - \mu}{\sigma})^2 - (\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}})^2$

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_n^2. Por otro lado, $(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}})^2$ sigue una distribución χ^2 con 1 grado de libertad y por otro lado,

En efecto recordemos en primer lugar la distribución de $Y = Z^2$, en que $Z \sim \mathcal{N}(0,1)$.
Sea $\Phi(x)$ la función de distribución de $Z \sim \mathcal{N}(0,1)$ y $F(y)$ la distribución de $Y = Z^2$.
$F(y) = P(Y \leq y) = P(Z^2 \leq y) = P(-\sqrt{y} \leq Z \leq \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$.

Se deduce la función de densidad 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 χ_1^2).

Observando que la χ_1^2 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 $U = \sum_1^n Y_i = \sum_1^n Z_i^2$ en que las Z_i^2 son χ_1^2 independientes, entonces
$\Psi_U(t) = (\frac{1}{1-2t})^{n/2}$, que es la f.g.m. de una distribución $Gamma(\frac{n}{2}, \frac{1}{2})$.

Se deduce así la función de densidad de U la v.a. χ_n^2, una Ji-cuadrado con n g.l.:

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

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

Corolario 2.1 *La suma de k v.a. 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 2.5 *Si $X_1, X_2, ..., X_n$ 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. y una transformación ortogonal $\underline{Y} = B\underline{X}$ tal que la primera fila de B es igual a $(1/\sqrt{n}, ..., 1/\sqrt{n})$. Se tiene entonces que:

- $Y_1 = \sqrt{n}\bar{X}_n$
- $\sum Y_i^2 = \sum X_i^2 = \sum (X_i - \bar{X}_n)^2 + n\bar{X}_n^2$
  $Y_2^2 + ... + Y_n^2 = nS_n^2$
- $(Y_1 - \sqrt{n}\mu)^2 + Y_2^2 + ... + Y_n^2 = (X_1 - \mu)^2 + ... + (X_n - \mu)^2$

La densidad conjunta de $Y_1, ..., Y_n$ es entonces proporcional a:

$$exp\{-(y_1 - \mu\sqrt{n})^2 + Y_2^2 + ... + Y_n^2\}/2\sigma^2$$

Luego $Y_1^2, ..., Y_n^2$ son independientes y

$$\sqrt{n}\bar{X}_n = Y_1 \sim \mathcal{N}(\sqrt{n}\mu, \sigma^2)$$

$$nS_n^2/\sigma^2 = Y_2^2 + ... + Y_n^2\}/\sigma^2 \sim \chi^2_{n-1}$$

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

Teorema 2.6 *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 las X_i provienen de una distribución normal.*

La demostración se deduce del teorema 2.4 y del corolario 2.1.

Definemos a continuación la distribución **t de Student** (Student es un seudónimo utilizado por el estadístico inglés W. S. Gosset para publicar), que tiene muchas aplicaciones en inferencia estadística como la distribución χ^2.

Definición 2.2 *Si X e Y son dos v.a. independientes, $X \sim \mathcal{N}(0,1)$ e $Y \sim \chi^2_n$, entonces la v.a. $T = \frac{X}{\sqrt{\frac{Y}{n}}}$ tiene una distribución t de Student a n grados de libertad.*

Buscamos la función de densidad de la v.a. 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 $f(x,y) = f_1(x)f_2(y)$.

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

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

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

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

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

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

Se observa que la función de densidad de T es simétrica y $E(T) = 0$ y $var(T) = \frac{n}{n-1}$ para $n \geq 2$. Ademas para $n = 1$ se tiene la distribución de Cauchy y para n 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.

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

es una t de Student con $n-1$ grados de libertad denotada t_{n-1}.

2.4.5 Valores extremos

Es importante estudiar más precisamente la distribución F considerando entre que valores podrían estar los valores muestrales, si hay simétria, entre otro.

Se define $X_{(1)}, ..., X_{(n)}$ **los estadísticos de orden**, como los valores muestrales ordenados de menor a mayor: $(X_{(1)} \leq X_{(2)} ... \leq X_{(n)})$. Los estadísticos de orden son variables aleatorias y nos interesamos en particular en $X_{(1)} = min\{X_1, ..., X_n\}$ y $X_{(n)} = max\{X_1, ..., X_n\}$.

En el curso de Probabilidades se estudio las distribuciones de estos estadísticos de orden en función de la distribución de población F(x) de X. En particular:

- 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.

2.4.6 Cuantilas muestrales

Definición 2.3 *Dada una función de distribución $F(x)$ de X, se llama cuantila de orden α al valor x_α tal que $F(x_\alpha) = \alpha$.*

En el caso empírico, 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 **cuartilas** a $x_{1/4}$ y $x_{3/4}$ y **intervalo intercuartila** a $x_{3/4} - x_{1/4}$.

Se observara que para una distribución discreta o empírica F_n una cuantila 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 muy interesantes para estudiar la asimetría de una distribución.