Universidad Tecnológica Nacional

Facultad Regional Resistencia

Probabilidad y Estadística

Apuntes de Cátedra

Año 2019

Sobre este apunte

Este apunte corresponde al curso de Probabilidad y Estadística que se dicta en la currícula de Ingeniería en Sistemas de Información, Ingeniería Química e Ingeniería Electromecánica de la Facultad de Ingeniería de la Universidad Tecnológica Nacional.

Esta edición corresponde a una versión actualizada de los apuntes utilizados durante años anteriores en la misma asignatura, elaborados por el Prof. Mario Garber. A partir del año 2014, es modificado y actualizado por el Ing. Guillermo Castro.

Capítulo 1
Conceptos Preliminares

1.1. La Estadística en la Ciencia y la Ingeniería

La finalidad de la ingeniería (en forma general) radica en la resolución de problemas de interés social mediante la aplicación eficiente de principios científicos. Para lograr esto, los ingenieros mejoran un producto existente (o proceso) o bien diseñan uno nuevo mediante la utilización de lo que se conoce como **método científico**, el cual puede formularse como la aplicación de una serie de pasos:

1. Describir en forma clara y concisa el problema.
2. Identificar, al menos aproximadamente, los principales factores que afectan al problema o que puedan jugar un papel importante en su solución.
3. Modelar el problema. Establecer las limitaciones o suposiciones inherentes a él.
4. Realizar experimentos y recolectar datos para validar el modelo o las conclusiones (pasos 2 y 3).
5. Refinar el modelo en función de los datos observados.
6. Manipular el modelo para desarrollar una solución del problema.
7. Conducir un experimento apropiado para confirmar que la solución propuesta es efectiva y eficiente.
8. Extraer conclusiones o realizar recomendaciones en base a los resultados obtenidos.

Estos pasos se encuentran representados en la figura (1.1). Es importante destacar la existencia de una fuerte interacción entre el problema, los factores que pueden afectar su solución, el modelo del fenómeno y la validación experimental del modelo y de la solución propuesta. Los pasos 2-4 pueden requerir la realización de ciclos iterativos hasta alcanzar la solución final. En consecuencia, es necesario saber planificar "experimentos" eficientes, recolectar datos, analizar e interpretar los mismos y comprender como los datos observados están relacionados con el modelo propuesto del problema bajo estudio.

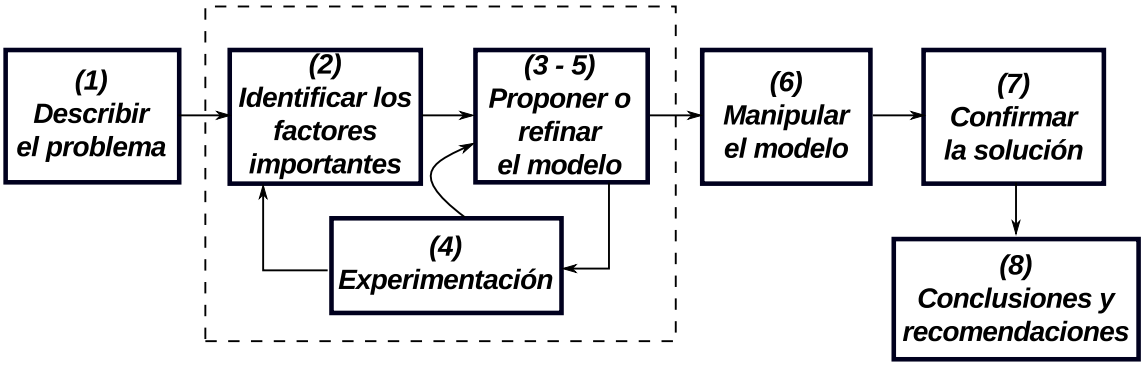


Figura 1.1: Pasos para la aplicación del método científico.

La **Estadística** y la **Probabilidad** juegan un papel predominante en la Ciencia y la Ingeniería actuales, particularmente en lo que se refiere a la recopilación y análisis de datos de cualquier tipo debido a la *componente aleatoria* que en muchas ocasiones interviene. Luego, en función del análisis realizado, el investigador o ingeniero debe tomar decisiones respecto de su objeto de análisis para lo cual debe contar con las herramientas adecuadas.

Estadística

El campo de la **Estadística** comprende la **recolección**, **ordenamiento**, **presentación** y **resumen** de datos, para posteriormente efectuar el **análisis**, extraer **conclusiones** y tomar, sobre esta base, **decisiones**. Las primeras cuatro acciones indicadas en la definición anterior conforman lo que usualmente se conoce como **"Estadística Descriptiva"**, mientras que las últimas tres forman parte de la **"Inferencia Estadística"**.

Por "componente aleatoria" nos referimos a que observaciones sucesivas de un sistema, experimento o fenómeno no producen exactamente el mismo resultado. Si analizamos por ejemplo el rendimiento de un auto ¿se realiza siempre el mismo kilometraje para cada tanque lleno? Sabemos que esto no siempre es así, siendo posible verificar algunos casos donde la variación puede resultar considerable. Esta variación va a depender de muchos factores, como ser el *tipo de entorno* en el cual se conduce (ciudad, autopista, campo), los cambios en las *condiciones del vehículo* en función del paso del tiempo (presión y estado de los neumáticos, condición del motor, desgastes de válvulas y correas, etc), el *tipo de combustible* que se emplea (menor o mayor octanaje) e incluso las *condiciones climáticas*. Estos factores representan potenciales fuentes de variación en el rendimiento, el cual puede considerarse como una **variable aleatoria** $\boldsymbol{X}$ (veremos detalladamente su definición en los capítulos posteriores), utilizando un modelo matemático como el siguiente:

$$X = \mu + \epsilon$$

donde μ es una constante y ϵ una *perturbación aleatoria*. La constante μ representaría los factores que permanecen invariantes en cada observación mientras que las pequeñas modificaciones del entorno, vehículo, combustible, etc, modifican el valor de ϵ.

Definiciones básicas

A continuación, enunciaremos algunas definiciones básicas sobre los conceptos que utilizaremos a lo largo de este libro:

Población

Se denomina **población** al conjunto de elementos, *finito o infinito*, que responden a una determinada característica.

Luego, el concepto de población en Estadística va más allá de la definición demográfica, esto es, la población de seres humanos exclusivamente. En Estadística, una población puede estar constituida por elementos de cualquier tipo, no solamente de seres humanos. Por ejemplo, se puede hablar de la población de viviendas de un barrio, de la población de alumnos de una Facultad, de la población de tornillos producidos por una fábrica, etc.

Muestra

Se denomina **muestra** al subconjunto de elementos pertenecientes a una población, utilizada para realizar estudios o investigaciones referidas a toda la población pero en menor tiempo y a un menor costo que si se la estudiara en forma exhaustiva.

Es importante aquí enfatizar que la economía conseguida al trabajar con una muestra se ve contrastada con una pérdida de exactitud, la cual sólo puede ser conseguida en estudios sobre la población completa. Es decir, la diferencia entre trabajar con la población o con una muestra puede relacionarse con la diferencia entre trabajar con un *censo* o con un *relevamiento* sobre una porción de la población. Un estudio basado en una muestra no garantiza exactitud pero permite realizar análisis estadísticos más rápidos y económicos que deben ser siempre acompañados con la información del grado de precisión con el que se ha trabajado.

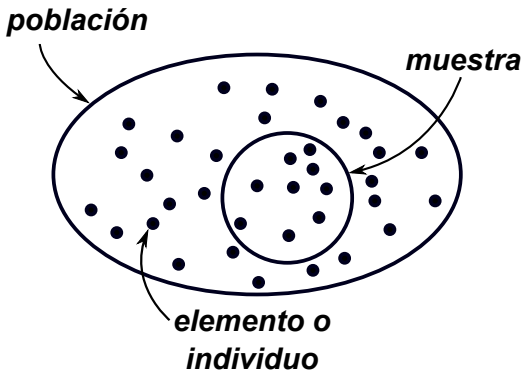


Figura 1.2: Representación gráfica de la población y una muestra.

Variable

Se denomina **variable** a una característica concreta de los individuos que conforman una población.

Una variable puede tomar diferentes valores dentro de un rango de variación determinado, denominado también "campo de variación". Por ejemplo:

- la *tensión de rotura* a la compresión o tracción de un determinado material,
- la *cantidad* de bits transmitidos en una red durante una cierta cantidad de tiempo,
- la *temperatura* de reacción de una determinada mezcla.

Estas variables pueden ser **cuantitativas** o **cualitativas**. Las variables cuantitativas se expresan por medio de cantidades (número, distancia, altura, peso, etc) mientras que las cualitativas representan atributos o cualidades (colores, sexo, razas, etc).

A su vez, las variables cuantitativas pueden clasificarse como **discretas** o **continuas**. Las continuas son aquellas que pueden tomar cualquier valor numérico dentro de su campo de variación. Por ejemplo, son variables continuas las alturas, las distancias, los pesos, etc. Las variables discretas sólo pueden tomar algunos valores dentro de su campo de variación: número de productos aceptados, resultado al arrojar una moneda o un dado, etc. En forma general, las variables continuas están relacionadas con operaciones de *medición*, mientras que las variables discretas se relacionan con operaciones de *conteo*.

Todas las variables utilizadas en Estadística son consideradas **aleatorias** y representan el resultado de un experimento, también denominado *aleatorio* dado que se realiza en condiciones de incertidumbre, y el cual se presenta con una cierta probabilidad de ocurrencia. Todas aquellas variables que no son de naturaleza aleatoria son denominadas **determinísticas**.

1.2. Muestreo

Muestreo

Método estadístico que permite realizar investigaciones referidas a una población, finita o infinita, seleccionando una muestra representativa de ésta.

Debido a que se trabaja con un subconjunto de la población (muestra), es posible caracterizar a las variables (cuantitativas o cualitativas) en menor tiempo y menor costo que si se estudiara a la población completa (censo), lo que finalmente terminaría siendo lento y caro, cuando no imposible a veces.

Sin embargo, como ya se ha mencionado anteriormente, no es posible garantizar la exactitud que sólo podría obtenerse consultando a toda la población. Además, en algunas ocasiones el muestreo es el único procedimiento para realizar cierto tipo de investigaciones en

las cuales las verificaciones son destructivas (prueba sobre la calidad de lámparas, fósforos, pruebas de impacto en vehículos, etc).

Un muestreo debe ser **representativo** y **confiable**. Por representativo se entiende que la muestra debe integrarse con una proporción de elementos similares a la composición existente en la población. La confiabilidad resulta ser, en cierto modo, consecuencia de lo anterior y constituye además un principio sobre la seguridad de que la toma de la información no producirá sesgos o errores en los resultados que se obtienen, tomando en consideración que un dato erróneo en la muestra incide en la confiabilidad mucho más que uno en la población.

En forma general, los tipos de muestreo pueden clasificarse como **intencionales** o bien **probabilísticos** o **aleatorios**. En un muestreo intencional, los elementos que integran la muestra son seleccionados arbitrariamente, lo cual permite efectuar la selección con rapidez pero sin garantizar la condición de representatividad mencionada anteriormente. No obstante esto suele ser utilizado en algunas ocasiones, con resultados poco confiables.

Por otro lado, en un muestreo probabilístico todos los elementos de la población analizada (también conocida como *población objetivo*) tienen una probabilidad asociada (igual o distinta según el caso) de ser elegidos para su inclusión en la muestra. En este tipo de muestreo se dan las condiciones básicas para el cumplimiento de la representatividad, porque todos los elementos están en condiciones de integrar la muestra sin excluir ninguno anticipadamente.

Tamaño de la muestra

El tamaño de la muestra, $\boldsymbol{n}$, no es arbitrario sino que depende de varios factores:

- **Variabilidad de la población:** En el caso extremo de que todos los elementos de la población fueran iguales, un único elemento sería suficiente. El tamaño de la muestra necesariamente deberá aumentar si la variabilidad de los elementos en la población aumenta y comienzan a diferenciarse entre ellos cada vez más. En consecuencia, existe una relación directa entre la variabilidad poblacional y el tamaño de la muesta.

- **Tamaño de la población ($\boldsymbol{N}$):** Una muestra de determinado tamaño puede ser suficiente para investigar una cierta población pero si el número de elementos que la constituyen duplicara su número, si bien sería necesario incrementar el tamaño de la muestra, no podría asegurarse que también deba incrementarse al doble (podría ser incrementada a más o a menos del doble). Esto significa que el tamaño de la muestra debe crecer en el mismo sentido que el tamaño de la población pero en una proporción que puede llegar a ser diferente, dependiendo de la variabilidad que asuma la población con su nuevo tamaño. En definitiva, el tamaño de la población también influye en forma directa en el tamaño de la muestra.

- **Margen de error o tolerancia ($\boldsymbol{d}$):** Si se desea realizar una investigación muestral y se exige que ella conduzca a un resultado sin errores, se deberá tomar un tamaño muestral igual al de la población, ya que esta es el única manera de no introducir errores. Es decir que con un margen de error cero (o tolerancia cero) el tamaño n de la muestra deberá ser igual al tamaño N de la población. Si se admite un margen de error mayor que cero, el tamaño de la muestra puede disminuir. En conclusión, a

medida que crece el margen de error admitido, disminuye el tamaño de la muestra, por lo que existe una relación inversa entre ellas.

En base a este análisis, es posible observar que el tamaño de la muestra es función de la variabilidad y el tamaño de la población, como así también del margen de error (o tolerancia) con el que se está dispuesto a trabajar. De estos tres elementos, la variabilidad poblacional es generalmente un valor que no puede conocerse anticipadamente.

Para poder resolver esta dificultad, se debe seleccionar una muestra inicial, denominada **muestra piloto**, de tamaño arbitrario $\boldsymbol{n_h}$. Este tamaño debe ser razonable, aunque mínimo, como para obtener una muestra compuesta por un primer conjunto de elementos (aplicando el criterio de selección probabilística) que permitirá conseguir una primera aproximación al valor de la variabilidad que se precisa conocer. Una vez obtenido, ese valor es utilizado en el cálculo de n para definir el número de elementos que se deberán seleccionar en la muestra, pudiendo ocurrir que:

a) $n_h < n \Rightarrow$ se aumenta n_h hasta llegar a n.

b) $n_h = n \Rightarrow$ se mantiene n_h.

c) $n_h > n \Rightarrow$ se mantiene n_h.

Selección probabilística

Como se mencionó anteriormente, mediante este procedimiento es posible asignar a cada elemento de la población una determinada probabilidad de ser incluido en la muestra. La selección puede ser realizada por medios manuales, asignando a los elementos de la población una numeración y luego efectuando un sorteo entre ellos por algún método apropiado (un bolillero por ejemplo) o por medios automáticos, instrumentando en una computadora un procedimiento de selección mediante un algoritmo que realice la tarea. Si la selección probabilística es correctamente implementada, es posible utilizar una muestra relativamente pequeña para posteriormente realizar inferencias sobre una población arbitrariamente grande.

Para la selección probabilística es muy común utilizar números aleatorios, los cuales puede obtenerse mediante tablas (método en desuso) o bien mediante medios computacionales. Las tablas de números aleatorios, que generalmente venían incorporadas como anexos en libros de estadística, estaban constituidas por arreglos numéricos de enteros dispuestos en filas y columnas agrupados de a 25 dígitos (a razón de 5 filas y 5 columnas por grupo). Por ejemplo, un grupo de 25 dígitos podría ser:

23874
07854
96453
17590
52086

Para iniciar la selección de números aleatorios se realizaba un ingreso aleatorio a la tabla, seleccionando al azar la página, la columna y la fila que permitirán encontrar el primer grupo de cinco dígitos que constituye el primer número aleatorio, de la serie de n

valores que se precisa. Una vez obtenido el primer grupo se continúa con los siguientes, considerando a la tabla como un texto que se debe leer de izquierda a derecha. Por ejemplo, sean los siguientes números aleatorios seleccionados para conformar una muestra de tamaño $n = 6$:

$$23874 \quad 56730 \quad 05628 \quad 34902 \quad 17472 \quad 96173$$

A continuación se convierte a estos seis grupos en números decimales agregando un cero a la izquierda, determinando así valores que pueden variar entre cero y uno:

$$0{,}23874 \quad 0{,}56730 \quad 0{,}05628 \quad 0{,}34902 \quad 0{,}17472 \quad 0{,}96173$$

Los números decimales así construidos se simbolizan con $\boldsymbol{x_{ai}}$ (número aleatorio i-ésimo) y se utilizan en la siguiente fórmula, cuya aplicación permite obtener todos los elementos e_i de la muestra:

$$e_i = \text{int}(N \cdot x_{ai} + 1) \qquad \text{para} \quad 1 \leq i \leq n \tag{1.1}$$

donde $\text{int}(x)$ es una función que devuelve el número entero más grande que sea inferior al argumento x. Con lo cual, si el tamaño de la población fuera $N = 120$, los seis elementos que integrarían la muestra son:

$$\begin{aligned}
e_1 &= \text{int}(120 \cdot 0{,}23874 + 1) = \text{int}(28{,}648 + 1) = 29 \\
e_2 &= \text{int}(120 \cdot 0{,}56730 + 1) = \text{int}(68{,}076 + 1) = 69 \\
e_3 &= \text{int}(120 \cdot 0{,}05628 + 1) = \text{int}(6{,}7536 + 1) = 7 \\
e_4 &= \text{int}(120 \cdot 0{,}34902 + 1) = \text{int}(41{,}882 + 1) = 42 \\
e_5 &= \text{int}(120 \cdot 0{,}17472 + 1) = \text{int}(20{,}966 + 1) = 21 \\
e_6 &= \text{int}(120 \cdot 0{,}96173 + 1) = \text{int}(115{,}407 + 1) = 116
\end{aligned}$$

con lo cual, los elementos número 29, 69, 7, 42, 21 y 116 conformarían la muestra.

Existen varios sistemas de selección probabilística, de los cuales destacaremos tres:

- **Muestreo simple al azar (MSA):** si es posible listar todos los elementos de una población de características homogéneas, es decir, de baja variabilidad, esta metodología es la forma más simple de obtener una muestra aleatoria. Consiste en seleccionar la muestra considerando que cada elemento de la población tiene la misma probabilidad de ser seleccionado. Es decir, una vez identificados los N elementos, se elige una muestra de tamaño n mediante alguno de los métodos comentados anteriormente, lo que convierte a este sistema en un procedimiento rápido y eficiente.

- **Muestreo estratificado:** cuando la población tiene una variabilidad importante, conviene dividirla en h estratos, los cuales conforman grupos de elementos internamente homogéneos (o con baja variabilidad) aunque heterogéneos entre los diferentes estratos conformados. Se calcula el tamaño n_i de cada estrato y se selecciona una muestra en cada estrato, de modo tal que el total de elementos se obtiene haciendo:

$$n = \sum_{i=1}^{h} n_i$$

Un ejemplo simple demuestra la utilidad del muestreo estratificado. Supongamos que se desea obtener una muestra aleatoria de alumnos que concurren a la universidad de tamaño $n = 100$ de una población de tamaño $N = 1000$, donde se sabe que hay 500 varones y 500 mujeres exactamente. Al observar esta situación vemos que, al menos teóricamente, al utilizar un *MSA* es posible seleccionar una muestra con ningún varón o ninguna mujer, por lo cual ésta carecería de representatividad. Para resolver esta situación conviene separar a la población en dos estratos varón-mujer y en cada uno de ellos realizar un *MSA* de 50 alumnos, con lo cual se garantiza que la proporción de géneros en la población se mantiene en la muestra.

Este sistema además permite asignar probabilidades diferentes a los elementos de distintos estratos, para darle una representación particular a cada estrato, como podría hacerse en el caso de seleccionar una muestra de establecimientos industriales donde los estratos se pueden constuir partiendo del número de empleados y asignando al estrato de mayor cantidad de empleados (que finalmente representarían a las empresas más importantes) una probabilidad mayor.

- **Muestreo sistemático:** Este sistema es conveniente utilizarlo cuando la población se encuentra ordenada en una secuencia donde se pueda asegurar que el atributo a analizar es aleatorio. Si no existe un orden debido a la participación de alguna variable, este sistema es tan conveniente como el muestreo simple al azar pero operativamente más cómodo. Consiste en definir el tamaño de la muestra y posteriormente dos valores denominados $\boldsymbol{a}$ (arranque) y $\boldsymbol{p_r}$ (progresión), donde:

$$p_r = \frac{N}{n} \qquad \text{y} \qquad 1 \leq a \leq p_r$$

En primer lugar se obtiene p_r y luego se calcula a mediante la siguiente fórmula:

$$a = \text{int}(p_r \cdot x_a + 1)$$

donde x_a es un número aleatorio. La muestra se conforma eligiendo a los elementos a partir de a y adicionando sucesivamente p_r.

Existe un potencial problema al utilizar un muestreo sistemático y yace en la posibilidad de que la población presente características repetitivas o cíclicas de la variable bajo estudio. Por ejemplo, si se desean extraer muestras de la producción de una máquina de la cual se conoce que introduce una falla cada 50 productos, si el muestreo sistemático coincide con esta frecuencia, los datos obtenidos no serán representativos de la población.

Capítulo 2
Estadística Descriptiva

Generalmente, luego de la realización de un experimento o de la observación de un fenómeno aleatorio, se obtiene una gran cantidad de datos con los cuales se precisa realizar un determinado tipo de análisis para posteriormente emitir conclusiones. Para poder presentarlos y analizarlos objetivamente sin "perderse" en una inmensidad de datos, es preciso resumirlos en forma adecuada. Para ello es posible utilizar tanto herramientas gráficas como matemáticas.

2.1. Distribución de frecuencias

Los datos recolectados a partir de una muestra se encuentran generalmente en forma **no ordenada** por lo cual el primer procedimiento es, precisamente, el **ordenamiento** de los datos. El ordenamiento es simplemente un arreglo convencional de los datos obtenidos en una investigación muestral, colocándolos por ejemplo, de menor a mayor o viceversa. Con los datos ordenados es posible determinar el **rango**, el cual es la primer medida estadística que puede calcularse:

$$R = x_M - x_m \tag{2.1}$$

donde x_M y x_m son el valor máximo y mínimo, respectivamente, del conjunto de datos.

Cuadro de distribución de frecuencias

El cuadro de distribución de frecuencias es una forma de presentación de los datos que facilita su tratamiento conjunto, posibilitando una comprensión diferente sobre ellos. A partir de su construcción, los datos pierden individualidad (se deja de conocer el valor de cada uno) debido a que se presentan agrupados en *clases* o *categorías* denominados **intervalos de clase**. A modo de ejemplo, a continuación se presentan los cuadros de distribución de frecuencias de distintas variables:

Tabla 2.1: Cantidad de expedientes iniciados en una oficina

N° de expedientes	f_i	**VL**	x_i	F_i	h_i (%)	H_i (%)
$1-6$	1	0,5	3,5	1	5,0	5,0
$7-12$	4	6,5	9,5	5	20,0	25,0
$13-18$	7	12,5	15,5	12	35,0	60,0
$19-24$	6	18,5	21,5	18	30,0	90,0
$25-30$	2	24,5	27,5	20	10,0	100,0
	$\sum f_i = \mathbf{20}$				$\sum h_i = \mathbf{100{,}0}$	

Tabla 2.2: Nivel de acidez del vino depositado en toneles en una bodega.

Acidez (pH)	f_i	**VL**	x_i	F_i	h_i (%)	H_i (%)
$1{,}00-1{,}49$	5	1,00	1,25	5	5	5
$1{,}50-1{,}99$	18	1,50	1,75	23	18	23
$2{,}00-2{,}49$	42	2,00	2,25	65	42	65
$2{,}50-2{,}99$	27	2,50	2,75	92	27	92
$3{,}00-3{,}49$	8	3,00	3,25	100	8	100
	$\sum f_i = \mathbf{100}$				$\sum h_i = \mathbf{100}$	

Tabla 2.3: Consumo de energía eléctrica en las viviendas de un barrio.

Consumo (KWh)	f_i	**VL**	x_i	F_i	h_i (%)	H_i (%)
$5{,}00-9{,}99$	5	5,00	7,50	5	5	5
$10{,}00-14{,}99$	18	10,00	12,50	23	18	23
$15{,}00-19{,}99$	42	15,00	17,50	65	42	65
$20{,}00-24{,}99$	27	20,00	22,50	92	27	92
$25{,}00-29{,}99$	8	25,00	27,50	100	8	100
	$\sum f_i = \mathbf{100}$				$\sum h_i = \mathbf{100}$	

La construcción de cada uno de los cuadros anteriores implica un cierto orden de procedimientos como así también nuevas definiciones, a saber:

a) Intervalo de clase. Es un intervalo de variación de los datos entre dos valores dados. Constituye la primer columna del cuadro, la cual lleva como título el nombre de la variable bajo estudio. Existen fórmulas que permiten obtener una guía para la elección del *número de intervalos de clase* (NI):

$$NI = 1 + 3{,}3 \log n \quad \text{(Fórmula de Sturges)}$$
$$NI = 2\sqrt[3]{n} \quad \text{(Fórmula de Rice)}$$

donde n es el número total de elementos en la muestra. Estas fórmulas son meramente orientativas y sólo sirven como guía en la elección de un NI óptimo. Normalmente una

distribución debería tener entre 5 y 12 intervalos de clase, dependiendo directamente de la cantidad de datos.

b) Frecuencia absoluta (f_i). Segunda columna del cuadro y representa la cantidad de casos que pertenecen a cada clase.

c) Límite inferior del intervalo de clase (LI). Es el menor valor de cada intervalo de clase. En cada uno de los intervalos, los límites inferiores son los valores ubicados a la izquierda. El LI del primer intervalo de clase debe ser, como mínimo, igual o menor que x_m, el cual es el menor valor del conjunto ordenado de datos.

d) Límite superior del intervalo de clase (LS). Es el mayor valor de cada intervalo de clase.

e) Verdadero límite o límite real del intervalo de clase (VL). El valor de este límite depende de si la variable es continua o discreta. Si la variable es continua, se conviene que los verdaderos límites coinciden con los límites inferiores, mientras que si es discreta este límite se obtiene haciendo la semisuma del límite superior del intervalo anterior y el límite inferior del intervalo considerado, es decir:

$$VL_i = \frac{LS_{i-1} + LI_i}{2}$$

f) Amplitud o tamaño del intervalo de clase (c). Diferencia, en valor absoluto, entre dos verdaderos límites consecutivos:

$$c_i = |\, VL_i - VL_{i+1}|$$

Una distribución de frecuencias que tiene todos sus intervalos de igual amplitud se denomina *equiespaciada* ($c_i = c = cte$). En este caso existe una relación entre el número de intervalos NI, el rango R y la amplitud c, expresada mediante la ecuación:

$$NI = \frac{R}{c} \Rightarrow c = \frac{R}{NI}$$

Esto significa que, conocido el rango, puede establecerse indistintamente uno de los otros dos valores: o la amplitud c o el número de intervalos NI. Si se decide construir una distribución con una amplitud determinada, se aplica la primer ecuación para obtener NI. En cambio, si se desea construir la distribución con un número de intervalos determinados, la amplitud se puede obtener mediante la segunda ecuación.

g) Punto medio del intervalo de clase (x_i). Se calcula haciendo la semisuma entre dos VL consecutivos, es decir:

$$x_i = \frac{VL_i + VL_{i+1}}{2}$$

Si la distribución de frecuencias fuera equiespaciada, a partir del primer punto medio se pueden obtener los siguientes sumándoles sucesivamente la amplitud c.

h) Frecuencia acumulada creciente o "menor que" (F_i). Total de elementos menores o iguales a un límite superior cualquiera LS_i. Se obtiene por adición sucesiva de las f_i desde el primer intervalo hasta el último. El resultado final debe coincidir con n.

i) Frecuencia relativa (h_i). Relación entre las frecuencias absolutas y el total de elementos n:

$$h_i = \frac{f_i}{n}$$

Suele expresarse en forma de porcentajes.

j) Frecuencia relativa acumulada (H_i). Su concepto es similar al de F_i pero en su cálculo intervienen las frecuencias relativas h_i.

2.2. Representación gráfica de la distribución de frecuencias

Histograma

Un **histograma** es un gráfico de la distribución de frecuencias, el cual se construye con rectángulos (barras) de superficie proporcional al producto de la amplitud por la frecuencia absoluta (o relativa) de cada uno de los intervalos de clase. A continuación se pueden observar los histogramas correspondientes a los ejemplos anteriores.

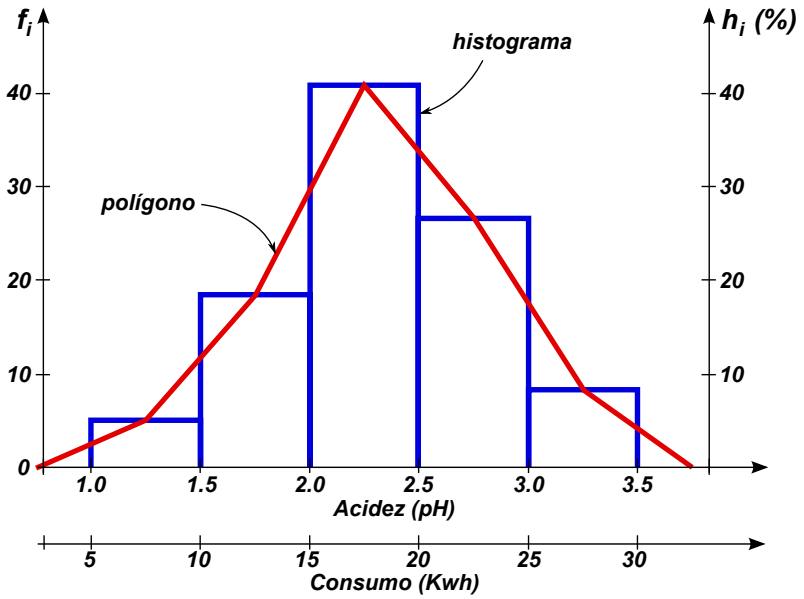


Figura 2.1: Histograma y polígono de frecuencias absolutas (izquierda) y relativas (derecha). Variable continua.

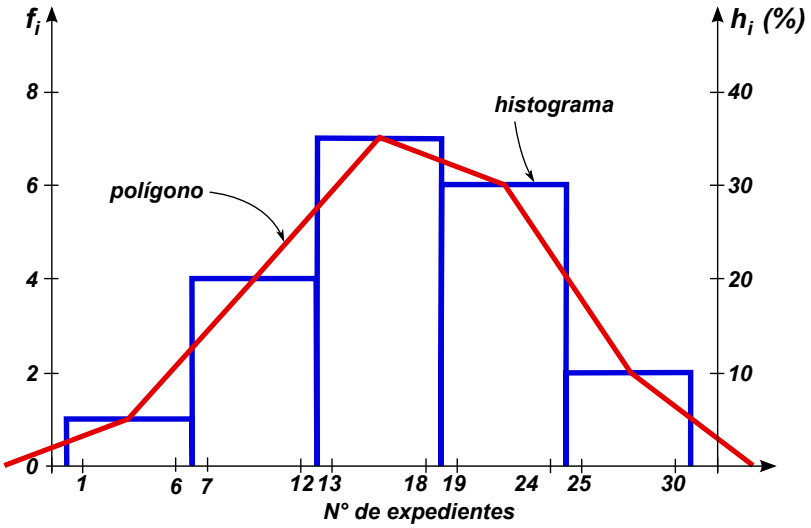


Figura 2.2: Histograma y polígono de frecuencias absolutas (izquierda) y relativas (derecha). Variable discreta.

Tanto en la figura (2.1) como en la (2.2) se han incluido dos ejes de ordenadas para indicar las frecuencias absolutas y relativas a la izquierda y derecha del gráfico, respectivamente. Esto permite obtener una representación de las dos escalas de frecuencia sin necesidad de repertir el gráfico. En la figura (2.2), la cual corresponde a la variable discreta "número de expedientes iniciados en una oficina", donde puede observarse con claridad que los rectángulos correspondientes a cada intervalo de clase se han construido entre los verdaderos límites.

Polígono de frecuencias

El polígono de frecuencias se obtiene a partir del histograma, uniendo los puntos medios de los lados opuestos a las bases de los rectángulos, incluyendo además el punto medio del intervalo de clase inmediato anterior al primer intervalo y el punto medio del intervalo inmediato superior al último intervalo, obteniendo así una figura poligonal cerrada con superficie similar a la del histograma.

Este gráfico permite obtener una representación aproximada y sumamente esquemática de la distribución teórica de la variable bajo estudio, la cual no puede ser nunca determinada exactamente a partir de muestras. Si en un histograma se disminuyera la amplitud de los intervalos (con lo cual se produciría un aumento del número de los mismos) al mismo tiempo que se aumentara el número total de observaciones, las variaciones del polígono serían cada vez más suaves. Cuando la amplitud tienda a cero, el número de intervalos deberá tender a infinito (junto con la cantidad de valores en la muestra), con lo cual el polígono se convertirá en una línea continua, denominada "distribución teórica de probabilidad".

Gráfico de frecuencias acumuladas. Ojivas

El gráfico de frecuencias acumuladas, también conocido como *ojiva*, es una representación de las frecuencias acumuladas y se construye, al igual que el histograma, con rectángulos de base y altura proporcionales a la amplitud y frecuencia (absoluta o relativa) de cada intervalo, respectivamente. La diferencia radica en que el inicio de cada rectángulo,

a partir del segundo intervalo, se desplaza hacia arriba, hasta coincidir con el nivel exacto de terminación del anterior. Cada punto determinado se une en forma consecutiva, obteniendo una línea poligonal creciente, denominada **ojiva creciente** o **"menor que"** dado que cada punto de coordenadas (x_i, F_i) o (x_i, H_i) sobre la ojiva representa la cantidad o porcentaje de valores menores a x_i.

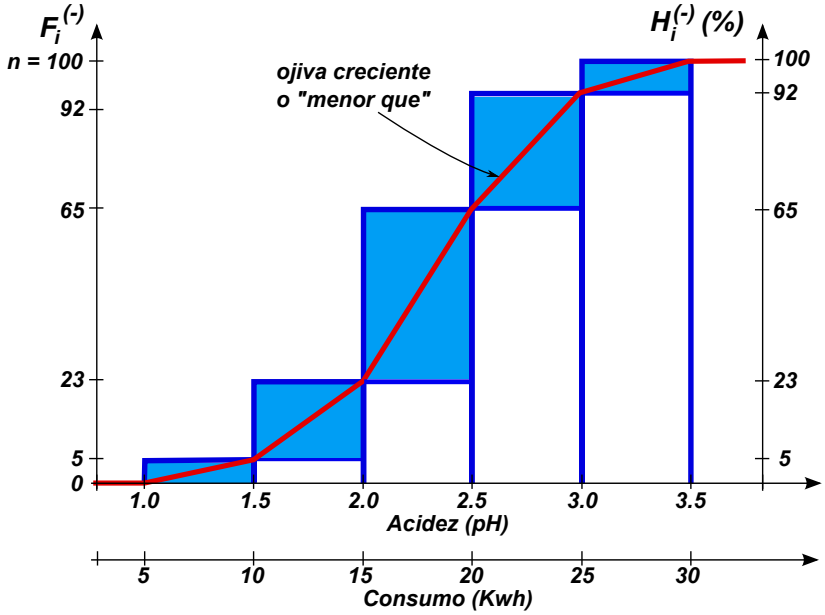


Figura 2.3: Frecuencias absolutas (izquierda) y relativas (derecha) acumuladas crecientes.

Un gráfico análogo al anterior puede construirse si en lugar se acumular las frecuencias partiendo desde cero se toma la cantidad total de datos o el 100 % de los datos según se trabaje con frecuencias absolutas o relativas y se van restando sucesivamente las superficies equivalentes a las frecuencias de cada intervalo, como se muestra en la tabla (2.4) y en la figura (2.4). Ahora se obtiene una línea poligonal decreciente llamada **ojiva decreciente** o **"mayor que"** donde para cada punto de coordenadas (x_i, F_i) o (x_i, H_i) sobre la ojiva se puede obtener la cantidad o porcentaje de valores mayores a x_i. Para diferenciar entre las frecuencias acumuladas crecientes y decrecientes se utiliza el supraíndice $(-)$ para las crecientes ("menor que") y el supraíndice $(+)$ para las decrecientes ("mayor que").

Tabla 2.4: Nivel de acidez del vino depositado en toneles en una bodega

Acidez (pH)	f_i	$F_i^{(-)}$	$F_i^{(+)}$	h_i (%)	$H_i^{(-)}$ (%)	$H_i^{(+)}$ (%)
1,00 – 1,49	5	5	100	5 %	5 %	100
1,50 – 1,99	18	23	95	18 %	23 %	95
2,00 – 2,49	42	65	77	42 %	65 %	77
2,50 – 2,99	27	92	35	27 %	92 %	35
3,00 – 3,49	8	100	8	8 %	100 %	8

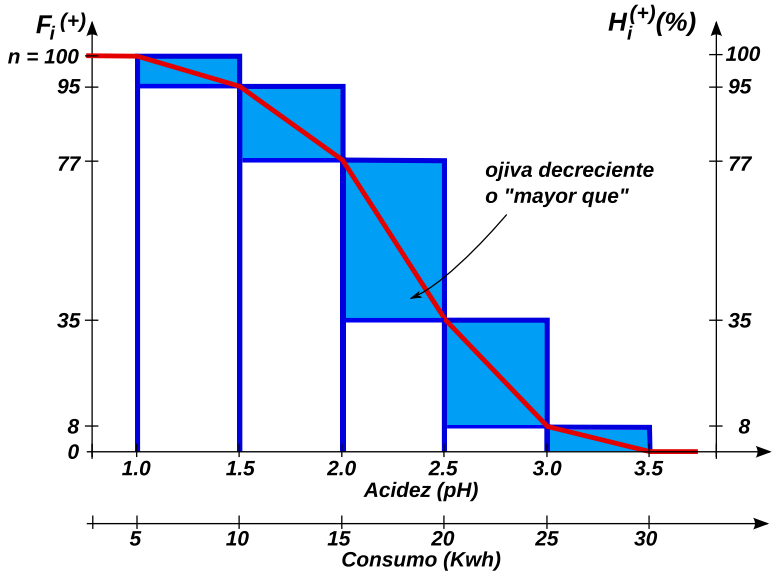


Figura 2.4: Frecuencias absolutas (izquierda) y relativas (derecha) acumuladas decrecientes.

Algunos comentarios

Cuando anteriormente se hizo referencia a los intervalos de clase se mencionó que el número de intervalos y el formato de la distribución de frecuencias dependían del investigador o del objetivo de la investigación. A lo largo de este capítulo se presentaron varios ejemplos de distribuciones de frecuencias, todas con las siguientes características:

- **Son equiespaciadas:** el tamaño o amplitud de todos los intervalos es constante. Esto significa que es posible construir distribuciones de frecuencias con intervalos de clase no equiespaciados, donde el tamaño de cada intervalo (c_i) resulte variable. Que la distribución sea equiespaciada resulta una importante ventaja para la organización del trabajo estadístico, debido a la simplificación que eso significa en el uso de las fórmulas matemáticas y en los cálculos a realizar.

- **Tienen intervalos cerrados:** esto significa que en todos los intervalos existen, perfectamente identificados, los Límites Inferior y Superior. Esto no significa que no puedan existir distribuciones donde algún intervalo no posea o Límite Inferior o Superior o ambos. El siguiente ejemplo ilustra este comentario:

Intervalos	f_i
menos de 450	3
450 - 499.99	5
500 - 599.99	12
600 - 699.99	10
700 - 799.99	4
800 - 899.99	-
900 - 999.99	-
1000 - 1099.99	-
1100 - 1199.99	-
1200 - 1299.99	-
1300 - 1399.99	-
1400 - 1499.99	-
1500 - 1599.99	1

Supongamos que el cuadro de la izquierda hace referencia a los ingresos mensuales del personal de una empresa y que incluye a un directivo cuyo sueldo es de $1590. Se puede observar que la distribución tiene trece intervalos de clase, donde el primero es abierto y los últimos ocho son cerrados, de los cuales siete tienen frecuencias absolutas nulas. Esta forma no resulta muy satisfactoria, por lo que es muy apropiado y conveniente construir una distribución de frecuencias como la presentada más abajo, agrupando los últimos intervalos en uno solo.

Intervalos	f_i
menos de 450	3
450 - 499.99	5
500 - 599.99	12
600 - 699.99	10
700 o más	5

En este caso, tanto el primero como el último intervalo de clase son abiertos (no tienen alguno de los dos límites), mientras que los demás intervalos son cerrados.

Se han resuelto algunos de los inconvenientes de la presentación anterior, ya que se han eliminado los intervalos con frecuencia absoluta igual a cero. Pero ahora aparece otra dificultad, que es la imposibilidad de calcular alguno de los elementos que forman parte del cuadro de distribución de frecuencias. En particular, no pueden calcularse los puntos medios, lo cual constituye un impedimento para efectuar posteriores pasos en el trabajo estadístico. No obstante esto, a menudo es necesario trabajar con distribuciones de intervalos de clase abiertos.

Por último debe observarse que una representación gráfica no es, en sí misma, un resultado final, sino una herramienta para describir, explorar y comparar datos. Las preguntas que debemos hacernos para extraer información de las mismas debieran ser:

- ¿Cuál es el valor aproximado del centro de la distribución y cuál es el rango de valores?
- ¿Están los valores distribuidos de manera uniforme? ¿La distribución se encuentra sesgada (ladeada) hacia la izquierda o hacia la derecha?
- ¿La distribución tiene picos? ¿Es simétrica?

2.3. Medidas de Posición

Las medidas de posición son valores que se calculan a partir del conjunto de datos y tienen la particularidad de ser representativos del mismo. Su nombre proviene de que indican sobre el eje de las abcisas la "posición" de todo el conjunto. También suelen denominarse **medidas de tendencia central** dado que muchas de las medidas de posición se ubican en el centro del conjunto de datos, aunque cabe aclarar que no todas las medidas de posición son medidas de tendencia central. De manera general, las podemos clasificar en dos tipos fundamentales:

a) Promedios. Se denominan así a las medidas de posición en cuyo cálculo intervienen todos los valores disponibles de la variable con la que se está trabajando. Por ejemplo, la *media aritmética* o promedio, la *media geométrica*, la *media armónica*, la *media cuadrática*.

b) Otras medidas de posición. Son todas aquellas en cuyo cálculo no intervienen todos los valores disponibles de la variable. Por ejemplo, la *mediana*, el *modo*, los *cuartiles*.

Media aritmética

La media aritmética, $\overline{\boldsymbol{x}}$, es la medida de posición por excelencia debido a la sencillez de su cálculo, al fácil manejo algebraico y a las amplias e interesantes propiedades que posee.

Media ponderada

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} x_i f_i}{n} \qquad \text{con } k \leq n \tag{2.2}$$

Esta fórmula se denomina **general** o **ponderada** dado que las f_i, que representan las frecuencias de aparición de los valores de la variable, ponderan ("pesan") a cada uno de ellos. Por ejemplo, en la tabla (2.5) tenemos valores de la variable x_i junto con sus respectivas frecuencias de aparición o repetición f_i para dos casos distintos:

$\boldsymbol{x_i}$	$\boldsymbol{f_i}$	$\boldsymbol{x_i f_i}$
1	1	1
3	3	9
5	5	25
10	2	20
	$\sum \boldsymbol{f_i} = \mathbf{11}$	$\sum \boldsymbol{x_i f_i} = \mathbf{55}$

$\boldsymbol{x_i}$	$\boldsymbol{f_i}$	$\boldsymbol{x_i f_i}$
1	1	1
3	1	3
5	1	5
10	1	10
	$\sum \boldsymbol{f_i} = \mathbf{4}$	$\sum \boldsymbol{x_i f_i} = \mathbf{19}$

Tabla 2.5: Operaciones para el cálculo de la media aritmética.

Para el ejemplo de la izquierda de la tabla (2.5) se obtiene $\overline{x} = 55/11 = 5$, mientras que para los datos a la derecha resulta $\overline{x} = 19/4 = 4{,}75$. Es posible observar en este último

caso que los valores de la primer y tercer columna coinciden, con lo cual también lo harán sus sumatorias, y como la sumatoria de las frecuencias resulta igual a la cantidad total de elementos observados, n, la ecuación (2.2) puede escribirse en forma simplificada como:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.3}$$

la cual se denomina **fórmula simple de la media aritmética**. Como característica importante debe notarse que la media aritmética posee la misma unidad de medida que la variable bajo estudio.

Media muestral y media poblacional - Simbología.

Es importante distinguir dos conceptos diferentes: el de **media muestral** y el de **media poblacional**. Si bien sus fórmulas de cálculo no son diferentes desde el punto de vista conceptual, están referidas a la muestra y a la población, respectivamente, por lo cual también se representan mediante símbolos diferentes. Mientras se ha visto que en el caso de la media aritmética muestral se utiliza el símbolo $\overline{\boldsymbol{x}}$, la media poblacional se simboliza con $\boldsymbol{\mu_x}$, donde el subíndice indica la variable con la que se está trabajando[1]. Si se desea obtener la fórmula de cálculo simple de la media aritmética poblacional, se tendría:

$$\mu_x = \frac{\sum_{i=1}^{N} x_i}{N} \tag{2.4}$$

donde puede notarse que no difiere sustancialmente de la fórmula simple de la media aritmética muestral, salvo que el promedio se está realizando ahora sobre todos los elementos de la población, $\boldsymbol{N}$. En los trabajos e investigaciones estadísticas muy pocas veces es posible calcular la media aritmética poblacional.

Propiedades de la media aritmética.

A continuación enumeraremos ciertas propiedades de la media aritmética que posteriormente permitirán arrojar conclusiones sobre su importancia como *estadística*[2]. Muchas de estas propiedades serán utilizadas en los capítulos posteriores de este apunte[3].

1) **La media aritmética de un conjunto definido de datos es un valor constante.** Esta propiedad no requiere demostraciones, sólo resulta apropiado explicar sencillamente que, dado un conjunto de datos, si éste no se modifica su media aritmética permanece constante.

2) **La sumatoria de los desvíos es igual a cero.** Los "desvíos", d_i, se obtienen haciendo la diferencia entre los valores de la variable y algún valor arbitrario A, es decir $d_i = x_i - A$. Si ese valor es la media aritmética, luego $d_i = x_i - \overline{x}$, denominado

[1] En muchas ocasiones utilizaremos directamente el símbolo μ, dado que generalmente se utiliza la variable x.

[2] Las medidas de posición y dispersión comúnmente son también denominadas "estadísticas".

[3] Se recomienda además la lectura del apéndice §A.1 sobre las propiedades del operador sumatoria.

simplemente *desvío*. Para todo otro valor distinto a la media aritmética se debe aclarar con respecto a qué valor se calculan los desvíos. La demostración de la 2^{da} propiedad resulta entonces:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\overline{x} = \sum_{i=1}^{n}x_i - n\overline{x} = \sum_{i=1}^{n}x_i - n\frac{\sum_{i=1}^{n}x_i}{n} = 0.$$

3) La sumatoria de los desvíos al cuadrado, entre los valores de la variable y un valor constante y arbitrario A, es un mínimo si $A = \overline{x}$. Esta propiedad es complementaria de la anterior y es posible demostrarla construyendo una función $\varphi(x, A) = \sum_{i=1}^{n}(x_i - A)^2$. Para un conjunto de valores x_i determinado, es posible hallar un punto crítico de la función aplicando las reglas correspondientes de derivación:

$$\frac{\partial\varphi}{\partial A} = \frac{\partial\sum_{i=1}^{n}(x_i - A)^2}{\partial A} = -2\sum_{i=1}^{n}(x_i - A) = 0.$$

Puede observarse que la derivación se ha realizado con respecto a la variable A, la cual es arbitraria y puede tomar cualquier valor entre $-\infty$ y $+\infty$. Luego, como $-2 \neq 0$, debe serlo la sumatoria, de allí que

$$\sum_{i=1}^{n}(x_i - A) = 0 \Rightarrow \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}A = 0,$$

y como A es constante para el operador sumatoria:

$$\sum_{i=1}^{n}x_i - nA = 0 \Rightarrow A = \frac{\sum_{i=1}^{n}x_i}{n} = \overline{x}$$

Por consiguiente, $A = \overline{x}$ es un punto crítico y se verifica que corresponde a un mínimo de la función φ:

$$\frac{\partial^2\varphi}{\partial A^2} = \frac{\partial\left[-2\sum_{i=1}^{n}(x_i - A)\right]}{\partial A} = 2n > 0$$

$\forall\ n$, dado que $n > 0$ (si $n = 0$ no hay muestra).

Esta última comprobación matemática es de alguna manera un resultado lógico, dado que al tener la función φ todos sus términos positivos, el punto crítico $A = \overline{x}$ debe ser una cota mínima dado que corresponde al centro del conjunto de datos, por definición de la media aritmética. Esto conducirá a que cualquier valor de A distinto de $\overline{x}$ (mayor o menor) arroje resultados mayores de la función φ.

4) Media aritmética de variables transformadas algebraicamente.

a) Si a todos los valores de una variable les sumamos o restamos un valor constante y arbitrario *A*, se obtiene una nueva variable cuya media aritmética será igual a la de la variable original sumada o restada por *A*. Supongamos tener una variable x_i que tiene una media $\overline{x}$, y un valor arbitrario A, con los cuales se construye una variable w_i, es decir, $w_i = x_i \pm A$. Luego, la media aritmética de la nueva variable w_i es:

$$\overline{w} = \frac{1}{n}\sum_{i=1}^{n} w_i = \frac{1}{n}\sum_{i=1}^{n}(x_i \pm A) \Rightarrow \frac{1}{n}\left[\sum_{i=1}^{n} x_i \pm nA\right] = \overline{x} \pm A.$$

b) Si a todos los valores de una variable los multiplicamos o dividimos por valor constante y arbitrario *c*, se obtiene una nueva variable cuya media aritmética será igual a la de la variable original multiplicada o dividida por *c*. Considerando sólo la multiplicación (la demostración por la división es análoga) tenemos $w_i = c\ x_i$, por consiguiente:

$$\overline{w} = \frac{1}{n}\sum_{i=1}^{n} w_i = \frac{1}{n}\sum_{i=1}^{n}(c\ x_i) \Rightarrow \frac{c}{n}\sum_{i=1}^{n} x_i = c\ \overline{x}.$$

c) Consideraremos ahora una combinación del caso a) y b). Se trata de una variable $u_i = \dfrac{x_i - A}{c}$, con lo cual:

$$\overline{u} = \frac{1}{n}\sum_{i=1}^{n} u_i = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - A)}{c} \Rightarrow \frac{1}{n}\frac{1}{c}\left[\sum_{i=1}^{n} x_i - nA\right] = \frac{\overline{x} - A}{c}$$

verificándose de esta manera que la media de una variable transformada algebraicamante mantiene la transformación de la variable original.

5) La media aritmética de la suma o diferencia de dos variables es la suma o diferencia de sus correspondientes medias aritméticas. Sean dos variables x_i e y_i, que tienen sus medias aritméticas iguales a $\overline{x}$ e $\overline{y}$, respectivamente. La nueva variable es $w_i = x_i \pm y_i$, luego:

$$\overline{w} = \frac{1}{n}\sum_{i=1}^{n} w_i = \frac{1}{n}\sum_{i=1}^{n}(x_i \pm y_i) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i \pm \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{x} \pm \overline{y}.$$

Desventaja de la media aritmética.

Supongamos que tenemos un sistema de masas con distintos pesos en un plano, como el que se muestra en la figura (2.5). Sabemos que las ecuaciones de equilibrio son:

$$\begin{aligned} \sum f_x &= 0 \\ \sum f_y = 0 &\Rightarrow R = \sum f_i \\ \sum M_i^0 = 0 &\Rightarrow R\ x_R = \sum f_i\ x_i \end{aligned}$$

donde f_x, $f_y \equiv f_i$ son las fuerzas en las direcciones x e y, respectivamente, R es la fuerza resultante (equilibrante) del sistema y x_i la distancia de las fuerzas al origen de coordenadas considerado, siendo x_R la distancia de la equilibrante del sistema al origen. Considerando las últimas dos ecuaciones tenemos que

$$x_R = \frac{\sum_{i=1}^{k} f_i \; x_i}{R} = \frac{\sum_{i=1}^{k} f_i \; x_i}{\sum_{i=1}^{k} f_i} \tag{2.5}$$

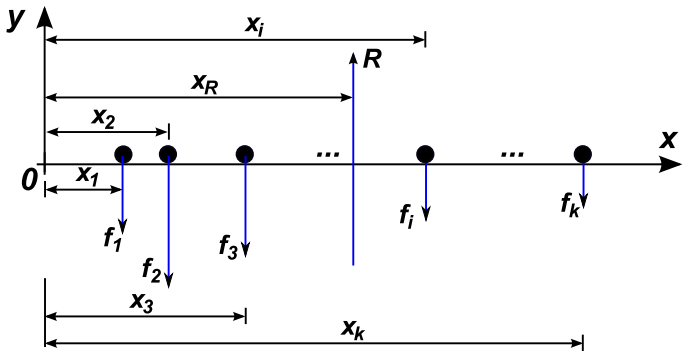


Figura 2.5: Analogía del cálculo de la resultante de un sistemas de fuerzas coplanares con el cálculo de la media aritmética.

La ecuación (2.5) nos remite a la definición de la media aritmética, ecuación (2.2). Es interesante notar que la media aritmética indica entonces el "punto de equilibrio" o "centro de gravedad" de un conjunto de valores, siendo por este motivo la medida de posición más utilizada. Sin embargo, esta observación también denota una característica de la media aritmética y es que resulta afectada por *valores extremos*, es decir, valores muy bajos o muy altos de la variable aleatoria pueden conducir a que el valor de la media no sea representativo de la ubicación del "centro" del conjunto de datos. Analizando la figura (2.5) se observa que cuanto más alejadas del conjunto de fuerzas se encuentren f_1 o f_k, la incidencia sobre el punto de aplicación de la resultante será mayor, dado que aumenta el "brazo de palanca" de la fuerza.

Cálculo de la media aritmética en distribuciones de frecuencias.

El cálculo de la media aritmética en distribuciones de frecuencia se realiza partiendo de las siguientes condiciones y supuestos:

- Una vez conformado el cuadro de distribuciones de frecuencias ya no se conocen los valores individuales de los datos.
- Los intervalos de clase tienen un límite inferior y un límite superior, pero estos no pueden ser tomados en cuenta para realizar el cálculo de la media aritmética.
- Los puntos medios de los intervalos de clase se convierten en los valores de la variable que permitirán realizar el trabajo de cálculo, aplicando la fórmula ponderada de la media aritmética, ecuación (2.2).

Continuando con los ejemplos anteriores, vemos que solamente precisamos de las primeras tres columnas de la distribución, dado que las restantes no son necesarias para este procedimiento de cálculo. Se adiciona una nueva columna, $x_i f_i$, que servirá para realizar la tarea de multiplicar cada punto medio por su correspondiente frecuencia absoluta y posteriormente efectuar la suma de esos resultados.

Tabla 2.6: Cantidad de expedientes iniciados en una oficina

N° de expedientes	f_i	F_i	x_i	$x_i f_i$
$1-6$	1	1	3,5	3,5
$7-12$	4	5	9,5	38,0
$13-18$	7	12	15,5	108,5
$19-24$	6	18	21,5	129,0
$25-30$	2	20	27,5	55,0
	$\sum f_i = \mathbf{20}$			$\sum x_i f_i = \mathbf{334,0}$

$$\Rightarrow \overline{x} = \frac{334}{20} = 16{,}7 \text{ expedientes.}$$

Tabla 2.7: Nivel de acidez del vino depositado en toneles en una bodega

Acidez (pH)	f_i	F_i	x_i	$x_i f_i$
$1{,}00-1{,}49$	5	5	1,25	6,25
$1{,}50-1{,}99$	18	23	1,75	31,50
$2{,}00-2{,}49$	42	65	2,25	94,50
$2{,}50-2{,}99$	27	92	2,75	74,25
$3{,}00-3{,}49$	8	100	3,25	26,00
	$\sum f_i = \mathbf{100}$			$\sum x_i f_i = \mathbf{232{,}50}$

$$\Rightarrow \overline{x} = \frac{232{,}5}{100} = 2{,}325 \text{ pH.}$$

Tabla 2.8: Consumo de energía eléctrica en las viviendas de un barrio

Consumo (KWh)	f_i	F_i	x_i	$x_i f_i$
$5{,}00-9{,}99$	5	5	7,50	38,50
$10{,}00-14{,}99$	18	23	12,50	225,00
$15{,}00-19{,}99$	42	65	17,50	735,00
$20{,}00-24{,}99$	27	92	22,50	607,50
$25{,}00-29{,}99$	8	100	27,50	220,00
	$\sum f_i = \mathbf{100}$			$\sum x_i f_i = \mathbf{1826{,}00}$

$$\Rightarrow \overline{x} = \frac{1826}{100} = 18{,}26 \text{ Kwh.}$$

Mediana

Mediana

La **mediana**, $\boldsymbol{M_e}$, es una medida descriptiva del centro de un conjunto de datos y está determinada por el valor de la variable que divide al conjunto de datos o distribución en dos partes iguales, dejando por debajo y por arriba de ella igual número de elementos.

Dado un conjunto de valores *no agrupados*, la mediana es un valor tal que el 50 % de los valores son *menores* o iguales a ella y el 50 % restante son *mayores* o iguales. Para determinarla, deben ordenarse las observaciones o mediciones del fenómeno estudiado (de mayor a menor o viceversa); si n es un número impar, la mediana es el valor de la observación que aparece en la posición $(n+1)/2$ y si es par, la mediana se define como el promedio de los valores de las observaciones que aparecen en los lugares centrales del arreglo, es decir, los valores que se encuentran en las posiciones $n/2$ y $n/2+1$.

- *a)* $2, 5, 8, 8, 9, 12, 15 \Rightarrow Me = 8$.
- *b)* $2, 5, 8, 8, 9, 12, 15, 18 \Rightarrow Me = 8{,}5$.
- *c)* $2, 5, 8, 8, 9, 12, 15, 100 \Rightarrow Me = 8{,}5$. Es posible observar que en este último ejemplo el valor de M_e es igual al de *b)*. Por lo tanto, es posible observar que la mediana no es afectada por valores extremos, como si ocurre en el caso de la media aritmética.

Para datos *agrupados* la mediana es el valor de la abscisa correspondiente a la línea vertical que divide al histograma en dos áreas iguales o bien, si se trabaja con el gráfico de frecuencias acumuladas "menor que" $(F_i^{(-)})$, el valor de la abscisa cuya frecuencia relativa acumulada es del 50 % (o $n/2$ si son frecuencias absolutas). Es posible deducir la ecuación que permite obtener el valor de M_e cuando se trabaja con el histograma de frecuencias absolutas siguiendo el siguiente procedimiento:

- Conocido n, se obtiene el valor $n/2$ y se identifica el intervalo de clase donde se encontrará M_e. Para ello se observa en qué intervalo de clase la frecuencia absoluta acumulada es igual o supera por primera vez al valor $n/2$, identificando los verdaderos límites inferior (VLI) y superior (VLS) del mismo. En la figura (2.6) dicho intervalo es el cuarto, donde la mediana es el valor de la abscisa correspondiente al segmento $\overline{\boldsymbol{LM}}$ que divide al histograma en dos áreas iguales.
- La frecuencia del intervalo de clase donde se encuentra M_e, denominada f_m, debe ser dividida en dos frecuencias, f_m^{izq} y f_m^{der}, tales que:

$$F_a + f_m^{izq} = f_m^{der} + f_{m+1} + f_{m+2} + f_k = n/2$$

es decir, la suma de las frecuencias absolutas a la izquierda y derecha de M_e debe resultar exactamente igual a $n/2$, con lo cual se verifica que las áreas a ambos lados del segmento $\overline{\boldsymbol{LM}}$ son iguales. F_a representa la *frecuencia acumulada anterior* al intervalo donde se encuentra M_e, es decir, $F_a = f_1 + f_2 + f_3$.

- Luego, se puede observar que en el intervalo de la mediana se cumple la siguiente relación:

$$\frac{\overline{AM}}{f_m^{izq}} = \frac{\overline{AB}}{f_m} \Rightarrow \overline{AM} = f_m^{izq}\frac{\overline{AB}}{f_m} \tag{2.6}$$

en donde se supone un incremento lineal dentro del intervalo.

- Finalmente, el valor buscado se obtiene haciendo $M_e = VLI + \overline{AM}$ por lo cual, considerando la ecuación (2.6) y teniendo en cuenta que $\overline{AB} = c$ (amplitud del intervalo) y $f_m^{izq} = n/2 - F_a$:

$$M_e = VLI + c\frac{n/2 - F_a}{f_m} \tag{2.7}$$

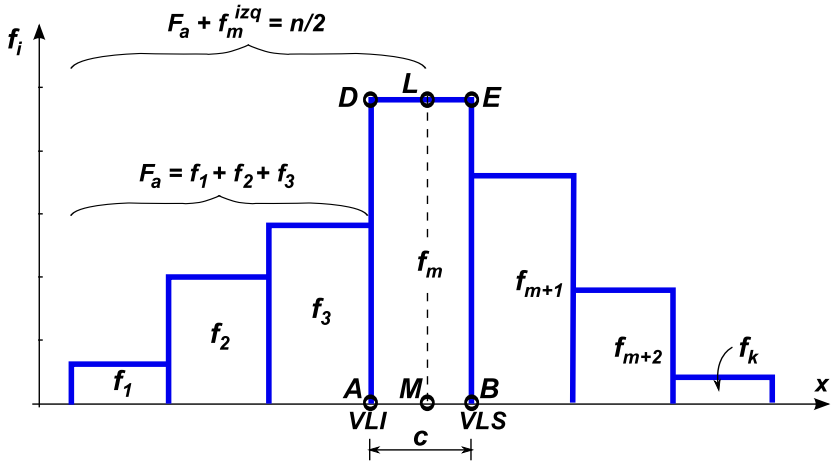


Figura 2.6: Determinación de la mediana para datos agrupados utilizando el histograma de frecuencias absolutas.

Otra forma de calcular la mediana es utilizando el polígono de frecuencias relativas acumuladas. En la figura (2.7) la mediana es la abscisa del punto $\boldsymbol{P}$, cuya ordenada es el 50 %. Para calcular ese valor podemos hacer uso de la semejanza entre los triángulos $\boldsymbol{RQP}$ y $\boldsymbol{RST}$:

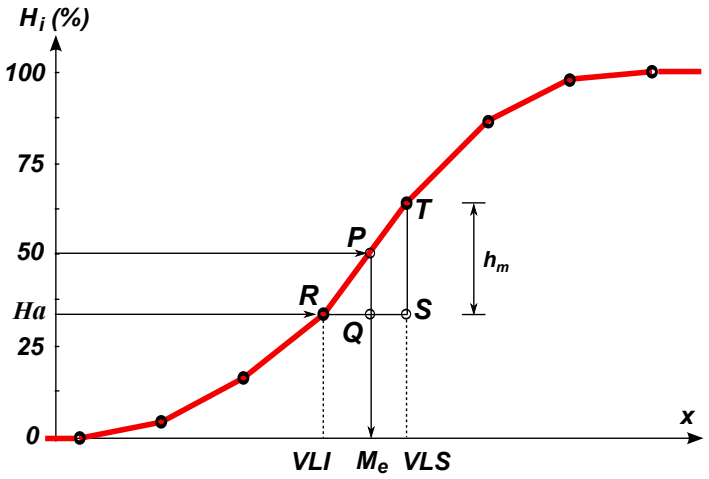


Figura 2.7: Determinación de la mediana para datos agrupados utilizando el histograma de frecuencias absolutas.

$$\frac{\overline{RQ}}{\overline{RS}} = \frac{\overline{PQ}}{\overline{TS}} \quad \Rightarrow \quad \frac{M_e - VLI}{c} = \frac{50\,\% - H_a}{h_m}$$

con lo cual la ecuación final resulta:

$$M_e = VLI + c\frac{50\,\% - H_a}{h_m} \tag{2.8}$$

donde H_a es la frecuencia relativa acumulada anterior y h_m es la frecuencia relativa del intervalo donde se encuentra M_e (ambas expresadas como porcentajes). Se puede observar claramente la similitud con la ecuación (2.6).

Aplicación

Siguiendo con los ejemplos, seguimos el siguiente procedimiento:

- Se obtiene el valor $n/2$.
- Se determina cuál es el intervalo cuya frecuencia acumulada "menor que" es igual o supera por primera vez a $n/2$.
- A partir de esta determinación, se otorga a los elementos de la fórmula los valores que corresponden a VL, F_a, f_m y c.

Ejemplo a) $n/2 = 10 \rightarrow 3^{er}\ IC \Rightarrow\ M_e = 12{,}5 + 6\dfrac{10-5}{7} = 16{,}79$ expedientes.

Ejemplo b) $n/2 = 50 \rightarrow 3^{er}\ IC \Rightarrow\ M_e = 2 + 0{,}5\dfrac{50-23}{42} = 2{,}32$ pH.

Ejemplo c) $n/2 = 50 \rightarrow 3^{er}\ IC \Rightarrow\ M_e = 15 + 5\dfrac{50-23}{42} = 18{,}21$ Kwh.

Modo

Modo

El **modo o moda** (M_o) es el valor de la variable al cual le corresponde la máxima frecuencia absoluta.

La palabra modo es, en realidad, una transformación académica de la palabra *moda*, utilizada normalmente para indicar algo que se suele utilizar con gran frecuencia.

Ejemplos:

(a)

x_i	f_i
1	1
3	3
5	4
8	1

(b)

x_i	f_i
1	2
4	5
8	5
11	3
12	1
100	1

(c)

x_i	f_i
1	1
4	1
8	1
11	1
12	1

(a) El modo es igual a 5, ya que a ese valor le corresponde la máxima frecuencia.

(b) Existen dos modos: 4 y 8. El conjunto de datos en este caso es *bimodal*.

(c) Como todos los valores de la variable tienen igual frecuencia, no existe un valor modal. Por consiguiente, el conjunto es *amodal*.

Para poder deducir la fórmula de cálculo del Modo para un conjunto de datos agrupados se recurre a un análisis similar al realizado para la mediana. En este caso se debe tener presente que M_o se encontrará en aquel intervalo de clase que posea la máxima frecuencia absoluta. Considerando el histograma de frecuencias absolutas de la figura (2.8) se puede efectuar la estimación del valor modal siguiendo las siguientes pautas:

- En la figura (2.8) se presenta solamente el sector del histograma que corresponde a los siguientes tres intervalos de clase: el que posee la frecuencia absoluta máxima y los dos intervalos vecinos, el anterior y el posterior.
- El modo debe encontrarse en el intervalo de clase que posee la máxima frecuencia y cumple con la siguiente condición: si la frecuencia absoluta del intervalo anterior (f_{ant}) es mayor que la frecuencia absoluta del intervalo posterior (f_{post}), M_o se encontrará a la izquierda del punto medio del intervalo modal, es decir, más cerca del verdadero límite inferior. Si en cambio f_{post} es mayor que f_{ant}, el modo se encontrará a la derecha del punto medio del intervalo que lo contiene, es decir, más cerca del verdadero límite superior. Este último caso es el observado en la figura (2.8).

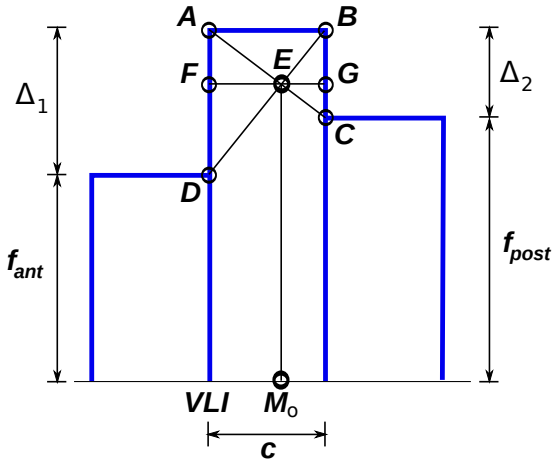


Figura 2.8: Determinación de la mediana para datos agrupados utilizando el histograma de frecuencias absolutas.

- Para determinar la posición del modo, se trazan dos segmentos, el primero une los puntos $\boldsymbol{A}$ y $\boldsymbol{C}$ y el segundo une los puntos $\boldsymbol{B}$ y $\boldsymbol{D}$, definiendo en su intersección el punto $\boldsymbol{E}$.

- La posición del punto $\boldsymbol{E}$ dependerá de las alturas de los rectángulos anterior y posterior al rectángulo central, por lo que el punto $\boldsymbol{E}$, proyectado hacia el eje de abscisas, indicará la ubicación de $\boldsymbol{M_o}$.

- Por el punto $\boldsymbol{E}$ se traza una línea paralela al eje de las abscisas, con lo cual se determinan los puntos $\boldsymbol{F}$ y $\boldsymbol{G}$ y los triángulos $\boldsymbol{AED}$ y $\boldsymbol{BEC}$, semejantes entre sí por ser opuestos por el vértice $\boldsymbol{E}$ entre las paralelas $\overline{\boldsymbol{AD}}$ y $\overline{\boldsymbol{CB}}$, por lo que se cumple la siguiente relación:

$$\frac{\overline{EF}}{\overline{AD}} = \frac{\overline{EG}}{\overline{BC}} \tag{2.9}$$

- Vemos que el segmento $\overline{\boldsymbol{AD}}$ es igual a la frecuencia modal menos la frecuencia anterior y que el segmento $\overline{\boldsymbol{BC}}$ es la frecuencia modal menos la frecuencia posterior:

$$\overline{AD} = f_{modal} - f_{ant} = \Delta_1$$
$$\overline{BC} = f_{modal} - f_{post} = \Delta_2$$

- Se comprueba además que el segmento $\overline{\boldsymbol{EF}}$ es igual al modo menos el verdadero límite inferior de su intervalo de clase y que el segmento $\overline{\boldsymbol{EG}}$ es el verdadero límite superior menos el modo:

$$\overline{EF} = M_o - \text{VLI}$$
$$\overline{EG} = \text{VLS} - M_o = (\text{VLI} + c) - M_o$$

- Con estas consideraciones, se parte de la ecuación (2.9) para obtener la fórmula de M_o:

$$\frac{M_o - \text{VLI}}{\Delta_1} = \frac{(\text{VLI} + c) - M_o}{\Delta_2}$$
$$\Rightarrow \ \Delta_2\ (M_o - \text{VLI}) = \Delta_1\ (\text{VLI} + c - M_o)$$
$$\Rightarrow \ M_o(\Delta_1 + \Delta_2) = \text{VLI}\ (\Delta_1 + \Delta_2) + c\ \Delta_1$$

por lo cual,

$$M_o = \text{VLI}\ + c\ \frac{\Delta_1}{\Delta_1 + \Delta_2} \tag{2.10}$$

Aplicación:

Utilizamos la ecuación (2.10) con los ejemplos vistos:

Ejemplo a) $IC_{modal} : 3^{er}\ IC \Rightarrow\ M_o = 12{,}5 + 6\frac{3}{3+1} = 17$ expedientes.

Ejemplo b) $IC_{modal} : 3^{er}\ IC \Rightarrow\ M_o = 2 + 0{,}5\frac{24}{24+15} = 2{,}31$ pH.

Ejemplo c) $IC_{modal} : 3^{er}\ IC \Rightarrow\ M_e = 15 + 5\frac{24}{24+15} = 18{,}08$ Kwh.

Cuantiles

Cuantiles

Se denominan **cuantiles** a aquellos valores de la variable que dividen al conjunto, serie de datos o distribución, una vez ordenados de menor a mayor o viceversa, en **k** partes iguales.

Los cuantiles son medidas de posición *no* central, dado que determinan los valores de la variable que dividen en distintas partes al conjunto de datos. Estas partes a su vez, contienen cada una la misma cantidad de observaciones de la variable o frecuencias.

Los cuantiles más utilizados son: los **cuartiles** ($\boldsymbol{Q_i}$), que dividen al conjunto de datos en cuatro partes, cada una de las cuales agrupa el 25 % de los datos; los **deciles** ($\boldsymbol{D_i}$) que lo dividen en diez partes, donde cada parte agrupa el 10 % de los datos, y los **percentiles** ($\boldsymbol{P_i}$), que dividen al conjunto en cien partes, agrupando cada parte el 1 % correspondiente. Por ejemplo, dado el siguiente conjunto ordenado de datos,

$$1 - 3 - \textcircled{5} - 5 - 8 - \textcircled{8} - 9 - 9 - \textcircled{10} - 10 - 11$$

si se lo divide en dos partes iguales tomando el valor central **8**, se conforman dos grupos de cinco valores, a la derecha y a la izquierda del valor central. A su vez, el grupo de la derecha tiene como valor central el número **5** y el de la izquierda al número **10**. Los valores así determinados constituyen los cuartiles del conjunto de datos:

$$Q_1 = 5; \quad Q_2 = 8; \quad Q_3 = 10.$$

Se comprueba además que la *mediana* coincide con el cuartil segundo ($M_e = Q_2$) y que los cuartiles son tres. Puede verificarse fácilmente que los deciles son nueve y que los percentiles son noventa y nueve. Es decir que la cantidad de cuantiles es $(k-1)$, siendo k el número de partes en el cual se divide al conjunto.

A partir del hecho de que M_e y Q_2 coinciden, es posible establecer una fórmula de cálculo para los cuartiles de datos agrupados partiendo de la ecuación que permite el cálculo de M_e, ecuación (2.7). Es decir,

$$Q_2 = \text{VLI} + c\,\frac{n/2 - F_a}{f_{q_2}} \tag{2.11}$$

Análogamente, para los cuartiles primero y tercero:

$$Q_1 = \text{VLI} + c\,\frac{n/4 - F_a}{f_{q_1}} \quad \text{y} \quad Q_3 = \text{VLI} + c\,\frac{3n/4 - F_a}{f_{q_3}} \tag{2.12}$$

De igual manera, la ecuación general para el cálculo de los deciles y percentiles:

$$\begin{aligned} D_i &= \text{VLI} + c\,\frac{\frac{i}{10}n - F_a}{f_{d_i}} \\ P_i &= \text{VLI} + c\,\frac{\frac{i}{100}n - F_a}{f_{p_i}} \end{aligned} \tag{2.13}$$

Aplicación:

Para los ejemplos vistos, los cuantiles Q_1 y Q_3 son:

Ejemplo a) $Q_1 = 6{,}5 + 6\dfrac{5-1}{4} = 12{,}5$ expedientes y

$$Q_3 = 18{,}5 + 6\frac{15-12}{6} = 21{,}5 \text{ expedientes.}$$

Ejemplo b) $Q_1 = 2 + 0{,}5\dfrac{25-23}{42} = 2{,}02$ pH y

$$Q_3 = 2{,}5 + 0{,}5\frac{75-65}{27} = 2{,}69 \text{ pH.}$$

Ejemplo c) $Q_1 = 15 + 5\dfrac{25-23}{42} = 15{,}19$ Kwh y

$$Q_3 = 20 + 5\frac{75-65}{27} = 21{,}85 \text{ Kwh.}$$

Gráfico de caja

En 1977, John Wilder Tukey introdujo por primera vez un tipo de gráfico capaz de resumir información utilizando cinco medidas estadísticas: x_m, Q_1, M_e, Q_3 y x_M, conocido

como gráfico de caja (también conocido como *boxplot* o *caja de Tukey*), figura (2.9).

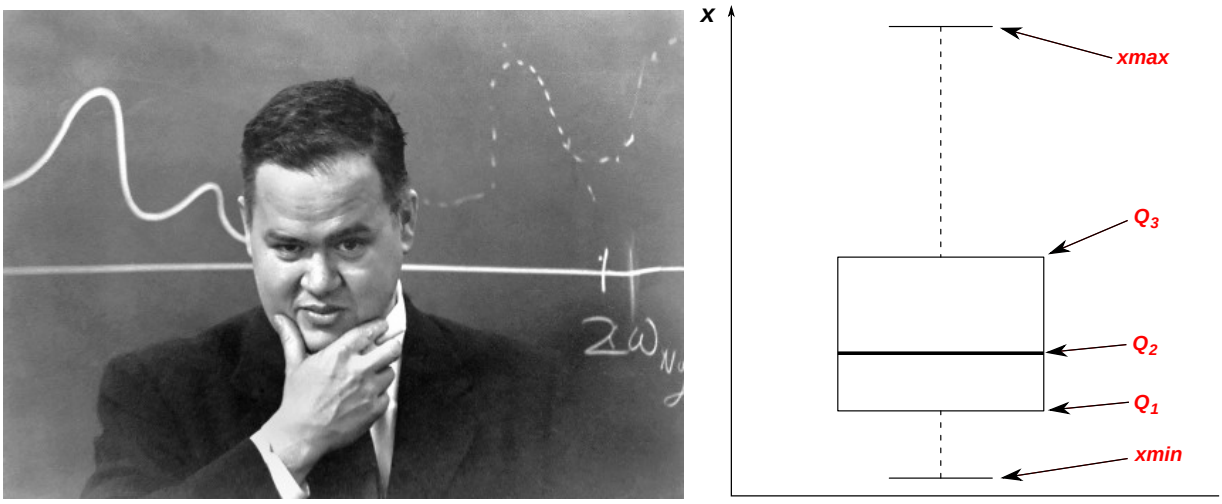


Figura 2.9: *Izq.:* John W. Tukey (1915-2000) - *Der.:* Gráfico de caja o *boxplot*.

En un gráfico de caja es posible visualizar si los datos presentan asimetría o si existen los denominados datos *anómalos* (también denominados *outliers*). El gráfico consiste en una caja cuya base coincide con el valor del primer cuartil y su lado superior con el tercer cuartil, mientras que el segmento interno representa el valor de la mediana. Desde los bordes superior e inferior se desarrollan dos segmentos, cuyos extremos representan los valores máximos y mínimos, respectivamente, del conjunto de datos. Estos segmentos son conocidos como "bigotes" (*whiskers*).

Para individualizar los potenciales datos anómalos, se conviene adoptar como longitud máxima de los bigotes de la caja una distancia de 1,5 veces la altura de la caja, es decir $1{,}5(Q_3 - Q1)$. Cualquier punto por fuera de esa distancia es considerado un dato anómalo y es representado por un punto, figura (2.10). Vemos también que esta construcción gráfica permite la comparación directa entre distintas muestras.

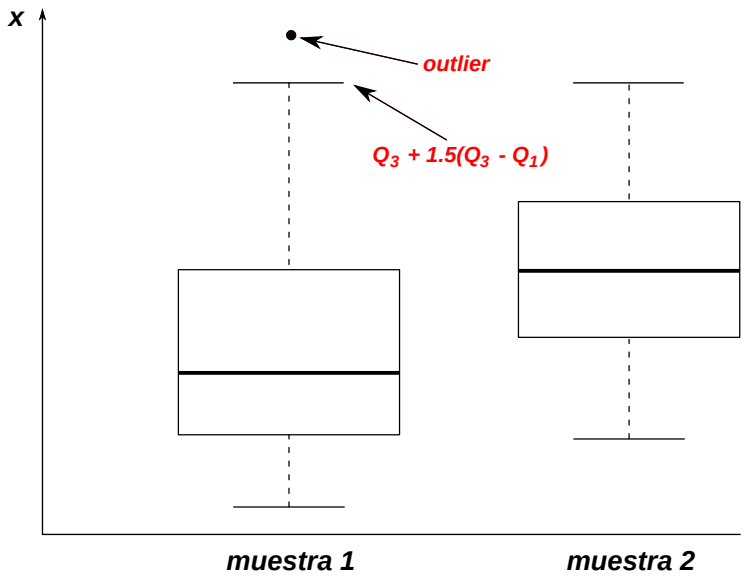


Figura 2.10: Comparación entre dos muestras utilizando boxplots. En la muestra 1 se detecta un *outlier* utilizando la convención $1{,}5(Q_3 - Q1)$.

Relaciones entre las medidas de posición - Asimetría

A lo largo de este Capítulo se definieron varias medidas de posición: la media aritmética, la mediana, el modo y los cuantiles. A continuación se verá que entre las tres primeras mantienen entre sí una relación sumamente interesente, que depende de la forma que adopta el histograma correspondiente.

En el caso de una distribución de frecuencias con histograma perfectamente simétrico, el polígono de frecuencias puede imaginarse, en un caso límite e ideal, como un curva continua que tiene la forma "de campana" que se presenta en el centro de la figura (2.11). En ese caso puede verse claramente que coinciden las tres medidas de posición mencionadas anteriormente.

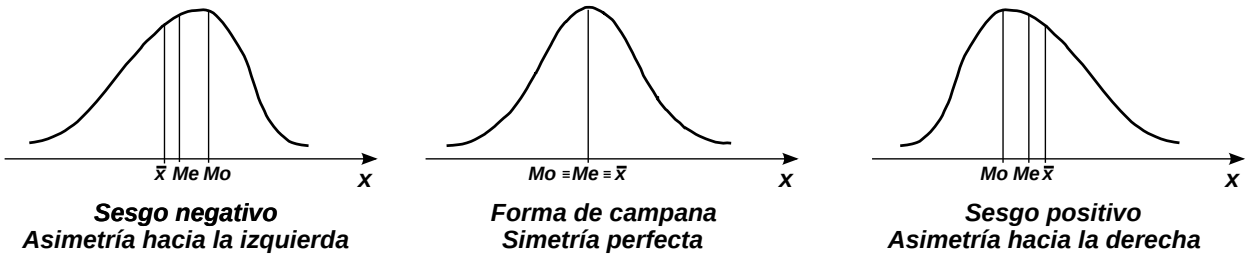


Figura 2.11: Determinación de la mediana para datos agrupados utilizando el histograma de frecuencias absolutas.

Ahora bien, si el histograma condujera a un polígono de frecuencias que diera lugar a una forma asimétrica, las tres medidas de posición ya no coincidirán en un mismo valor y sus posiciones dependerán de la forma final de la curva. En la figura (2.11) pueden verse las dos alternativas de asimetría, hacia la izquierda y hacia la derecha, verificándose lo siguiente:

- Las tres medidas de posición no coinciden.
- El modo está ubicado en el punto del eje de abscisas que corresponde a la altura máxima de la gráfica.
- La mediana está ubicada de tal modo que la superficie bajo la curva a la izquierda de ella es igual a la superficie bajo la curva a su derecha.

Con lo cual es posible concluir que:

- Si existe simetría perfecta: $\overline{x} = M_e = M_o$.
- Si existe asimetría hacia la izquierda: $\overline{x} < M_e < M_o$.
- Si existe asimetría hacia la derecha: $\overline{x} > M_e > M_o$.

En base a esta evidencia, Karl Pearson figura (2.12), propuso la siguiente fórmula empírica (no posee demostración teórica), la cual relaciona las tres medidas de posición:

$$\overline{x} - M_o = 3(\overline{x} - M_e) \tag{2.14}$$

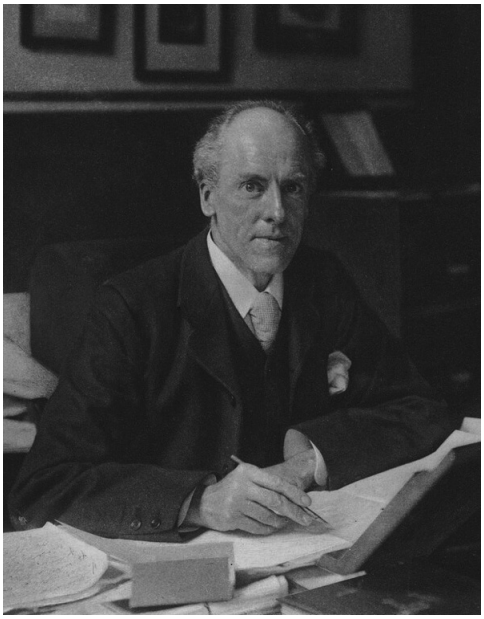


Figura 2.12: El matemático inglés Karl Pearson (1857-1936) jugó un papel predominante en el desarrollo de la estadística moderna.

Conveniencia de las medidas de posición

Las diferentes medidas de posición analizadas tienen, cada una, un conjunto de ventajas y desventajas que deben ser consideradas al determinar cuándo resulta más conveniente utilizar una u otra:

a) Si la distribución de frecuencias es simétrica o aproximadamente simétrica, resulta conveniente utilizar la media aritmética, por las distintas propiedades que posee.

b) Si la distribución de frecuencias es notoriamente asimétrica, es más apropiado utilizar la mediana o el modo, ya que ambas no se encuentran afectadas por los valores extremos del conjunto de datos.

c) Si la distribución de frecuencias tiene todos sus intervalos cerrados, no existen inconvenientes para calcular cualquiera de las tres medidas de posición. Para decidir cuál conviene se tomará en cuenta lo indicado en los puntos *a)* y *b)*.

d) Si la distribución tiene algún intervalo de clase abierto (el primero, el último o ambos), como no puede determinarse el/los punto/s medio/s tampoco será posible calcular la media aritmética, pero sí la mediana y el modo, dado que en sus fórmulas de cálculo no intervienen los puntos medios, sólo los verdaderos límites.

2.4. Medidas de dispersión

Para comprender cuál es la utilidad de las medidas de dispersión, se presenta un ejemplo sumamente sencillo. Sean dos variables aleatorias X e Y para las cuales se dispone de los siguientes datos:

$\boldsymbol{x}$	$\boldsymbol{y}$
40	20
40	30
40	50
40	60
$\overline{\boldsymbol{x}} = 40$	$\overline{\boldsymbol{y}} = 40$

Tabla 2.9: Valores de las variables aleatorias X e Y

Podemos ver que si bien las medias aritméticas de ambas variables son idénticas, provienen de conjuntos de datos completamente diferentes. Esto significa:

- que las medidas de posición por sí solas no son suficientes para determinar las características de un conjunto de datos.
- que se requiere de una medida adicional que permita calcular el alejamiento de los valores de la variable respecto de algún valor de referencia.

De esta manera surge la necesidad de calcular medidas de dispersión para ampliar adecuadamente la información referida al conjunto de datos bajo estudio.

Rango

El rango ($\boldsymbol{R}$) fue presentado anteriormente en el tema "Distribución de Frecuencias" y se definió como la diferencia entre los valores extremos del conjunto de datos ordenados, es decir, $R = x_M - x_m$. Esta medida de dispersión es muy sencilla de calcular, lo que constituye una ventaja, pero a su vez tiene dos desventajas que lo hacen desaconsejable como medida de dispersión:

- Para calcularlo no se toma como referencia ningún valor considerado central.
- No es factible su cálculo (exacto) en una distribución de frecuencias.

Ejemplo: considerando los valores de la tabla (2.9), los valores del rango serían:

$$R_x = 40 - 40 = 0 \qquad R_y = 60 - 20 = 40$$

Otro tipo de rango, denominado **rango semi-intercuartílico**, $\boldsymbol{R_q}$, utiliza dos de los cuartiles:

$$R_q = \frac{Q_3 - Q_1}{2} \tag{2.15}$$

cuyo resultado mide la distancia (o semi-distancia) entre dos medidas de posición particulares. Un mayor o menor valor de R_q indica una mayor o menor distancia entre los dos cuartiles, es decir, una mayor o menor dispersión.

Desvío medio

Para cuantificar la dispersión de un conjunto de valores, una posibilidad interesante surge al considerar los desvíos de cada una de las variables respecto de la media aritmética para construir una medida, en especial porque los desvíos constituyen una manera natural de medir el alejamiento de los valores de la variable respecto a un valor central. Sin embargo, de acuerdo a la segunda propiedad de la media aritmética, hemos visto que la suma de los desvíos se anula, es decir, $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$, lo cual en principio imposibilita su utilización como cuantificador de la dispersión, a menos que se encuentre una forma apropiada de evitar esa nulidad.

El desvío medio, $\boldsymbol{DM}$, aparece como una de esas alternativas y se expresa como:

$$DM = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}| \tag{2.16}$$

donde se observa que al utilizar la suma de los valores absolutos de los desvíos se evita que el resultado final se anule, obteniendo una nueva medida que tiene como ventaja el hecho de que toma como referencia un valor central.

Ejemplo: considerando los valores de la tabla (2.9), los valores del desvío medio serían:

$$DM_x = 0$$
$$DM_y = \frac{|20-40| + |30-40| + |50-40| + |60-40|}{4} = \frac{60}{4} = 15$$

Variancia

Definición

Si $x_1, x_2, \ldots, x_n$ constituye una muestra de tamaño n, la **variancia** es

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2.17}$$

Esta medida de dispersión utiliza, al igual que el desvío medio, al desvío como base para sus cálculos, con la diferencia de que opera con sus valores al cuadrado y no con sus valores absolutos. Su fórmula ponderada es:

$$S_x^2 = \frac{1}{\sum_{i=1}^{k} f_i}\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i \qquad \text{con } k \leq n \tag{2.18}$$

que se convierte en la fórmula simple si todas las f_i son iguales a la unidad. Puede verse claramente que, si bien los desvíos continúan siendo los elementos básicos para el cálculo

de la medida de dispersión, se ha utilizado el artificio de elevarlos al cuadrado, por lo que todos se convierten en valores positivos y su suma no es nula.

Ejemplo: el valor de la variancia de los valores de la tabla (2.9) es:

$$S_x^2 = 0,$$
$$S_y^2 = \frac{(20-40)^2 + (30-40)^2 + (50-40)^2 + (60-40)^2}{4} = \frac{1000}{4} = 250$$

con lo cual se pueden realizar las siguientes observaciones:

- en el caso de la variable x, todas las medidas de dispersión calculadas dieron como resultado un valor cero.
- en el caso de la variable y, mediante la variancia se obtiene un resultado excesivo frente a las otras medidas de dispersión ya vistas, debido a que modifica la escala de las variables al elevar los desvíos al cuadrado.

Desvío o desviación estándar

Para obtener una medida de dispersión en la misma escala que la variable analizada se utiliza el desvío o desviación estándar (o típica) ($\boldsymbol{S_x}$):

$$S_x = \sqrt{S_x^2} = \sqrt{\frac{1}{\sum_{i=1}^{k} f_i} \sum_{i=1}^{k} (x_i - \overline{x})^2 f_i} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{2.19}$$

El desvío estándar mide como se alejan, **en promedio**, los valores de la variable respecto de su propia media aritmética.

Ejemplo: nuevamente, para el caso analizado, los desvíos estándar son:

$$S_x = 0,$$
$$S_y = \sqrt{250} = 15{,}81$$

Nota: como ya se hizo referencia, es importante destacar que la unidad de medida de la variancia es la misma unidad que corresponde a la variable bajo estudio elevada al cuadrado, mientras que la unidad de medida del desvío estándar es la misma que la de la variable (y de la media aritmética).

Variancia - Desarrollo teórico y conceptual.

Variancia poblacional y variancia muestral - Simbología.

Un símbolo general para indicar a la variancia es la letra V. De modo que, por ejemplo, el símbolo $V(x)$ indica la variancia de la variable x. De igual manera, los símbolos para las variancias poblacional y muestral varían entre sí del mismo modo en que varían los símbolos para las medias. En este caso, los símbolos S_x^2 y S_x se reservan para la variancia y el desvío

estándar muestrales, mientras que para los símbolos poblacionales se utiliza la letra griega σ, siendo σ_x^2 y σ_x la variancia y el desvío estándar poblacionales, respectivamente. Cuando la población es finita y está formada por N valores, la variancia poblacional puede definirse como

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2. \tag{2.20}$$

Así como la media muestral puede emplearse para hacer inferencias sobre la media poblacional, la variancia muestral puede utilizarse para realizar inferencias sobre la variancia poblacional.

Fórmula de trabajo de la variancia.

La fórmula de trabajo se obtiene mediante una serie de operaciones algebraicas y permite encontrar un procedimiento diferente y en algunos casos más conveniente, para calcular la variancia. Se parte de la fórmula ponderada y se desarrolla el cuadrado del binomio que se encuentra entre paréntesis,

$$\begin{aligned} S_x^2 &= \frac{1}{\sum_{i=1}^{k} f_i}\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i = \frac{1}{n}\sum(x_i^2 - 2x_i\overline{x} - \overline{x}^2)f_i \\ &= \frac{1}{n}\sum(x_i^2 f_i - 2x_i\overline{x}f_i + \overline{x}^2 f_i) = \frac{1}{n}\left(\sum x_i^2 f_i - 2\overline{x}\sum x_i f_i + \overline{x}^2\sum f_i\right) \\ &= \frac{\sum x_i^2 f_i}{n} - 2\overline{x}\frac{\sum x_i f_i}{n} + \overline{x}^2\frac{n}{n} = \frac{\sum x_i^2 f_i}{n} - 2\overline{x}^2 + \overline{x}^2 = \frac{\sum x_i^2 f_i}{n} - \overline{x}^2 \end{aligned}$$

Luego, las fórmulas de trabajo son:

$$\begin{aligned} S_x^2 &= \frac{\sum_{i=1}^{k} x_i^2 f_i}{\sum_{i=1}^{k} f_i} - \overline{x}^2 \qquad \text{para la forma ponderada,} \\ S_x^2 &= \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2 \qquad \text{para la forma simple.} \end{aligned} \tag{2.21}$$

Propiedades de la variancia.

1) **La variancia de un conjunto definido de datos es un valor constante mayor o igual a cero.**

   Esta propiedad tiene similar connotación y objetivo que la primer propiedad de la media aritmética. Además, es un valor positivo (mayor o igual a cero) dado que la fórmula de la variancia está compuesta por la sumatoria de valores elevados al cuadrado.

2) **La variancia de una constante es igual a cero.**

Sea una constante a cuya media aritmética es $\overline{a} = a$. Luego, la variancia de a será

$$S_a^2 = \frac{1}{n}\sum_{i=1}^{n}(a - \overline{a})^2 = 0$$

Puede verse que esta propiedad se verificó empíricamente cuando se calcularon las medidas de dispersión (la variancia entre ellas) de la variable X de la tabla (2.9), y que por tener todos sus valores iguales a 40, constituye una constante.

3) **La variancia es una medida mínima si se la compara con cualquier otra similar que se calcule tomando como referencia alguna medida de posición diferente de la media aritmética.**

Esta propiedad remite a la tercer propiedad de la media aritmética, la cual expresa que la sumatoria de los desvíos al cuadrado entre los valores de la variable y la media aritmética es un mínimo, lo cual equivale a decir que si los desvíos se calcularan respecto de cualquier otro valor que no fuera la media aritmética, la sumatoria de ellos al cuadrado daría un resultado mayor.

4) **Variancia de variables transformadas algebraicamente.**

a) **Si a todos los valores de una variable se le suma o resta un valor constante y arbitrario A, se obtiene una nueva variable, cuya variancia será igual a la de la variable original.**

Sea x_i una variable con media $\overline{x}$ y variancia S_x^2 y sea A un valor constante arbitrario. Se construye la variable $w_i = x_i \pm A$, y recordando que $\overline{w} = \overline{x} \pm A$:

$$\begin{aligned} S_w^2 &= \frac{1}{n}\sum_{i=1}^{n}(w_i - \overline{w})^2 = \frac{1}{n}\sum[x_i \pm A - (\overline{x} \pm A)]^2 \\ &= \frac{1}{n}\sum(x_i \pm A - \overline{x} \mp A)^2 = \frac{1}{n}\sum(x_i - \overline{x})^2 = S_x^2 \end{aligned}$$

b) **Si a todos los valores de una variable se les multiplica (o divide) por un valor constante y arbitrario c, se obtiene una nueva variable cuya variancia será igual a la de la variable original multiplicada o dividida por c^2.**

Sea x_i una variable con media $\overline{x}$ y variancia S_x^2 y sea c un valor constante y arbitrario. Se construye la variable $w_i = c\ x_i$ (se deducirá la propiedad para el caso del producto pero resulta sencillo ver que la demostración es equivalente para el caso del cociente) y recordando que $\overline{w} = c\overline{x}$:

$$S_w^2 = \frac{1}{n}\sum_{i=1}^{n}(w_i - \overline{w})^2 = \frac{1}{n}\sum(cx_i - c\overline{x})^2 = \frac{1}{n}\sum c^2(x_i - \overline{x})^2 = c^2\frac{1}{n}\sum(x_i - \overline{x})^2 = c^2 S_x^2$$

c) Para una transformación algebraica que combina los casos a) y b), $w_i = \frac{x_i - A}{c}$, recordando que $\overline{w} = \frac{\overline{x} - A}{c}$, se obtiene:

$$S_w^2 = \frac{1}{n}\sum_{i=1}^{n}(w_i - \overline{w})^2 = \frac{1}{n}\sum\left[\left(\frac{x_i - A}{c}\right) - \left(\frac{\overline{x} - A}{c}\right)\right]^2$$
$$= \frac{1}{c^2}\frac{1}{n}\sum(x_i - A - \overline{x} + A)^2 = \frac{1}{c^2}\frac{1}{n}\sum(x_i - \overline{x})^2 = \frac{1}{c^2}S_x^2$$

5) **Variancia de la suma o diferencia de dos variables.**

a) **Caso de la suma.** Sean dos variables x_i e y_i, con medias $\overline{x}$ e $\overline{y}$ y variancias S_x^2 y S_y^2, conocidas. Se construye una nueva variable $w_i = x_i + y_i$, recordando que $\overline{w} = \overline{x} + \overline{y}$:

$$S_w^2 = \frac{1}{n}\sum_{i=1}^{n}(w_i - \overline{w})^2 = \frac{1}{n}\sum\left[(x_i + y_i) - (\overline{x} + \overline{y})\right]^2$$
$$= \frac{1}{n}\sum\left[(x_i - \overline{x}) + (y_i - \overline{y})\right]^2 = \frac{1}{n}\sum\left[(x_i - \overline{x})^2 + (y_i - \overline{y})^2 + 2(x_i - \overline{x})(y_i - \overline{y})\right]$$
$$= \frac{1}{n}\Big[\sum(x_i - \overline{x})^2 + \sum(y_i - \overline{y})^2 + 2\sum(x_i - \overline{x})(y_i - \overline{y})\Big]$$
$$= \frac{1}{n}\sum(x_i - \overline{x})^2 + \frac{1}{n}\sum(y_i - \overline{y})^2 + 2\frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y}) = S_x^2 + S_y^2 + 2S_{xy}$$
$$= V(x) + V(y) + 2\mathrm{Cov}(xy)$$

Como resultado de esta última demostración se verifica que **la variancia de la suma de dos variables es igual a la suma de sus variancias más dos veces $\mathbf{Cov(x, y)}$**, símbolo de la covariancia entre x e y (también simbolizado con S_{xy}). Luego,

$$\mathrm{Cov}(xy) = S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{2.22}$$

La covariancia mide la relación lineal promedio existente entre dos variables x_i e y_i, mediante la suma de los productos de los desvíos de ambas variables entre sí y resulta una consecuencia de la demostración de la variancia de la suma de dos variables. Más adelante se hará un análisis más detallado de este término.

b) **Caso de la diferencia.** Sean dos variables x_i e y_i, con medias $\overline{x}$ e $\overline{y}$ y variancias S_x^2 y S_y^2, conocidas. Este caso es sencillo de demostrar si se parte de la demostración anterior, es decir, construyendo una nueva variable $w_i = x_i - y_i$ y recordando que $\overline{w} = \overline{x} - \overline{y}$, con lo cual se obtiene:

$$S_d^2 = V(x) + V(y) - 2\mathrm{Cov}(xy)$$

con lo cual se verifica que la variancia de la diferencia de dos variables es igual a la suma de sus variancias menos dos veces la covariancia entre ambas variables.

Fórmula de trabajo de la covariancia.

Se ha indicado anteriormente que la covariancia aparece cuando se deduce la expresión de la variancia de la suma o de la diferencia de dos variables. Así como se realizó para la variancia, es posible obtener una fórmula de trabajo,

$$\begin{aligned}\text{Cov}(xy) &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \frac{1}{n}\sum(x_iy_i-\overline{x}y_i-x_i\overline{y}+\overline{x}\ \overline{y}) \\ &= \frac{1}{n}\sum x_iy_i-\overline{x}\frac{1}{n}\sum y_i-\overline{y}\frac{1}{n}\sum x_i+n\frac{1}{n}\overline{x}\ \overline{y} \\ &= \frac{1}{n}\sum x_iy_i-\overline{x}\ \overline{y}-\overline{y}\ \overline{x}+\overline{x}\ \overline{y} = \frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{x}\ \overline{y}\end{aligned} \tag{2.23}$$

Cálculo de la variancia en distribuciones de frecuencia

Sobre la base de los ejercicios anteriores, cuyas medias aritméticas fueron calculadas anteriormente, se presentan a continuación los cálculos de las variancias utilizando la fórmula de trabajo. Como en el caso del cálculo de la media aritmética, los puntos medios representan los valores de la variable, con los cuales se construye la columna $x_i^2 f_i$, cuya suma permite obtener uno de los datos necesarios para el cálculo de la variancia.

Ejemplo a)

N° de expedientes	$\boldsymbol{f_i}$	$\boldsymbol{x_i}$	$\boldsymbol{x_i f_i}$	$\boldsymbol{x_i^2 f_i}$
$1-6$	1	3,5	3.5	12.25
$7-12$	4	9,5	38	361
$13-18$	7	15,5	108.5	1681.75
$19-24$	6	21,5	129	2773.5
$25-30$	2	27,5	55	1512.5
	$\boldsymbol{\sum f_i = 20}$		$\boldsymbol{\sum x_i f_i = 334{,}0}$	$\boldsymbol{\sum x_i^2 f_i = 6341}$

$$\begin{aligned}\overline{x} &= 16{,}7 \text{ expedientes} \\ S_x^2 &= \frac{6341}{20} - (16{,}7)^2 = 38{,}16 \text{ (expedientes)}^2 \\ S_x &= \sqrt{38{,}16} = 6{,}18 \text{ expedientes}\end{aligned}$$

Ejemplo b)

Acidez (pH)	f_i	x_i	$x_i f_i$	$x_i^2 f_i$
1,00 – 1,49	5	1,25	6,25	7,8125
1,50 – 1,99	18	1,75	31,50	55,125
2,00 – 2,49	42	2,25	94,50	212,625
2,50 – 2,99	27	2,75	74,25	204,188
3,00 – 3,49	8	3,25	26,00	84,500
	$\sum f_i = 100$		$\sum x_i f_i = 232{,}50$	$\sum x_i^2 f_i = 564{,}25$

$$\overline{x} = 2{,}325 \text{ pH}$$
$$S_x^2 = \frac{564{,}25}{100} - (2{,}325)^2 = 0{,}111 \text{ (pH)}^2$$
$$S_x = \sqrt{0{,}111} = 0{,}333 \text{ pH}$$

Ejemplo c)

Consumo (Kwh)	f_i	x_i	$x_i f_i$	$x_i^2 f_i$
5,00 – 9,99	5	7,50	38,50	281,25
10,00 – 14,99	18	12,50	225,00	2812,50
15,00 – 19,99	42	17,50	735,00	12862,50
20,00 – 24,99	27	22,50	607,50	13068,00
25,00 – 29,99	8	27,50	220,00	6050,00
	$\sum f_i = 100$		$\sum x_i f_i = 1826{,}00$	$\sum x_i^2 f_i = 35074{,}25$

$$\overline{x} = 18{,}26 \text{ Kwh}$$
$$S_x^2 = \frac{35074{,}25}{100} - (18{,}26)^2 = 17{,}315 \text{ (Kwh)}^2$$
$$S_x = \sqrt{17{,}315} = 4{,}161 \text{ Khw}$$

Dispersión Relativa: Coeficiente de Variación

Como se mencionó anteriormente, el desvío estándar es una medida que describe de qué modo se alejan, en promedio, los valores de una variable respecto de una medida de posición convencional, como por ejemplo la media aritmética. De esta manera, se utiliza para medir el alejamiento interno promedio de los valores de una variable respecto de una medida de posición, es decir, es una medida de dispersión **absoluta**.

Ahora bien, si se desea comparar las dispersiones de dos (o más) distribuciones o conjuntos de datos, la variancia y el desvío estándar no resultan ser medidas apropiadas, por su carácter absoluto. En consecuencia, para efectuar comparaciones respecto de la dispersión

de varias distribuciones de frecuencias es necesario contar con una medida de dispersión relativa, denominada "coeficiente de variación" (CV):

$$CV(\,\%) = \frac{S_x}{\overline{x}}\ 100 \tag{2.24}$$

medida que resulta adimensional y que con frecuencia es expresada como porcentaje. Nótese que para obtener un CV siempre positivo, debemos expresar $\overline{x}$ en valor absoluto y además deberá ser $\overline{x} \neq 0$. Es más, $\overline{x}$ no deberá ser un número muy pequeño cercano a cero dado que distorsionaría el valor de CV.

Ejemplo: Sean dos variables x e y, con las siguientes medias y variancias:

$$\begin{aligned} \overline{x} &= 50 \qquad & \overline{y} &= 200 \\ S_x^2 &= 100 \qquad & S_y^2 &= 200 \end{aligned}$$

con estos datos se calculan los coeficientes de variación:

$$\begin{aligned} CV_x &= \frac{S_x}{\overline{x}} = \frac{10}{50} = 20\,\% \\ CV_y &= \frac{S_y}{\overline{y}} = \frac{14{,}145}{200} = 7\,\% \end{aligned}$$

Los resultados obtenidos permiten señalar que si bien la variable x tiene menor dispersión absoluta que la variable y, posee mayor dispersión relativa.

Variable Estandarizada

La variable estandarizada es una variable que se obtiene a partir de una transformación algebraica particular dada por $z_i = \frac{x_i - A}{c}$, donde $A = \overline{x}$ y $c = S_x$, por lo cual

$$z_i = \frac{x_i - \overline{x}}{S_x}$$

Esta transformación tiene la particularidad de que sin importar cuánto valen los valores de la variable x_i, como así también su media aritmética y desvío estándar, la media y la variancia de z_i serán siempre iguales a cero y uno, respectivamente. Es decir,

$$\overline{z} = \frac{\sum_{i=1}^{n} z_i}{n} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \overline{x}}{S_x} = \frac{1}{S_x\ n}\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

según lo visto en las propiedades de la media aritmética, y la variancia:

$$S_z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \overline{z})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{S_x} - 0\right)^2 = \frac{1}{S_x^2\ n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{S_x^2}{S_x^2} = 1$$

Asimetría

Asimetría

La asimetría es el grado de desviación hacia la derecha o hacia la izquierda que posee una distribución con respecto a una medida de posición determinada.

De acuerdo a lo visto anteriormente en las relaciones entre las medidas de posición, particularmente en la ecuación (2.14) propuesta por Pearson, si la media aritmética, la mediana y el modo coinciden en un mismo valor de la variable, existe una situación de simetría perfecta, mientras que si hay diferencias entre la media aritmética y las otras dos, se pueden presentar dos situaciones diferentes de asimetría. Una cuando la media aritmética es menor, tanto de la mediana como del modo, denominada *asimetría hacia la izquierda*, y la otra cuando la media aritmética es mayor que la mediana y el modo, definida como *asimetría hacia la derecha*.

Partiendo de la relación (2.14) es posible construir una forma de evaluar el grado de asimetría de una distribución por medio de una de las siguientes medidas:

$$\begin{aligned} As_1 &= \frac{\overline{x} - M_o}{S_x} \\ As_2 &= \frac{3(\overline{x} - M_e)}{S_x} \end{aligned} \tag{2.25}$$

En resumen:

- Si en una distribución de frecuencias las medidas de posición coinciden, los valores correspondientes a As_1 y As_2 son iguales a cero, por lo que el grado de asimetría es nulo y en consecuencia, la distribución es perfectamente simétrica.
- En una distribución, si la media aritmética fuera menor que la mediana o el modo, se obtendría una simetría hacia la izquierda. En ese caso el resultado particular de las medidas As_1 y As_2 sería negativo y por consiguiente se dice que la asimetría es hacia la izquierda o negativa.
- Si en cambio, la media aritmética fuera mayor que la mediana o el modo, el resultado de las medidas As_1 y As_2 sería positivo y por consiguiente se dice que la distribución posee una asimetría hacia la derecha o positiva.
- Anteriormente se había discutido sobre la conveniencia en el uso de una medida de posición sobre las otras. Para el caso de distribuciones asimétricas es posible obtener cierta orientación mediante las siguientes sugerencias:

$$\begin{aligned} &\text{Si} \quad -0{,}20 \leq A_s \leq +0{,}20 \Rightarrow \text{conviene utilizar la media aritmética.} \\ &\text{Si} \quad |A_s| > 0{,}20 \quad \Rightarrow \text{no conviene utilizar la media aritmética.} \end{aligned}$$

En el caso de no poder obtener los valores de la media aritmética o de la variancia (como por ejemplo sucede cuando se tienen distribuciones de frecuencia con intervalos abiertos), las medidas de asimetría dadas por las ecuaciones (2.25) no pueden utilizarse. En esos

caso puede utilizarse una medida de asimetría basada en los cuartiles, dado que éstos no precisan que todos los intervalos de clase sean cerrados:

$$As_q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \tag{2.26}$$

Capítulo 3
Ajustamiento lineal y Correlación

3.1. Análisis de regresión

En muchos problemas existen dos o más variables que se encuentran relacionadas de alguna manera y en donde resulta necesario explorar la naturaleza de esa relación. El *análisis de regresión* es una técnica estadística utilizada para el modelado y análisis de la relación existente entre dos o más variables. Esta herramienta es aplicada en numerosos campos de estudio:

- supongamos que en un proceso químico el rendimiento del producto está relacionado con la temperatura de operación del proceso. El análisis de regresión puede utilizarse para predecir el rendimiento a un valor determinado de temperatura. Este modelo también puede ser utilizado para optimización del proceso, como puede ser *hallar la temperatura que maximiza el rendimiento* o bien para su control.
- entre dos procesadores P_0 y P_1, el tiempo T_c que tarda en enviarse un mensaje entre ellos es función de la longitud del mensaje: $T_c = T_c(n)$, donde n es el número de bytes en el mensaje. Realizando un análisis de regresión entre el tiempo de comunicación y diferentes tamaños de mensajes es posible determinar la *latencia* y el *ancho de banda* de la red.
- es conocido que la rigidez a la flexión de un tirante de madera está relacionada con la densidad de la misma. Mediante un análisis de regresión es posible encontrar la densidad óptima para un determinado valor de rigidez requerida.

En todos estos ejemplos tenemos variables dependientes o *respuestas* (rendimiento, tiempo de comunicación, rigidez) y variables independientes o *predictoras* (temperatura, tamaño del mensaje, densidad).

El término *regresión* fue introducido en 1889 por Francis Galton, Fig. (3.1), en su trabajo *"Regression toward mediocrity in hereditary stature"*, donde comparó las alturas de una muestra de padres con sus respectivos hijos. Galton observó que si bien los hijos de padres altos tendían a ser altos, llegaban a no ser tan altos como aquellos. Igualmente, hijos de padres bajos tendían a ser bajos, pero no de estatura tan baja como sus padres (el artículo original se encuentra disponible en http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf). A esta relación denominó *"regression to mediocrity"* (regresión a la media).

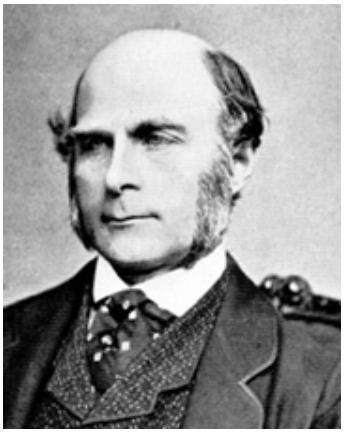


Figura 3.1: Francis Galton (1822 - 1911).

Sabemos que cuando dos variables X_i e Y_i están relacionadas por una expresión matemática de cualquier tipo (por ejemplo, $Y_i = a + bX_i$ o $Y_i = ae^{bX_i}$) se dice que entre ellas existe una *dependencia funcional*. Este tipo de dependencia es tal que a determinados valores de la variable X_i le corresponden valores *definidos* de la variable Y_i. Entiéndase aquí como "definidos" a que esta relación proporciona siempre el mismo valor de la respuesta ante una misma "predicción". Por el contrario, se dice que entre dos variables X_i e Y_i existe una *dependencia estadística* cuando se presupone que entre ambas existe algún tipo de relación mediante la cual, para determinados valores de la variable X_i existe una correspondencia no determinística con los valores de la variable Y_i.

Concepto de ajustamiento

Supongamos la existencia de dos variables X_i e Y_i donde se sabe o simplemente se supone que entre ellas existe algún tipo de relación que insinúe una dependencia estadística. Además de los ya mencionados anteriormente, podemos citar los siguientes casos:

- temperatura de la superficie asfáltica de una ruta y la deformación del pavimento.
- consumo de combustible de un vehículo y la distancia recorrida.
- velocidad del viento y voltaje producido por un generador eólico.

Para cada una de las variables bajo análisis es posible obtener n valores empíricos, es decir, n datos provenientes de observaciones o mediciones, los cuales se ordenan en una tabla con el formato de la figura (3.2) (*izq*). La tabla contiene n pares de datos empirícos de la forma (X_i, Y_i). Con ese conjunto de valores se construye la figura (3.2) (*der*) denominada **diagrama de dispersión**, la cual muestra la disposición de los n puntos en el plano.

X_i	Y_i
X_1	Y_1
X_2	Y_2
X_3	Y_3
...	...
...	...
X_n	Y_n

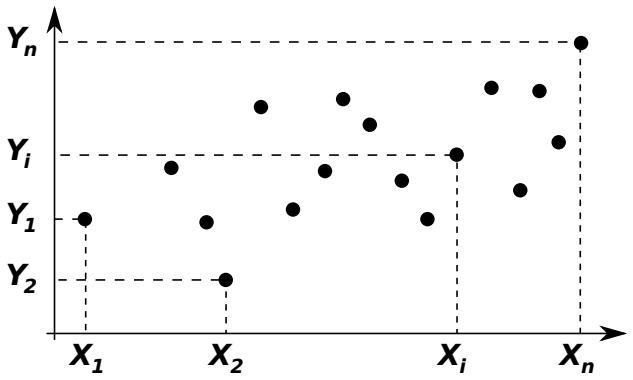


Figura 3.2: Valores empíricos para el análisis de regresión (*izq*) y diagrama de dispersión (*der*).

Definición

La teoría del ajustamiento lineal trata sobre los procedimientos destinados a ajustar linealmente los puntos del diagrama de dispersión, lo cual significa encontrar la ecuación de la función de primer grado (línea recta) que mejor explique la dependencia estadística existente, es decir, que mejor explique el comportamiento de los n puntos del diagrama.

Como observación, debe considerarse que este procedimiento difiere de los métodos de interpolación (hallar la función de grado $(n-1)$ que pase exactamente por todos esos puntos).

Por otro lado, además de describir linealmente la relación existente entre dos variables, otro de los objetivos del ajustamiento es la **estimación** o el **pronóstico**. Es decir que una vez hallada la expresión de la función matemática de primer grado, ella puede ser utilizada para estimar valores de la variable dependiente Y_i para valores seleccionados de la variable independiente X_i.

Importancia del diagrama de dispersión.

La construcción del diagrama de dispersión es imprescindible y debe ser la primera acción que realice el investigador cuando tiene los datos empíricos en su poder. Una vez construido, es sumamente conveniente observar la disposición de los puntos contenidos en él, lo que permite decidir si un ajuste lineal es procedente o si corresponde un ajuste de otro tipo, aunque debe aclararse que cualquiera sea la disposición de los puntos, todo diagrama de dispersión admite un ajustamiento del tipo lineal.

Si bien una disposición de puntos no lineal no estaría bien representada por una función de primer grado, esa función, se reitera, puede calcularse perfectamente sin inconvenientes. En todo caso, la decisión de que un ajuste sea lineal o no depende del investigador del problema, por lo que se insiste en que lo imprescindible es construir, en primer lugar, el diagrama de dispersión. Algunos casos de disposición no lineal se observan en la figura (3.3).

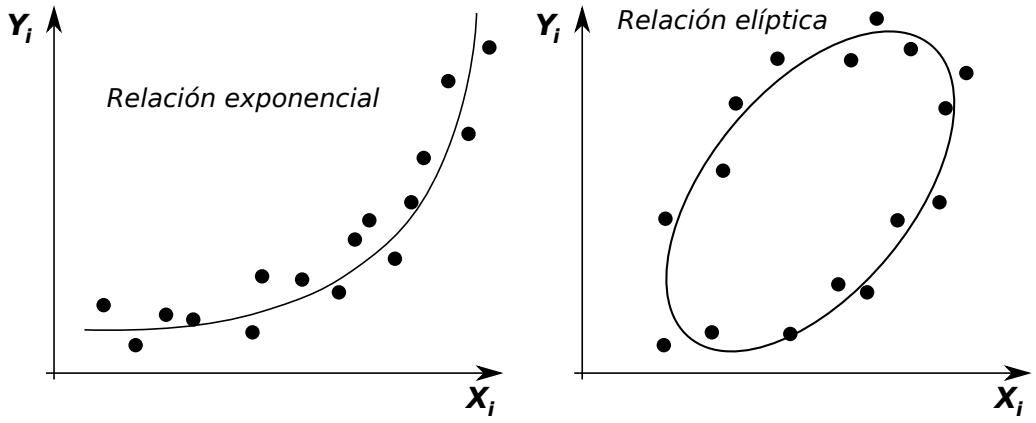


Figura 3.3: Distintas relaciones entre los valores empíricos.

Observando ambos diagramas queda perfectamente claro que los puntos no siguen una disposición lineal por lo que no sería apropiado un ajuste del tipo lineal. Existen otros tipos de análisis que permiten ajustamientos no lineales, los cuales no serán abordados en este apunte.

Método de los mínimos cuadrados

El método de los mínimos cuadrados es un procedimiento propuesto por el matemático alemán Gauss (1809)[1], quien sugirió un criterio objetivo para determinar cuál es la mejor recta de ajustamiento. De acuerdo a este criterio, esta recta es aquella que minimiza la sumatoria de los cuadrados de los desvíos existentes entre los puntos empíricos del diagrama de dispersión y la propia recta de ajustamiento.

Figura 3.4: Carl Friedrich Gauss (1777 - 1855).

La idea consiste en encontrar una función lineal del tipo $\hat{Y}_i = a_1 + b_1 X_i$ que cumpla con las condiciones sugeridas por Gauss. Para ello se deben analizar los desvíos o *residuos* del modelo de regresión. Un desvío e_i es la diferencia entre un punto empírico de ordenada Y_i y un punto teórico de ordenada $\hat{Y}_i$, es decir, $e_i = Y_i - \hat{Y}_i$. En la figura (3.5) se indica el desvío de un valor empírico cualquiera de coordenadas (X_i, Y_i).

[1]Existe una controversia acerca de la autoría del método, dado que el matemático francés Adrien-Marie Legendre (1752-1833) lo publicó en 1805.

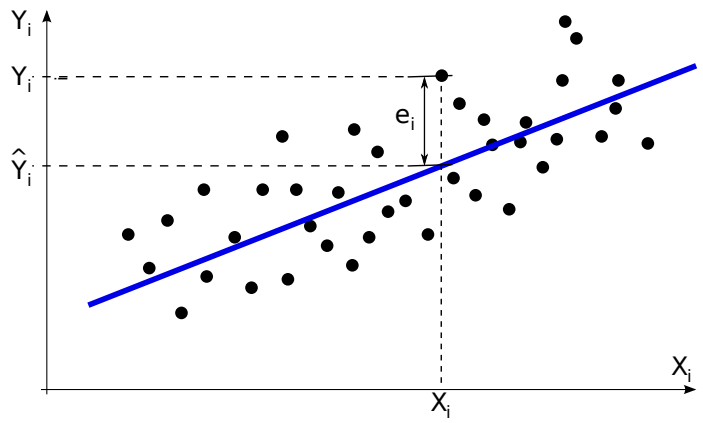


Figura 3.5: Desvío de un punto empírico respecto de la recta de ajuste lineal.

En la figura (3.5) se puede observar claramente como cualquier desvío e_i puede ser positivo (el punto empírico está por encima de la recta de ajuste), negativo (el punto empírico está por debajo de la recta de ajuste) o nulo (el punto empírico coincide con la recta de ajuste). Considerando ahora todos los posibles desvíos en el diagrama de dispersión, si los elevamos al cuadrado y los sumamos, se obtiene la siguiente expresión, que llamaremos RSS:

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a_1 - b_1\ X_i)^2.$$

Como ya se indicó, el postulado de Gauss expresa que la mejor recta de ajuste es aquella que minimiza estos desvíos al cuadrado. Si bien en un plano existen infinitas rectas, cada una con un par de parámetros a_1 y b_1, de todas ellas sólo una cumple con la condición impuesta por Gauss. Para hallar los coeficientes a_1 y b_1 que identifican a dicha recta se minimiza la función RSS:

$$\min[\text{RSS}] = \min\left[\sum_{i=1}^{n} (Y_i - a_1 - b_1\ X_i)^2\right],$$

por lo que en primer lugar se obtiene la derivada primera de RSS respecto de los coeficientes a_1 y b_1 e igualando a cero:

$$\frac{\partial \text{RSS}}{\partial a_1} = 2\sum_{i=1}^{n} (Y_i - a_1 - b_1\ X_i)(-1) = -2\sum_{i=1}^{n} (Y_i - a_1 - b_1\ X_i) = 0$$
$$\frac{\partial \text{RSS}}{\partial b_1} = 2\sum_{i=1}^{n} (Y_i - a_1 - b_1\ X_i)(-X_i) = -2\sum_{i=1}^{n} (Y_i X_i - a_1 X_i - b_1\ X_i^2) = 0.$$

Luego, como el factor $-2 \neq 0$ en ambas expresiones,

$$\frac{\partial \text{RSS}}{\partial a_1} = \sum_{i=1}^{n} (Y_i - a_1 - b_1\ X_i) = 0 \Rightarrow \sum Y_i - n a_1 - b_1 \sum X_i = 0 \tag{3.1}$$
$$\frac{\partial \text{RSS}}{\partial b_1} = \sum_{i=1}^{n} (Y_i X_i - a_1 X_i - b_1\ X_i^2) = 0 \Rightarrow \sum Y_i X_i - a_1 \sum X_i - b_1 \sum X_i^2 = 0 \tag{3.2}$$

y despejando $\sum Y_i$ de la ecuación (3.1) se obtiene:

$$\sum_{i=1}^{n} Y_i = na_1 + b_1 \sum_{i=1}^{n} X_i \tag{3.3}$$

expresión denominada **primera ecuación normal de Gauss**. Análogamente, despejando $\sum Y_i X_i$ de la ecuación (3.2):

$$\sum_{i=1}^{n} Y_i X_i = a_1 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \tag{3.4}$$

denominada **segunda ecuación normal de Gauss**. Ambas ecuaciones normales conforman un sistema de dos ecuaciones lineales con dos incógnitas (a_1 y b_1). Aplicando el método de los determinantes, las incógnitas se calculan de la siguiente manera:

$$a_1 = \frac{\Delta a_1}{\Delta} \qquad b_1 = \frac{\Delta b_1}{\Delta} \tag{3.5}$$

donde Δa_1, Δb_1 y Δ son los valores de los determinantes:

$$\Delta a_1 = \begin{vmatrix} \sum Y_i & \sum X_i \\ \sum Y_i X_i & \sum X_i^2 \end{vmatrix} = \sum Y_i \sum X_i^2 - \sum Y_i X_i \sum X_i$$

$$\Delta b_1 = \begin{vmatrix} n & \sum Y_i \\ \sum X_i & \sum Y_i X_i \end{vmatrix} = n \sum Y_i X_i - \sum Y_i \sum X_i$$

$$\Delta = \begin{vmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{vmatrix} = n \sum X_i^2 - \left(\sum X_i \right)^2.$$

Luego, reemplazando estas expresiones en las ecuaciones (3.5):

$$a_1 = \frac{\sum Y_i \sum X_i^2 - \sum Y_i X_i \sum X_i}{n \sum X_i^2 - \left(\sum X_i \right)^2} \tag{3.6}$$

$$b_1 = \frac{n \sum Y_i X_i - \sum Y_i \sum X_i}{n \sum X_i^2 - \left(\sum X_i \right)^2} \tag{3.7}$$

donde se puede observar que ambos coeficientes son calculados a partir de expresiones basadas exclusivamente en los datos obtenidos empíricamente. Lo que resta analizar es si el punto crítico obtenido corresponde a un máximo, a un mínimo o a un punto de ensilladura. Para ello calculamos el determinante del Hessiano,

$$\begin{aligned} |H(a_1, b_1)| = \begin{vmatrix} \mathrm{RSS}_{a_1 a_1} & \mathrm{RSS}_{a_1 b_1} \\ \mathrm{RSS}_{b_1 a_1} & \mathrm{RSS}_{b_1 b_1} \end{vmatrix} &= \frac{\partial^2 \mathrm{RSS}}{\partial a_1^2} \frac{\partial^2 \mathrm{RSS}}{\partial b_1^2} - \frac{\partial^2 \mathrm{RSS}}{\partial b_1 \, \partial a_1} \frac{\partial^2 \mathrm{RSS}}{\partial a_1 \, \partial b_1} \\ &= \frac{\partial^2 \mathrm{RSS}}{\partial a_1^2} \frac{\partial^2 \mathrm{RSS}}{\partial b_1^2} - \left(\frac{\partial^2 \mathrm{RSS}}{\partial a_1 \, \partial b_1} \right)^2 \end{aligned} \tag{3.8}$$

donde,

$$\frac{\partial^2 \text{RSS}}{\partial a_1^2} = \frac{\partial}{\partial a_1}\big[-2\sum(Y_i - a_1 - b_1\ X_i)\big] \Rightarrow \frac{\partial^2 \text{RSS}}{\partial a_1^2} = 2n \tag{3.9}$$

$$\frac{\partial^2 \text{RSS}}{\partial b_1^2} = \frac{\partial}{\partial b_1}\big[-2\sum_{i=1}^{n}(Y_i X_i - a_1 X_i - b_1\ X_i^2)\big] \Rightarrow \frac{\partial^2 \text{RSS}}{\partial b_1^2} = 2\sum X_i^2 \tag{3.10}$$

$$\frac{\partial^2 \text{RSS}}{\partial a_1\ \partial b_1} = \frac{\partial}{\partial b_1}\big[-2\sum(Y_i - a_1 - b_1\ X_i)\big] \Rightarrow \frac{\partial^2 \text{RSS}}{\partial a_1\ \partial b_1} = 2\sum X_i \tag{3.11}$$

Luego, reemplazando las ecuaciones (3.9), (3.10) y (3.11) en la expresión (3.8):

$$|H(a_1, b_1)| = 4n\sum X_i^2 - 4\bigg(\sum X_i\bigg)^2$$

en la cual vemos que:

$$|H(a_1, b_1)| = 4n^2\left[\frac{\sum X_i^2}{n} - \frac{\big(\sum X_i\big)^2}{n^2}\right] = 4n^2\left[\frac{\sum X_i^2}{n} - \overline{X}^2\right] = 4n^2 S_X^2 \geq 0 \tag{3.12}$$

donde el carácter ≥ 0 está dado por S_X^2, que es la varianza de la variable X_i. Además, por lo visto en el capítulo anterior, sabemos que $S_X^2 = 0 \Leftrightarrow X_i$ es una constante, por lo cual en ese caso no existirá una dependencia lineal entre las variables. Por lo tanto, al ser $|H(a, b)| = 4n^2 S_X^2 > 0$ y $\text{RSS}_{a_1 a_1} = 2n > 0$ estamos en presencia de un mínimo.

Además, es interesante observar que es posible arribar a la misma conclusión recordando la segunda propiedad de la media aritmética. En efecto, si se analiza la primera ecuación normal de Gauss,

$$\sum_{i=1}^{n} Y_i = na_1 + b_1\sum_{i=1}^{n} X_i \Rightarrow \sum_{i=1}^{n} Y_i - na_1 - b_1\sum_{i=1}^{n} X_i = 0$$
$$\Rightarrow \sum_{i=1}^{n}(Y_i - a_1 - bX_i) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0$$

vemos que en su recorrido a través del diagrama de dispersión, la recta de ajustamiento se comporta como una medida de posición aunque de **carácter dinámico** (no de carácter estático, como sería el caso de la media aritmética) ya que verifica que $\sum(Y_i - \hat{Y}_i) = 0$.

Desde el punto de vista de la estadística, la interpretación de los coeficientes es la siguiente:

- El coeficiente a_1 indica cuál es la cantidad promedio de la variable Y_i para un valor igual a cero de la variable X_i.
- El coeficiente b_1 indica cuál es la variación promedio de la variable Y_i correspondiente a una variación unitaria de la variable X_i.

Como se viene realizando hasta acá, el trabajo de cálculo de los coeficientes es posible organizarlo construyendo la tabla siguiente:

$\boldsymbol{X_i}$	$\boldsymbol{Y_i}$	$\boldsymbol{X_i^2}$	$\boldsymbol{X_iY_i}$
X_1	Y_1	X_1^2	X_1Y_1
X_2	Y_2	X_2^2	X_2Y_2
$\dots$	$\dots$	$\dots$	$\dots$
X_n	Y_n	X_n^2	X_nY_n
$\sum X_i$	$\sum Y_i$	$\sum X_i^2$	$\sum X_iY_i$

con la cual se obtienen todos los términos involucrados en el cálculo de los coeficientes.

Ejemplo. Sean dos variables X_i e Y_i para las cuales se dispone de los siguientes datos empíricos:

$\boldsymbol{X_i}$	$\boldsymbol{Y_i}$	$\boldsymbol{X_i^2}$	$\boldsymbol{X_iY_i}$
1	3	1	3
2	5	4	10
3	1	9	3
4	2	16	8
5	4	25	20
15	**15**	**55**	**44**

Examinando el diagrama de dispersión es posible verificar que la disposición de los puntos no resulta ser muy lineal, sin embargo se puede ver que no existe impedimento alguno para efectuar el cálculo de una función de ajustamiento lineal.

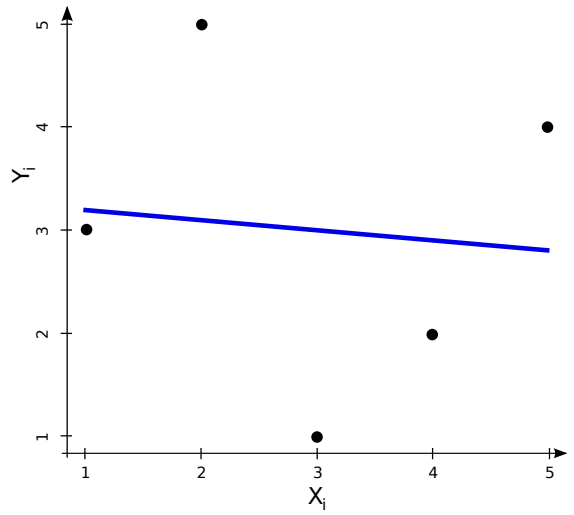


Figura 3.6: Diagrama de dispersión y recta de ajuste lineal.

Utilizando las ecuaciones (3.6) y (3.7):

$$a_1 = \frac{15 \cdot 55 - 44 \cdot 15}{5 \cdot 55 - (15)^2} = \frac{825 - 660}{275 - 225} = \frac{165}{50} = 3{,}3$$
$$b_1 = \frac{220 - 225}{50} = \frac{-5}{50} = -0{,}10$$

Por consiguiente, la ecuación de la recta de ajuste es $\hat{Y}_i = 3{,}3 - 0{,}10X_i$.

Método abreviado

Este método se basa en una transformación de la variable X_i a efectos de simplificar las ecuaciones que definen los coeficientes a_1 y b_1. Para ello, se define la variable $x_i = X_i - \overline{X}$ mediante la cual se verifica que

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (X_i - \overline{X}) = 0$$

por la segunda propiedad de la media aritmética. De esta manera, si se efectuara el desarrollo teórico para encontrar las fórmulas de los coeficientes con las variables x_i e Y_i en lugar de las variables X_i e Y_i, las ecuaciones normales que se obtendrían tendrían la forma:

$$\sum_{i=1}^{n} Y_i = na_1' + b_1' \sum_{i=1}^{n} x_i = na_1'$$
$$\sum_{i=1}^{n} Y_i \; x_i = a_1' \sum_{i=1}^{n} x_i + b_1' \sum_{i=1}^{n} x_i^2 = b_1' \sum_{i=1}^{n} x_i^2$$

De ambas ecuaciones normales así construidas y mediante un simple pasaje de términos, se obtienen las expresiones para calcular los nuevos coeficientes mediante el método abreviado:

$$a_1' = \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y} \qquad b_1' = \frac{\sum_{i=1}^{n} x_i \; Y_i}{\sum_{i=1}^{n} x_i^2}$$

La simple comparación entre éstas y las fórmulas obtenidas originalmente permite comprobar que el método abreviado reduce notoriamente su tamaño y complejidad. Utilizando los nuevos coeficientes la recta de ajuste resulta:

$$\hat{Y}_i = a_1' + b_1' x_i \tag{3.13}$$

Sin embargo, si bien el método abreviado permite calcular los coeficientes mediante fórmulas más breves, al concluir el cálculo no se obtienen a_1 y b_1. Para llegar a esos valores se parte de considerar las dos expresiones posibles para la recta de ajustamiento,

$$\hat{Y}_i = a_1 + b_1 X_i \tag{3.14}$$
$$\hat{Y}_i = a_1' + b_1' x_i \tag{3.15}$$

reemplazando en la ecuación (3.15) la transformación $x_i = X_i - \overline{X}$:

$$\hat{Y}_i = a'_1 + b'_1(X_i - \overline{X}) = a'_1 + b'_1 X_i - b'_1\ \overline{X} = (a'_1 - b'_1\ \overline{X}) + b'_1 X_i$$

con lo cual, al comparar con la ecuación (3.14) se observa que:

$$\begin{aligned} b_1 &= b'_1 \\ a_1 &= a'_1 - b'_1\ \overline{X} = \overline{Y} - b_1\overline{X} \end{aligned}$$

con lo cual se obtienen los verdaderos coeficientes a_1 y b_1 a partir de los calculados a'_1 y b'_1.

Interpretación gráfica del método abreviado

La base del método abreviado consiste en transformar la variable X_i en una variable centrada x_i. Ahora bien, desde el punto de vista gráfico, como construir una variable centrada significa restar $\overline{X}$ a todos los valores de la variable X_i, esa construcción implica correr todos los puntos empíricos $\overline{X}$ unidades hacia la izquierda o, lo que es lo mismo, correr el eje de las ordenadas $\overline{X}$ unidades hacia la derecha. En la figura (3.7) es posible visualizar lo señalado precedentemente. Puede verse que en el gráfico se han representado los n puntos empíricos y la recta de ajustamiento y que se han indicado dos ejes de abscisas que deben utilizarse alternativamente según se trabaje con la variable X_i o x_i , con lo que claramente se descubre la correspondencia entre los valores de ambas variables, de modo que el valor $\overline{X}$ en el eje X_i corresponde al valor cero en el eje x_i.

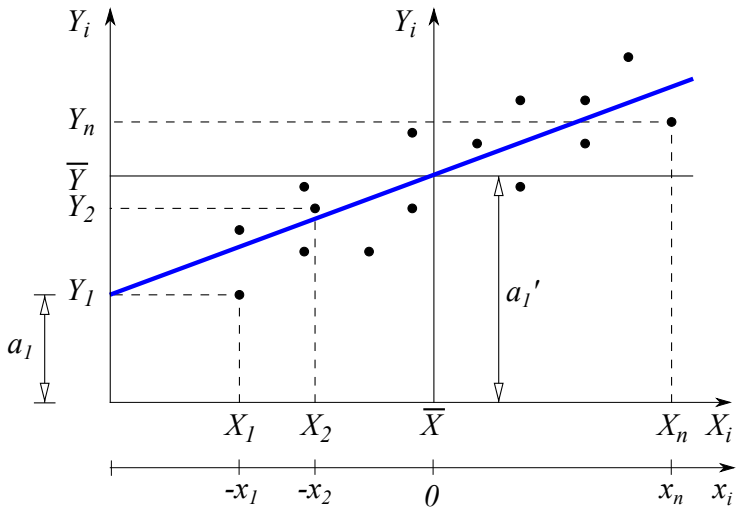


Figura 3.7: Interpretación gráfica del método abreviado.

También puede observarse que el eje Y_i se presenta tanto en su posición original como en una nueva posición, corrido hacia la derecha, donde en este caso su trazado coincide con $\overline{X}$. Como correr los ejes hacia uno u otro lado no modifica la pendiente de la recta, fácilmente se comprende que b'_1 es igual a b_1 (ambos valores representan la tangente del ángulo β) mientras que lo que sí se modifica con el corrimiento del eje Y_i es la ordenada al origen de la recta de ajustamiento, por lo que a' difiere del valor de a.

Si se amplía un sector del gráfico como en la figura (3.8), podemos demostrar trigonométricamente la ecuación que permite calcular a_1:

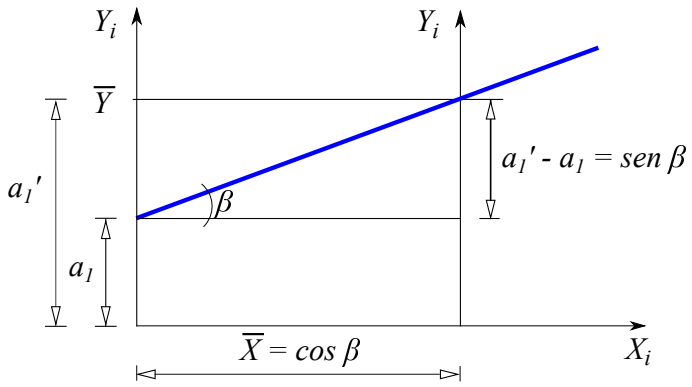


Figura 3.8: Interpretación gráfica del método abreviado.

$$\tan\beta = b_1 = \frac{a_1' - a_1}{\overline{X}} \Rightarrow a_1 = a_1' - b_1\overline{X} = \overline{Y} - b_1\overline{X}$$

Caso inverso (Y_i como variable independiente)

El caso inverso consiste en imaginar una alternativa que resulta sólo posible desde el punto de vista teórico, y que consiste en que la variable independiente sea Y_i en lugar de X_i. Se reitera que esta posibilidad sólo puede presentarse teóricamente dado que en cualquier problema de ajustamiento siempre se define anticipadamente cuál es la variable independiente y a ella se la simboliza comúnmente con X_i.

Sin embargo, una vez definida esta circunstancia, puede pensarse que el conjunto particular de datos con el que se está trabajando puede originar otro problema de ajustamiento, que denominaremos **caso inverso** en el que la variable independiente sea la simbolizada tradicionalmente con Y_i. Gráficamente, esto dá lugar a la aparición de una segunda recta de ajustamiento simbolizada como

$$\hat{X}_i = a_2 + b_2 Y_i \tag{3.16}$$

En general ocurre que $a_1 \neq a_2$, $b_1 \neq b_2$ y que el trazado de $\hat{Y}_i$ no coincide con el de $\hat{X}_i$. Las ecuaciones normales de la recta de ajustamiento inversa, $\hat{X}_i$, son similares a las correspondientes a $\hat{Y}_i$, pero con las variables cambiadas:

$$\begin{aligned} \sum_{i=1}^{n} X_i &= na_2 + b_2 \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i &= a_2 \sum_{i=1}^{n} Y_i + b_2 \sum_{i=1}^{n} Y_i^2 \end{aligned} \tag{3.17}$$

Finalmente, las fórmulas de los coeficientes a_2' y b_2' del caso inverso, calculados mediante el método abreviado, son

$$a'_2 = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X} \qquad b'_2 = \frac{\sum_{i=1}^{n} y_i\ X_i}{\sum_{i=1}^{n} y_i^2}$$

Representando gráficamente ambas ecuaciones, es decir, $\hat{Y}_i = f(X_i)$ y $\hat{X}_i = f(Y_i)$ se puede observar de que aparecen dos rectas que se intersectan en un punto en común. Si bien estas dos rectas son diferentes, el conjunto de datos que relaciona linealmente es el mismo, cambiando variables dependientes e independientes.

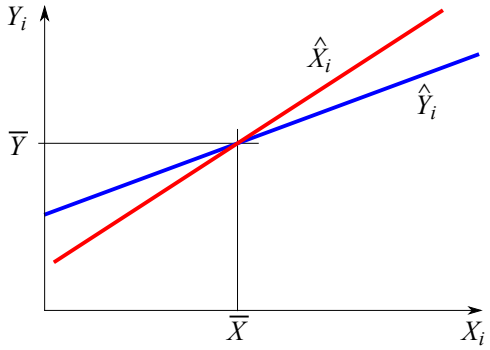


Figura 3.9: Representación del caso inverso.

En la figura (3.9) se observan las rectas de ajustamiento $\hat{Y}_i$ e $\hat{X}_i$ y se evidencia que el punto en común entre ambas es el punto de coordenadas $(\overline{X}, \overline{Y})$. Para determinar matemáticamente esta intersección, partiremos de la primer ecuación normal de Gauss para ambas rectas:

$$\sum_{i=1}^{n} Y_i = na_1 + b_1 \sum_{i=1}^{n} X_i \qquad \sum_{i=1}^{n} X_i = na_2 + b_2 \sum_{i=1}^{n} Y_i$$

y dividiendo por n:

$$\frac{\sum Y_i}{n} = a_1 + b_1\ \frac{\sum X_i}{n} \Rightarrow \overline{Y} = a_1 + b_1\ \overline{X}$$
$$\frac{\sum X_i}{n} = a_2 + b_2 \frac{\sum Y_i}{n} \Rightarrow \overline{X} = a_2 + b_2\ \overline{Y}$$

demostrando así que el punto de coordenadas $(\overline{X}, \overline{Y})$ satisface ambas ecuaciones, por lo que ambas rectas de ajustamiento pasan por ese punto.

3.2. Teoría de la Correlación Lineal

La teoría de la correlación lineal reúne un conjunto de procedimientos matemáticos que permiten calcular el coeficiente de correlación lineal (r) que permite medir:

- en forma directa, el grado de relación lineal entre dos variables X_i e Y_i.
- en forma indirecta, dado un diagrama de dispersión determinado, si un ajustamiento lineal es bueno o no.

Por consiguiente, el coeficiente de correlación lineal nos suministra un valor objetivo mediante el cual es posible decidir si resulta conveniente o apropiado considerar un ajuste lineal o, en caso contrario, buscar una solución diferente (algún ajustamiento no lineal).

Tipos de correlación lineal

En la figura (3.10) se muestran los diferentes tipos de correlación lineal. El gráfico de la izquierda presenta un conjunto de puntos con una *relación lineal directa* entre las dos variables bajo estudio, es decir que cuando cualquiera de las dos variables crece, le corresponde un crecimiento de la otra. Por otra parte, en el gráfico de la derecha el conjunto de puntos presenta una *correlación lineal inversa*, lo cual significa que cuando cualquiera de las dos variables crece, la otra decrece y viceversa. Los puntos del caso 3 muestran una situación en la que la correlación *lineal* es inexistente, pero podría existir una correlación de cualquier otro tipo. En ese caso, el valor del coeficiente de correlación lineal indicaría que lo que no existe es una relación lineal entre las variables, no significando que sea imposible la existencia de una relación (por ejemplo, circular) entre ellas. Aquí se puede observar el importante detalle que las rectas de ajustamiento $\hat{Y}_i$ y $\hat{X}_i$ se cruzan formando un ángulo de 90°.

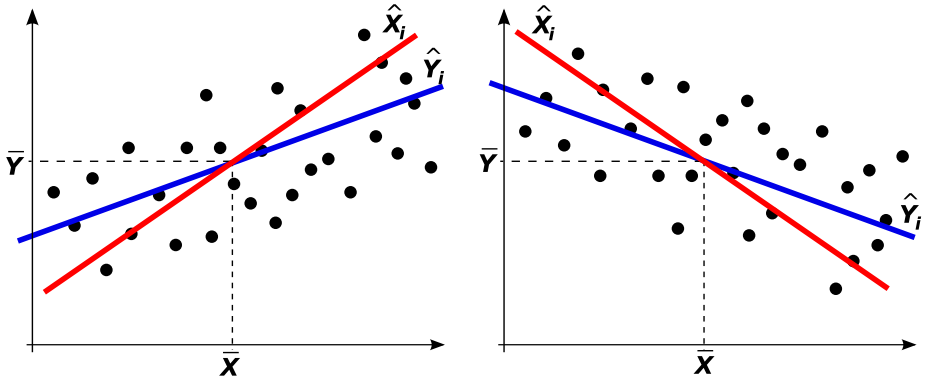


Figura 3.10: Correlación lineal directa (*izquierda*). Correlación lineal inversa (*derecha*).

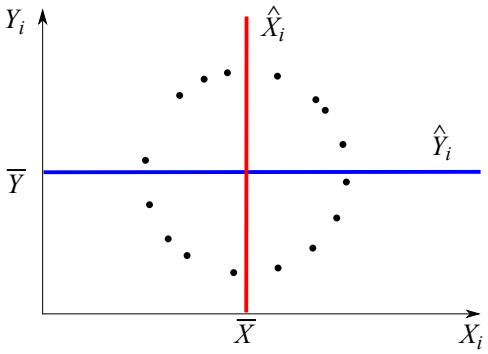


Figura 3.11: Correlación lineal nula.

Existe un caso extremo poco frecuente desde el punto de vista empírico denominado "correlación lineal perfecta" (directa o inversa) en el cual todos los puntos del diagrama de dispersión se encuentran perfectamente alineados y, por consiguiente, coinciden con las dos rectas de ajustamiento $\hat{Y}_i$ y $\hat{X}_i$, como se puede observar en la figura (3.12).

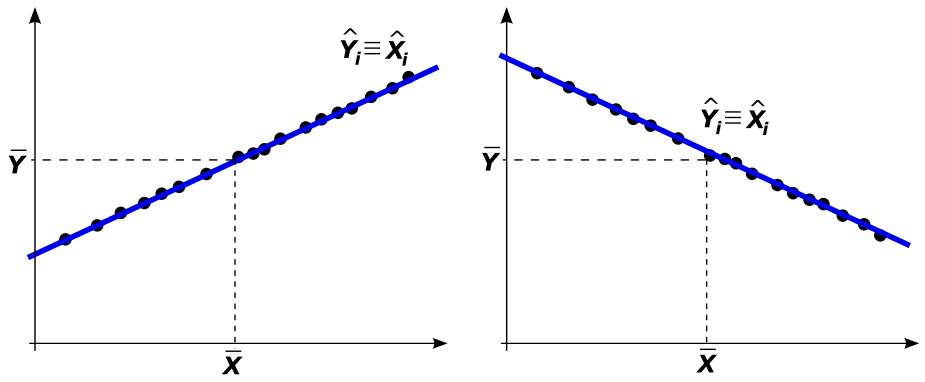


Figura 3.12: Correlación lineal directa perfecta (*izquierda*). Correlación lineal inversa perfecta (*derecha*).

Cálculo del coeficiente de correlación lineal

El coeficiente de correlación r puede calcularse mediante la fórmula de los momentos, propuesta por el matemático Pearson y cuya expresión simbólica es la siguiente:

$$r = \frac{COV(XY)}{DS(X)\ DS(Y)} = \frac{S_{xy}}{S_x\ S_y} \tag{3.18}$$

es decir,

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}\ \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{3.19}$$

Ahora bien, si se reemplazan cada uno de los términos de la ecuación (3.18) por sus correspondientes fórmulas de trabajo, se obtiene la fórmula de trabajo del coeficiente de correlación:

$$r = \frac{\frac{\sum_{i=1}^{n} X_i\ Y_i}{n} - \overline{X}\ \overline{Y}}{\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2}\ \sqrt{\frac{\sum_{i=1}^{n} Y_i^2}{n} - \overline{Y}^2}} \tag{3.20}$$

la cual es utilizada para calcular el valor de r en la mayoría de los casos. Trabajando algebraicamente es posible obtener una fórmula abreviada:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\ \sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{\sum_{i=1}^{n} x_i\ y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\ \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{3.21}$$

recordando que $x_i = X_i - \overline{X}$ e $y_i = Y_i - \overline{Y}$.

Para calcular r mediante la aplicación de la fórmula de trabajo, el cálculo puede ordenarse siguiendo el formato de la tabla utilizada para el ajustamiento lineal más una columna con los Y_i^2:

$\boldsymbol{X_i}$	$\boldsymbol{Y_i}$	$\boldsymbol{X_iY_i}$	$\boldsymbol{X_i^2}$	$\boldsymbol{Y_i^2}$
X_1	Y_1	X_1Y_1	X_1^2	Y_1^2
X_2	Y_2	X_2Y_2	X_2^2	Y_2^2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
X_n	Y_n	X_nY_n	X_n^2	Y_n^2
$\sum \boldsymbol{X_i}$	$\sum \boldsymbol{Y_i}$	$\sum \boldsymbol{X_iY_i}$	$\sum \boldsymbol{X_i^2}$	$\sum \boldsymbol{Y_i^2}$

Cálculo de r a partir del producto de las pendientes

Recordando el método abreviado visto en el Capítulo 3, podemos encontrar la pendiente b_1' de la recta de ajustamiento $\hat{Y}_i$ mediante la fórmula:

$$b_1' = \frac{\sum_{i=1}^{n} x_i \; Y_i}{\sum_{i=1}^{n} x_i^2}$$

donde, reemplazando por $Y_i = y_i + \overline{Y}$ se puede observar que

$$b_1' = b_1 = \frac{\sum_{i=1}^{n} x_i \; (y_i + \overline{Y})}{\sum_{i=1}^{n} x_i^2} = \frac{\sum x_i \; y_i + \overline{Y} \sum x_i}{\sum x_i^2} = \frac{\sum x_i \; y_i}{\sum x_i^2} \tag{3.22}$$

debido a que $\sum x_i = 0$. Con idéntico criterio, reemplazando $X_i = x_i + \overline{X}$ en la ecuación de b_2':

$$b_2' = b_2 = \frac{\sum_{i=1}^{n} x_i \; y_i}{\sum_{i=1}^{n} y_i^2} \tag{3.23}$$

Multiplicando entre sí las expresiones (3.22) y (3.23):

$$b_1 b_2 = \frac{\left(\sum_{i=1}^{n} x_i \; y_i\right)^2}{\sum_{i=1}^{n} x_i^2 \; \sum_{i=1}^{n} y_i^2} = r^2 \quad \Rightarrow \quad r = \pm\sqrt{b_1 b_2} \tag{3.24}$$

Expresión que, además de posibilitar el cálculo del coeficiente r mediante el producto de las pendientes de las rectas de ajustamiento, permite extraer dos conclusiones:

- *a)* las pendientes de las rectas de ajustamiento $\hat{Y}_i$ y $\hat{X}_i$ **tienen el mismo signo** (lo cual las hace crecientes o decrecientes simultáneamente) o **ambas son nulas**. De lo contrario, r no podría ser calculado.
- *b)* el signo del coeficiente r es, por convención, idéntico al de las pendientes. Si la relación es directa el signo de r será positivo, si es inversa será negativo.

Cálculo de r a partir de las variaciones

Variaciones, variancias y errores estándar

Sea un conjunto de datos apareados que contienen un punto muestral (X_i, Y_i), donde $\hat{Y}_i$ es el valor predicho de Y_i (punto teórico) y la media aritmética de los valores muestrales Y_i es $\overline{Y}$, tal como se muestra en detalle en la figura (3.13).

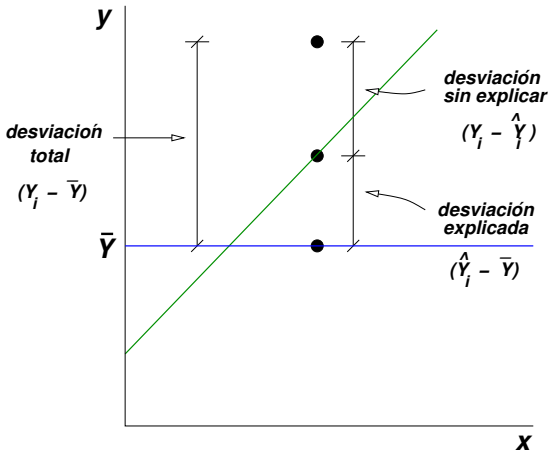


Figura 3.13: Desvío de un punto empírico respecto de la recta de ajuste lineal.

En base a este gráfico, es posible definir las siguientes *variaciones*:

- Tal como se ha definido en los capítulos anteriores, la desviación total o desvío de un punto particular (X_i, Y_i) es la distancia vertical $Y_i - \overline{Y}$ (distancia entre el punto (X_i, Y_i) y la línea horizontal que representa la media muestral $\overline{Y}$). Luego, la **variación total** es la sumatoria de los desvíos al cuadrado entre los puntos empíricos Y_i y la media aritmética $\overline{Y}$:

$$VT = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{3.25}$$

Vemos que si dividimos la expresión (3.25) por n se obtiene la variancia de la variable Y_i:

$$\frac{VT}{n} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = S_y^2 \tag{3.26}$$

- La desviación explicada es la distancia vertical $\hat{Y}_i - \overline{Y}$, entre el valor teórico $\hat{Y}_i$ y la media muestral $\overline{Y}$. Luego, la **variación explicada** es la sumatoria de los desvíos al cuadrado entre los puntos teóricos $\hat{Y}_i$ y la media aritmética $\overline{Y}$:

$$VE = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 \tag{3.27}$$

- La desviación no explicada es la distancia vertical $Y_i - \hat{Y}_i$, entre el punto (X_i, Y_i) y la recta de regresión (la distancia $Y_i - \hat{Y}_i$ también se conoce como *residual*). Luego, la **variación no explicada** es la sumatoria de los desvíos al cuadrado entre los puntos empíricos Y_i y los teóricos $\hat{Y}_i$:

$$\overline{VE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{3.28}$$

Dividiéndola por n se obtiene:

$$\frac{\overline{VE}}{n} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n} = S_{y,x}^2$$

Analizando esta expresión se puede observar que tiene la forma de una variancia, midiendo cómo se alejan los puntos del diagrama de dispersión respecto de la *media dinámica* que es la recta de ajustamiento $\hat{Y}_i$. Esta variancia $S_{y,x}^2$ se denomina "variancia del estimador de Y en X" o simplemente, "variancia del estimador". La raíz cuadrada de la variancia del estimador permite obtener el error estándar del estimador, $S_{y,x}$.

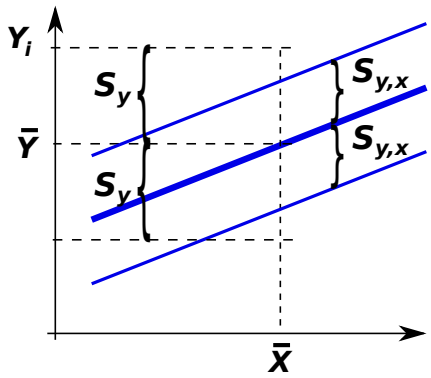


Figura 3.14: Comparación entre los desvíos estándar con respecto a la media aritmética y a la recta de regresión.

En la figura (3.14) se puede observar la diferencia entre los desvíos estándar S_y y $S_{y,x}$, donde la primera mide el alejamiento *en promedio* de los puntos empíricos con respecto a la media aritmética $\overline{Y}$ mientras la segunda mide el alejamiento *en promedio* de los puntos empíricos con respecto a la recta de ajustamiento $\hat{Y}_i$.

Justificación de las denominaciones de las variaciones

- Variación explicada: se denomina así porque en su cálculo intervienen los puntos teóricos $\hat{Y}_i$ y la media aritmética de la variable $\overline{Y}$, ambas *explicadas matemáticamente* mediante sus respectivas ecuaciones.
- Variación no explicada: se denomina así porque en su cálculo intervienen los puntos empíricos Y_i, cuya presencia en el diagrama de dispersión no se encuentra explicada por ningún modelo, ya que responden a datos originados en observaciones experimentales, y en consecuencia, sujetos al azar.
- Variación total: se denomina así porque resulta de la sumatoria de las variaciones anteriores.

$$\underbrace{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}_{\text{variación total}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2}_{\text{variación explicada}} + \underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}_{\text{variación no explicada}} \tag{3.29}$$

Ahora bien, observando detenidamente las expresiones correspondientes a cada una de las variaciones, se puede verificar que:

a) las tres son positivas, ya que son calculadas como sumas de desvíos al cuadrado. No pueden adoptar valores negativos.

b) tanto VE como $\overline{VE}$ pueden ser nulas (y por lo tanto también VT). Esto ocurre cuando los puntos teóricos coinciden con la media aritmética $\overline{Y}$ en el primer caso, y cuando los puntos empíricos coinciden con los teóricos $\hat{Y}_i$ en el segundo.

Relaciones entre las variaciones

Se ha mencionado precedentemente que la variación total es la suma de las variaciones explicada y no explicada, es decir, $VT = VE + \overline{VE}$, ecuación (3.29). Esta ecuación se puede demostrar partiendo de la siguiente relación:

$$Y_i - \overline{Y} = Y_i - \overline{Y} + \hat{Y}_i - \hat{Y}_i = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$$

elevando al cuadrado ambos miembros:

$$(Y_i - \overline{Y})^2 = \left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})\right]^2 = (Y_i - \hat{Y}_i)^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) + (\hat{Y}_i - \overline{Y})^2$$

y aplicando el operador sumatoria en ambos miembros:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

Analizando esta última expresión se debe observar que

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\overline{Y})=\sum(Y_i-a_1-b_1X_i)(a_1+b_1X_i-\overline{Y})=a_1\sum(Y_i-a_1-b_1X_i)+$$
$$+b_1\sum(Y_i-a_1-b_1X_i)X_i-\overline{Y}\sum(Y_i-a_1-b_1X_i)=0$$

dado que, según la primer y segunda ecuación normal de Gauss, ecuaciones (3.3) y (3.4), respectivamente, reescritas a continuación:

$$\sum_{i=1}^{n}Y_i=na+b\sum_{i=1}^{n}X_i\Rightarrow\sum_{i=1}^{n}Y_i-na-b\sum_{i=1}^{n}X_i=\sum_{i=1}^{n}(Y_i-a-bX_i)=0 \tag{3.30}$$

$$\sum_{i=1}^{n}Y_iX_i=a\sum_{i=1}^{n}X_i+b\sum_{i=1}^{n}X_i^2\Rightarrow\sum_{i=1}^{n}Y_iX_i-a\sum_{i=1}^{n}X_i-b\sum_{i=1}^{n}X_i^2=$$
$$=\sum_{i=1}^{n}(Y_i-a-bX_i)X_i=0 \tag{3.31}$$

Por lo cual, la variación total puede ser expresada como:

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2\Rightarrow VT=VE+\overline{VE} \tag{3.32}$$

Cálculo de las variaciones

Cada una de las variaciones puede ser calculada de forma independiente:

- Variación no explicada.

$$\begin{aligned}\overline{VE}&=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-a_1-b_1\ X_i)(Y_i-a_1-b_1\ X_i)\\&=\sum\left[Y_i\ (Y_i-a_1-b_1\ X_i)-a_1(Y_i-a_1-b_1\ X_i)-b_1\ X_i(Y_i-a_1-b_1\ X_i)\right]\\&=\sum Y_i^2-a_1\ \sum Y_i-b_1\ \sum X_i\ Y_i-a_1\sum(Y_i-a_1-b_1\ X_i)-\\&-b_1\ \sum(Y_i-a_1-b_1\ X_i)X_i\end{aligned}$$

los últimos dos terminos se anulan, según las ecuaciones (3.30) y (3.31), por lo tanto:

$$\overline{VE}=\sum_{i=1}^{n}Y_i^2-a_1\ \sum_{i=1}^{n}Y_i-b_1\ \sum_{i=1}^{n}X_i\ Y_i \tag{3.33}$$

- Variación explicada. Partiendo de la ecuación (3.32) y utilizando la (3.33):

$$VE = VT - \overline{VE} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \left[\sum_{i=1}^{n} Y_i^2 - a_1 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i\ Y_i\right]$$
$$= \sum (Y_i^2 - 2Y_i\overline{Y} + \overline{Y}^2) - \sum Y_i^2 + a_1 \sum Y_i + b_1 \sum X_i\ Y_i$$
$$= \sum Y_i^2 - 2\overline{Y} \sum Y_i + n\overline{Y}^2 - \sum Y_i^2 + a_1 \sum Y_i + b_1 \sum X_i\ Y_i$$

eliminando términos y reordenando,

$$VE = -2\overline{Y}n\frac{\sum Y_i}{n} + n\overline{Y}^2 + a_1 \sum Y_i + b_1 \sum X_i\ Y_i$$
$$= a_1 \sum Y_i + b_1 \sum X_i\ Y_i + n\overline{Y}^2 - 2n\overline{Y}^2$$
$$= a_1 \sum_{i=1}^{n} Y_i + b_1 \sum_{i=1}^{n} X_i\ Y_i - n\overline{Y}^2$$

Cálculo de r a partir de las variaciones

Si utilizamos el método abreviado, ecuación (3.13) y recordando que $a_1' = \overline{Y}$ y $b_1' = b_1$, podemos escribir la igualdad $\hat{Y}_i - \overline{Y} = b_1\ x_i$. En base a esta última igual es posible plantear:

$$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = b_1^2 \sum_{i=1}^{n} x_i^2$$

Ahora, tomando la ecuación (3.24) para el cálculo del coeficiente r en función de las pendientes, multiplicando y dividiendo por $\sum_{i=1}^{n} x_i^2$:

$$r^2 = \frac{\left(\sum_{i=1}^{n} x_i\ y_i\right)^2 \sum_{i=1}^{2} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2 \sum_{i=1}^{n} y_i^2} = \frac{b_1^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{VE}{VT} \tag{3.34}$$

Luego,

$$r = \pm\sqrt{\frac{VE}{VT}}$$

fórmula que permite extraer las siguientes conclusiones:

- De acuerdo con la demostración de $VT = VE + \overline{VE}$ y con el hecho de que las variaciones pueden ser nulas o positivas,

  si $VE = 0 \Rightarrow \overline{VE} = VT \Rightarrow r^2 = 0 \Rightarrow r = 0$

  si $VE = VT \Rightarrow \overline{VE} = 0 \Rightarrow r^2 = 1 \Rightarrow r = \pm 1$

- Si $r = +1 \Rightarrow r^2 = 1 \Rightarrow VE = VT \Rightarrow \overline{VE} = 0$. Esto implica que $Y_i = \hat{Y}_i$, o sea, los puntos empíricos coinciden con los teóricos. Luego, la correlación lineal es perfecta con pendiente positiva.

- Si $r = -1 \Rightarrow r^2 = 1 \Rightarrow VE = VT \Rightarrow \overline{VE} = 0$. La correlación lineal es perfecta con pendiente negativa.
- Si $r = 0 \Rightarrow r^2 = 0 \Rightarrow VE = 0 \Rightarrow \overline{VE} = VT$. Los puntos teóricos coinciden con la media aritmética. La correlación lineal es nula y las rectas de ajustamiento se cruzan a 90°.

De acuerdo a este análisis, tenemos que:

$$-1 \leq r \leq +1 \qquad \text{y} \qquad r^2 \leq +1$$

Coeficiente de determinación

Se denomina así al coeficiente de correlación lineal al cuadrado (r^2), e indica cuál es la proporción de la variación total que es explicada por el modelo aplicado en un determinado ajustamiento lineal:

$$r^2 = \frac{VE}{VT}$$

De esta manera, el coeficiente de determinación (que suele expresarse como porcentaje) es un indicador objetivo para determinar que porcentaje de la variación total es explicada por el modelo lineal. A modo exclusivamente orientativo, se presenta a continuación un cuadro en el que se presentan diferentes valores del coeficiente de correlación con su correspondiente valor del coeficiente de determinación y una calificación respecto de la calidad del ajustamiento lineal en cada caso.

r	**r^2**	**porcentaje explicado**	**calidad del ajustamiento lineal**
±0,90/ ± 1,00	0,81 a 1,00	81 % a 100 %	muy bueno
±0,80/ ± 0,90	0,64 a 0,81	64 % a 81 %	bueno
±0,70/ ± 0,80	0,49 a 0,64	49 % a 64 %	regular
±0,60/ ± 0,70	0,36 a 0,49	36 % a 49 %	malo
menos de |0,60|	menos de 0,36	menos del 36 %	muy malo

Correlación y dependencia estadística

En función de todo lo explorado en este capítulo es posible realizar las siguientes conclusiones:

a) La existencia de dependencia estadística entre dos variables implica que entre ellas existe algún grado de correlación, pero la inversa no es cierta: **la existencia de correlación entre dos variables no implica que exista dependencia estadística entre ellas**.

Esto quiere decir que la relación entre dos variables puede existir y ser alta, pero eso no significa que dependa estadísticamente una de la otra. Por ejemplo, entre las variables "número de fallecidos en una ciudad" y "cantidad de pájaros en la misma ciudad" puede existir un grado de relación lineal inversa muy estrecha (a

menor número de pájaros, mayor número de fallecidos) pero sin embargo entre ellas no existe ninguna dependencia. Lo que sí existe en este caso es una tercera variable, la *temperatura*, o en otras palabras, los meses del año para los cuales se toma la información (el investigador debe profundizar en su búsqueda para descubrir la posible existencia de esas variables ocultas cuando realiza una investigación de cualquier naturaleza), ya que se puede comprobar fácilmente que en los meses de baja temperatura, tradicionalmente los de invierno, aumenta el número de personas fallecidas y disminuye el número de aves debido a las migraciones. Por lo cual, las variables "número de personas fallecidas" y "cantidad de aves" son estadísticamente independientes.

b) La obtención de un resultado nulo para el coeficiente de correlación lineal r indica que las variables bajo estudio no tienen correlación lineal, o bien, que son linealmente independientes. Sin embargo entre ellas sí puede existir alguna correlacion del tipo no lineal.

c) Recordando que

$$V(x \pm y) = V(x) + V(y) \pm 2\ Cov(x, y) \qquad \text{y} \qquad r = \frac{Cov(x, y)}{S_x\ S_y}$$

es posible enunciar que

Si las variables son linealmente independientes, $r = 0$.

Si $r = 0$ luego $Cov(x, y) = 0$.

Si $Cov(x, y) = 0$ entonces $\mathbf{V}(\mathbf{x} \pm \mathbf{y}) = \mathbf{V}(\mathbf{x}) + \mathbf{V}(\mathbf{y})$

Luego, cuando dos variables X_i e Y_i son linealmente independientes, la variancia de su suma o diferencia es siempre igual a la suma de sus respectivas variancias.

Capítulo 4
Teoría de la Probabilidad

4.1. Antecedentes históricos

La teoría de la Probabilidad fue inicialmente propuesta por el matemático francés Blaise Pascal (1623-1662), quien en 1654, a raíz de consultas que le había planteado un integrante de la nobleza en un intercambio epistolar con Pierre de Fermat (1601-1665), se dedicó a resolver los problemas de probabilidad relacionados con los juegos de azar, muy de moda en las sociedades europeas de aquellos tiempos.

(a) Blaise Pascal (1623 - 1662).

(b) Pierre de Fermat (1601 - 1665).

Así surgió la primera definición de probabilidad, que actualmente se conoce como "definición clásica". Con posterioridad, otros matemáticos de los siglos XVII y XVIII se ocuparon de ampliar los límites del conocimiento en el tema de la Teoría de la Probabilidad. Entre los investigadores más reconocidos se encuentran el propio Fermat, Bernoulli, Gauss, Laplace, Poisson, entre otros.

Con la Revolución Francesa la Teoría de la Probabilidad sufrió un deterioro importante debido fundamentalmente a que, habiendo surgido a partir de intereses no demasiado bien vistos de la nobleza, fue considerado un producto casi despreciable por parte de los científicos e investigadores matemáticos de la época. Eso fue así hasta mediados del siglo XIX, cuando su estudio y análisis vuelven a recibir un fuerte impulso, de modo que a comienzos del siglo XX ya se la considera una herramienta trascendental, aplicada en varias ramas del campo científico.

Hoy en día no existe investigación alguna, en cualquier terreno, que no la considere y la aplique, habiéndose incorporado con naturalidad a los procesos analíticos de la física,

química, ingeniería, medicina, economía, entre otros.

4.2. Definición clásica de la probabilidad

Probabilidad

Suponiendo la existencia de un experimento aleatorio que puede dar lugar a la aparición de un suceso $\boldsymbol{A}$, que puede presentarse de $\boldsymbol{h}$ formas diferentes, todas ellas que le favorecen, de un total de $\boldsymbol{n}$ formas posibles de ocurrencia del experimento; se define a la probabilidad de ocurrencia del suceso A como la relación entre el número de casos favorables h y el número de casos posibles n, es decir:

$$P(A) = \frac{h}{n} \tag{4.1}$$

Ejemplo 4.1: Una moneda *legal* tiene dos lados, a los que llamamos "cara" (C) y "cruz" (X). En el experimento aleatorio "arrojar la moneda al aire", ¿Cuál es la probabilidad de obtener una cara?

Si el número de casos favorables al lado cara es igual a 1 y el total de lados posibles de aparecer es 2, la probabilidad simplemente es: $P(C) = 1/2$.

Ejemplo 4.2: Un dado *legal* tiene seis caras, luego la probabilidad de obtener un 3 en una tirada es: $P(3) = 1/6$.

Para indicar el resultado de una probabilidad en esta etapa de estudio, utilizaremos las formas fraccionarias respetando, de esta manera, el fundamento básico de la definición de Pascal, aunque no existe ningún inconveniente para convertir la forma fraccionaria en una forma decimal. Si bien el concepto de experimento aleatorio es, en principio, ambiguo, al recordar que la definición pascaliana se originó en el estudio de los juegos de azar, puede decirse que un **experimento aleatorio** puede ser tanto una de las repeticiones de algún juego de azar como cualquier otra cosa que, no siendo precisamente un juego, esté sujeta a las reglas del azar, de modo que el resultado que se puede presentar en sus realizaciones se encuentra sujeto a las condiciones de incertidumbre propias de la aleatoriedad.

Como ejemplos de experimentos aleatorios se pueden citar los siguientes:

- La definición del sexo de un bebé al momento de la concepción: varón o mujer.
- La ingesta de un medicamento para combatir una infección: efectivo o no efectivo.
- La asistencia de alumnos a las clases de Estadística: 20, 25, 32, etc.

Analicemos ahora los elementos que componen la fórmula de la definición clásica y veremos que:

- n siempre debe ser mayor o igual a 1, ya que si n es cero no hay realización del experimento.
- h siempre varía entre 0 y n, dado que si es igual a cero no hay casos favorables y como máximo todos los casos pueden ser favorables. Entonces:

$$\begin{aligned} &\text{Si } h = 0 \Rightarrow P(A) = 0 \quad \rightarrow \quad A \text{ se denomina } \textbf{suceso imposible} \\ &\text{Si } h = n \Rightarrow P(A) = 1 \quad \rightarrow \quad A \text{ se denomina } \textbf{suceso seguro o cierto} \end{aligned}$$

Luego se concluye que $\mathbf{0 \leq P(A) \leq 1}$: La probabilidad de ocurrencia de un suceso A es un número real perteneciente al intervalo $[0, 1]$. En las secciones siguientes definiremos dos conceptos importantes vinculados con la definición clásica de la probabilidad: *espacio muestral* y *eventos*.

Espacio muestral

Definición

El conjunto de todos los posibles resultados de un experimento se llama **espacio muestral**.

Por ejemplo, el espacio muestral del experimento que consiste en lanzar una moneda al aire tres veces es:

$$\Omega = \{(c, c, c); (c, c, x); (c, x, c); (c, x, x); (x, c, c); (x, c, x); (x, x, c); (x, x, x)\}.$$

El espacio muestral del experimento "*arrojar una vez un dado*" es:

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

En otro experimento le preguntamos a la primer persona que encontramos en la calle su mes de nacimiento:

$$\Omega = \{\text{Ene, Feb, Mar, Abr, May, Jun, Jul, Ago, Sep, Oct, Nov, Dic}\}$$

Usualmente se designa al espacio muestral con la letra griega $\mathbf{\Omega}$. La definición de un espacio muestral depende del objetivo del análisis. En el primer ejemplo vemos que si bien el espacio muestral no contiene muchos elementos, es muy fácil confundirse con las distintas combinaciones que pueden darse con los tres resultados, y más aún si se arroja un número mayor de veces la moneda. Para describir ordenadamente un espacio muestral es muy útil un esquema conocido como **diagrama de árbol**. Cuando un espacio muestral puede construirse en varios pasos o etapas, entonces cada una de las n_1 maneras de completar el primer paso puede representarse como una rama del árbol. Cada una de las maneras de completar el segundo paso puede representarse con n_2 ramas que comienzan donde terminan las ramas anteriores, y así sucesivamente.

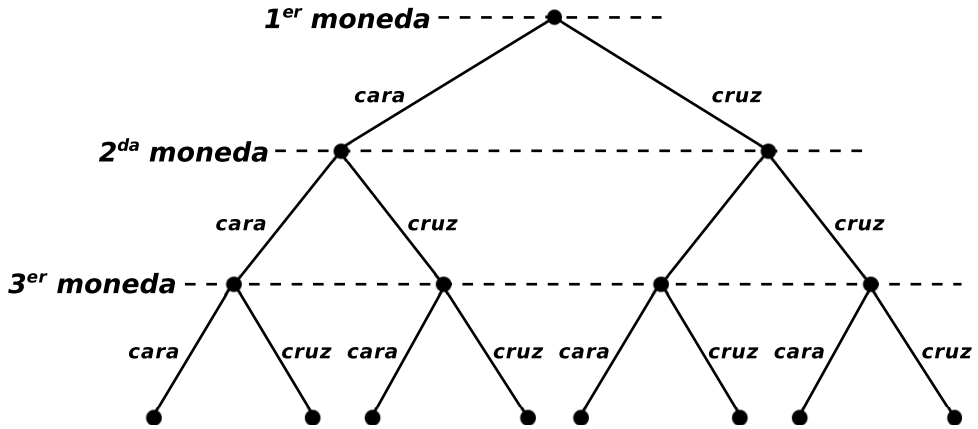


Figura 4.1: Diagrama de árbol para el ejemplo de la moneda arrojada tres veces. Nótese de que es posible pensarlo al problema como 3 monedas que se arrojan una sola vez.

Ejemplo 4.3: Considérese un experimento en el cual debemos seleccionar al azar una pieza de plástico, como por ejemplo un conector, y medir su espesor. Los posibles valores del espesor dependerán de la resolución del instrumento de medición, pero podemos definir al espacio muestral simplemente conteniéndolo en el conjunto de números reales positivos (sin incluir al cero):

$$\Omega = \Re^{+} = \{x/x > 0\}$$

debido a que no podemos obtener valores negativos del espesor. Si es conocido que todos los conectores tienen un espesor entre 10 y 11 milímetros, el espacio muestral podría ser:

$$\Omega = \{x/10 < x < 11\}$$

Si el objetivo del análisis es clasificar los espesores de las piezas como bajo, medio o alto:

$$\Omega = \{\text{bajo, medio, alto}\}$$

Si el objetivo del análisis es considerar sólo si las piezas verifican ciertas especificaciones de la fábrica, el especio muestral podría ser un conjunto de dos posibles resultados:

$$\Omega = \{\text{si, no}\}$$

indicando si la pieza verifica o no.

Un espacio muestral es **discreto** si consiste en un **conjunto finito** o **infinito numerable** de resultados posibles y es **continuo** si contiene un intervalo (finito o infinito) de números reales. Por ejemplo, si en una producción industrial extraemos un artículo para averiguar si es defectuoso o no, Ω puede constar de 2 elementos: D (defectuoso) o N (no defectuoso), por lo tanto el espacio muestral es discreto. El espacio muestral del experimento "medir la resistencia del acero" es continuo debido a que el resultado puede ser cualquier número real positivo dentro de cierto intervalo.

Eventos o Sucesos

Definición

Un **evento** A (respecto a un espacio muestral particular Ω asociado con un experimento) es simplemente un conjunto de resultados posibles contenidos en Ω.

En terminología de conjuntos, un evento es un *subconjunto* del espacio muestral Ω. Esto quiere decir que incluso Ω es un evento como así también lo es el conjunto vacío $\{\emptyset\}$. Cualquier resultado individual también es considerado un evento.

Volviendo al ejemplo de la moneda, si nos interesa que en los lanzamientos aparezcan *al menos* dos caras, el evento será:

$$A = \{(c, c, c); (c, c, x); (c, x, c); (x, c, c)\}.$$

En el caso del dado, si estamos interesados sólo en números pares:

$$A = \{2, 4, 6\}.$$

Si en el experimento de los cumpleaños nos interesan sólo aquellos meses con 31 días:

$$L = \{\text{Ene, Mar, May, Jul, Ago, Oct, Dic}\}.$$

Los eventos pueden combinarse de acuerdo a las operaciones usuales entre conjuntos. Por ejemplo, si definimos a R como el evento correspondiente a la ocurrencia de los meses en cuyos nombres se encuentra la letra "*r*":

$$R = \{\text{Ene, Feb, Mar, Abr, Sep, Oct, Nov, Dic}\}.$$

Por lo que el evento "*ocurrencia de un mes de 31 días cuyo nombre contiene a la letra r*" será

$$L \cap R = \{\text{Ene, Mar, Oct, Dic}\}$$

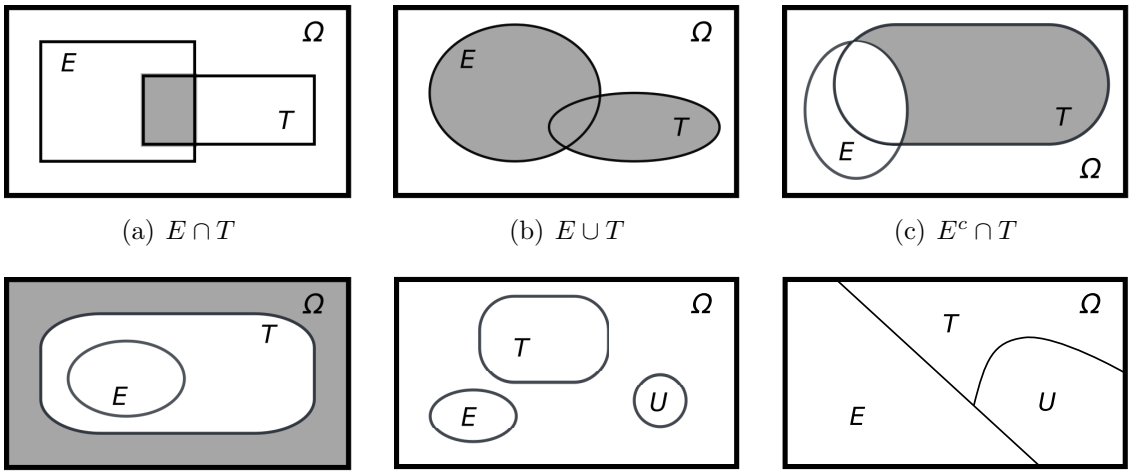
donde $L \cap R$ indica la **intersección** de L y R. La intersección de dos conjuntos L y R es el conjunto de todos los elementos que pertenecen tanto a L como a R. De igual manera, es posible plantear la **unión**, $L \cup R$, conjunto formado por todos los elementos que pertenecen exclusivamente a L, exclusivamente a R, o a ambos conjuntos. En consecuencia,

$$L \cup R = \{\text{Ene, Feb, Mar, Abr, May, Jul, Ago, Sep, Oct, Nov, Dic}\}$$

Dados dos conjuntos E y T, la unión e intersección entre ambos puede expresarse formalmente como:

$$E \cup T = \{x | x \in E \ \vee \ x \in T\} \qquad E \cap T = \{x | x \in E \ \wedge \ x \in T\}$$

Los conjuntos y las operaciones asociadas a ellos son fáciles de visualizar cuando se utilizan *diagramas de Venn*:



(a) $E \cap T$ (b) $E \cup T$ (c) $E^c \cap T$

(d) T^c (e) $\{E \cap T \cap U\} = \{\emptyset\}$ (f) $\{E \cup T \cup U\} = \Omega$

Figura 4.2: Ejemplos de diagramas de Venn.

Álgebra de conjuntos

Las operaciones entre conjuntos poseen varias propiedades, las cuales son consecuencias inmediatas de sus definiciones. Algunas de ellas son:

$$\begin{aligned} R \cup T &= T \cup R, & R \cup (T \cup W) &= (R \cup T) \cup W, \\ R \cap (T \cup W) &= (R \cap T) \cup (R \cap W), & R \cup (T \cap W) &= (R \cup T) \cap (R \cup W), \\ (R^c)^c &= R, & R \cap R^c &= \emptyset, \\ R \cup \Omega &= \Omega, & R \cap \Omega &= R. \end{aligned}$$

Dos propiedades particularmente útiles son las conocidas como **Leyes de De Morgan**, las cuales establecen que

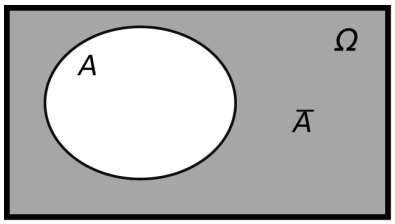
$$\overline{(A \cup B)} = \overline{A} \cap \overline{B} \qquad \text{y} \qquad \overline{(A \cap B)} = \overline{A} \cup \overline{B}$$

o bien, generalizando,

$$\left(\bigcup_n E_n \right)^c = \bigcap_n {E_n}^c \qquad (4.2) \qquad\qquad \left(\bigcap_n E_n \right)^c = \bigcup_n {E_n}^c \qquad (4.3)$$

Sucesos opuestos

Indiquemos con el símbolo $\overline{\boldsymbol{A}}$ (se lee "no A") al suceso o conjunto de sucesos que no son el suceso $\boldsymbol{A}$ (es decir, es el conjunto complementario). Si de los n casos posibles, h favorecen al suceso A, los restantes $(n - h)$ sucesos favorecerán al suceso A. Por consiguiente, partiendo de la definición clásica de la probabilidad:



$$P(\overline{A}) = \frac{n-h}{n} = 1 - \frac{h}{n} \qquad (4.4)$$

Si ahora sumamos las probabilidades de los sucesos A y $\overline{A}$, obtenemos:

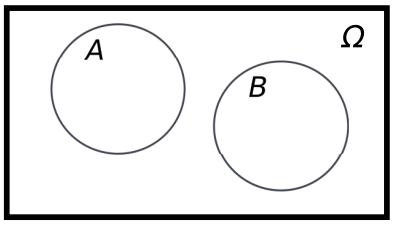
$$P(A) + P(\overline{A}) = \frac{h}{n} + \left(1 - \frac{h}{n}\right) = 1$$

Con lo que podemos decir: **La suma de las probabilidades de dos sucesos opuestos es igual a 1.**

Ejemplo 4.4: Los sucesos C y X en la tirada de una moneda son opuestos. Luego la $P(C) + P(X) = 1$, lo cual demuestra adicionalmente que la aparición de cara o cruz en una sola tirada de una moneda constituye un suceso seguro o, lo que es lo mismo, ambos sucesos conforman un **sistema completo** (ya que no existe otra solución posible para el experimento). También es opuesto el suceso que salga la cara del dado con el 3 a los otros cinco resultados (que no salga 3) posibles en la tirada de un dado.

Sucesos Mutuamente Excluyentes

Dos o más sucesos se denominan mutuamente excluyentes si la ocurrencia de cualquiera de ellos excluye la de los otros. De modo que si A y B son sucesos mutuamente excluyentes, entonces $P(A \cap B) = 0$.



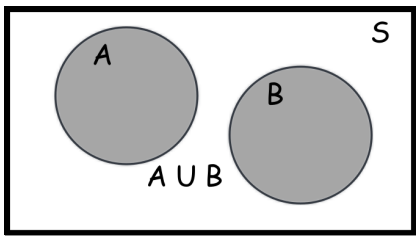
Cuando dos o más sucesos no son mutuamente excluyentes son **compatibles**. Haciendo una analogía con el álgebra de conjuntos, estos eventos serían conjuntos con elementos en común.

Regla de la suma

En una realización de un experimento aleatorio, la probabilidad de ocurrencia del suceso A ó del suceso B (A y B son sucesos mutuamente excluyentes) se resuelve mediante la siguiente ecuación:

$$P(A \cup B) = P(A \; o \; B) = P(A) + P(B) \qquad (4.5)$$

$(A \cup B)$ se interpreta como la "ocurrencia de A o de B", es decir, puede ocurrir A, B o ambos simultáneamente; y se lo indica con el símbolo $\cup$ (unión).



Demostración: Si A ocurrió f_1 veces y B ocurrió f_2 veces, siendo n la cantidad total de observaciones realizadas:

$$P(A \cup B) = \frac{f_1 + f_2}{n} = \frac{f_1}{n} + \frac{f_2}{n} = P(A) + P(B)$$

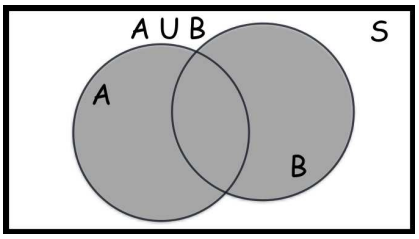
Vemos también que si A es el complemento de $\overline{A}$, tendremos:

$$P(A \cup \overline{A}) = \frac{f_1 + (n - f_1)}{n} = \frac{f_1}{n} + \frac{n - f_1}{n} = P(A) + P(\overline{A}) = 1$$

es decir, un **evento seguro**.

Ahora, si con $A + B$ se denota el suceso de que "ocurra A o B o ambos a la vez" (sucesos compatibles), entonces:

$$P(A \cup B) = P(A \text{ o } B) = P(A) + P(B) - P(A \cap B) \tag{4.6}$$



Para demostrarlo, definiremos 2 eventos mutuamente excluyentes A y $\overline{A} \cap B$. Entonces tendremos:

$$P\big[A \cup (\overline{A} \cap B)\big] = P(A) + P(\overline{A} \cap B)$$

Sabemos que $(\overline{A} \cap B) \cup (A \cap B) = B$, luego

$$P\big[(\overline{A} \cap B) \cup (A \cap B)\big] = P(\overline{A} \cap B) + P(A \cap B) = P(B)$$

dado que $(\overline{A} \cap B)$ y $(A \cap B)$ son sucesos mutuamente excluyentes. Entonces:

$$P(\overline{A} \cap B) = P(B) - P(A \cap B)$$

y como $A \cup (\overline{A} \cap B) = A \cup B$,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

que es análoga a la ecuación (4.6). Se puede observar que si $(A \cap B) = \emptyset$, obtenemos la ecuación (4.5).

Ejemplo 4.5: Se arroja un dado. ¿Cuál es la probabilidad de obtener un uno o un seis? Estos sucesos son *excluyentes*, por lo tanto la solución es:

$$P(1 \cup 6) = 1/6 + 1/6 = 1/3$$

Ejemplo 4.6: Se dispone de un mazo de cartas españolas. ¿Cuál es la probabilidad de sacar un As o una carta de espadas retirando del mazo una sola carta al azar?

La probabilidad de sacar un As es: $P(As) = 4/40 = 1/10$.

La probabilidad de sacar una carta de espadas es: $P(\text{esp}) = 10/40 = 1/4$.

Pero también existe la posibilidad de que salga el As de espadas: $P(As \text{ de esp}) = 1/40$.

Por lo que se trata de *sucesos compatibles* (no excluyentes). La probabilidad de sacar un *As* o una carta de espadas será:

$$P(As \cup \text{esp}) = 4/40 + 10/40 - 1/40 = 13/40.$$

Tres o más eventos

Pueden obtenerse ecuaciones para calcular la unión de tres o más eventos partiendo de las ecuaciones anteriores y con el empleo de algunas operaciones básicas entre conjuntos. Por ejemplo, para 3 eventos A, B y C:

$$P(A \cup B \cup C) = P\big[(A \cup B) \cup C\big] = P(A \cup B) + P(C) - P\big[(A \cup B) \cap C\big] \tag{4.7}$$

desarrollando ahora el término $P(A \cup B)$ de acuerdo a la ecuación (4.6) y utilizando la propiedad distributiva para operaciones entre conjuntos para simplificar $P\big[(A \cup B) \cap C\big]$:

$$\begin{aligned} P(A \cup B \cup C) =& P(A) + P(B) - P(A \cap B) + P(C) - P\big[(A \cap C) \cup (B \cap C)\big] \\ =& P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &+ P(A \cap B \cap C) \end{aligned} \tag{4.8}$$

Con esto se obtiene una fórmula para calcular la probabilidad de la unión de tres eventos. En forma general, la fórmula a emplear para calcular la probabilidad de la unión de un número cualquiera de eventos es:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) = &\sum_{i=1}^{n} P(E_i) - \sum_{i,j:i<j}^{n} P(E_i \cap E_j) + \sum_{i,j,k:i<j<k}^{n} P(E_i \cap E_j \cap E_k) - \ldots \\ &\ldots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} E_i\right) \end{aligned} \tag{4.9}$$

Probabilidad condicional, Sucesos independientes y dependientes

La probabilidad de ocurrencia de un suceso puede depender de otro (o de la no-ocurrencia). Si existe esa dependencia, la probabilidad asociada a ese evento se denomina "probabilidad condicional". Si A y B son dos sucesos, la probabilidad de que B ocurra dado que haya ocurrido A se denota como $P(B/A)$ y se llama **probabilidad condicional de *B* dado *A***.

Si la ocurrencia de A no afecta la probabilidad de ocurrencia de B, entonces $P(B/A) = P(B)$, y diremos que A y B son **sucesos independientes**, en caso contrario, se dirá que son **sucesos dependientes** o **condicionales**.

Por ejemplo:

- Al arrojar dos veces o más una moneda, el resultado de cada experimento *no depende* de lo ocurrido en el experimento anterior, así como tampoco influye en el resultado del experimento siguiente. Luego, al arrojar dos veces (o más) una moneda, se generan sucesos *independientes*.
- Al arrojar simultáneamente dos (o más) monedas, el resultado que se presenta en cada una de ellas *no depende* del resultado que se presenta en las otras, de modo que también en este caso los sucesos son independientes.
- Al seleccionar dos bolillas de un recipiente que contiene varias bolillas, los sucesos "elección de bolillas" son *independientes* si la elección se realiza *con reposición*, es decir, si se repone la primer bolilla extraída del recipiente previamente a la selección de la segunda. En cambio, si la elección se realiza *sin reposición*, los sucesos son *condicionales*, dado que al no reponer la primer bolilla extraída, se modificará el espacio muestral para la selección de la segunda.
- En el caso anterior, si la selección se realiza simultáneamente (extrayendo al mismo tiempo las dos bolillas) se considera que también se generan dos sucesos *condicionales*.

Regla de la multiplicación

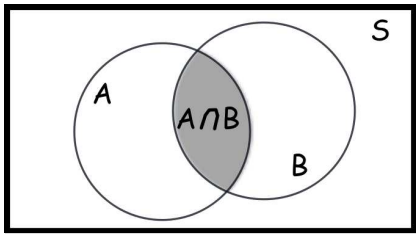
Si denotamos $(A \cap B)$ al suceso de que "A y B ocurran", entonces:

$$P(A \cap B) = P(A)\ P(B/A) = P(B)\ P(A/B) \tag{4.10}$$

Haciendo una analogía con la teoría de conjuntos, la probabilidad condicional puede interpretarse como la probabilidad de extraer un elemento del conjunto A que también pertenezca al conjunto B:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \tag{4.11}$$

También es posible asimilarlo como si fuera una reducción del espacio muestral, es decir, al asumir que B ya ha ocurrido, las posibilidades se restringen a los n_B elementos de B, de los cuales sólo $n_{A \cap B}$ elementos pertenecen al suceso "ocurrencia de A".
Podemos observar que $P(A)$ podría indicarse como $P(A/S) = \dfrac{P(A \cap S)}{P(S)}$,



luego $A \cap S = A$ y $P(S) = 1 \Rightarrow P(A/S) = P(A)$.

Ahora, si los sucesos estadísticamente independientes:

$$P(AB) = P(A \text{ y } B) = P(A)\ P(B)$$

es decir: $P(B/A) = P(B)$ y $P(A/B) = P(A)$.

Ejemplo 4.7: Se arroja una moneda dos veces. Hallar la probabilidad de obtener el suceso cara en la primera tirada y el suceso cruz en la segunda.

$$P(C_1 \text{ y } X_2) = P(C_1)\ P(X_2/C_1) = P(C_1)\ P(X_2) = 1/2 \cdot 1/2 = 1/4$$

En este ejemplo se menciona por primera vez el orden de aparición de los sucesos. En la formulación de la ecuación (4.10) se asume que el orden en que se presentan los sucesos está dado por el orden de la conjunción o intersección entre los eventos. Cuando éste no sea el caso, es decir, cuando se desee contemplar todas las situaciones posibles de presentación, se debe considerar que puede presentarse primero el evento A y luego el B o viceversa. Luego, la ecuación (4.10) debe escribirse:

$$\begin{aligned} P(A \cap B) &= P\big[(A_1 \cap B_2) \cup (B_1 \cap A_2)\big] \\ &= P(A_1)\ P(B_2/A_1) + P(B_1)\ P(A_2/B_1) \end{aligned} \tag{4.12}$$

si los sucesos fuesen condicionales, y

$$\begin{aligned} P(A \cap B) &= P\big[(A_1 \cap B_2) \cup (B_1 \cap A_2)\big] \\ &= P(A_1)\ P(B_2) + P(B_1)\ P(A_2) \end{aligned} \tag{4.13}$$

si los sucesos fuesen independientes.

Ejemplo 4.8: Se arroja una moneda dos veces. Hallar la probabilidad de obtener los sucesos cara y cruz.

$$\begin{aligned} P(C \text{ y } X) &= P\big[(C_1 \cap X_2) \cup (X_1 \cap C_2)\big] \\ &= P(C_1)\ P(X_2/C_1) + P(X_1)\ P(C_2/X_1) = 1/2 \cdot 1/2 + 1/2 \cdot 1/2 = 1/2 \end{aligned}$$

Ejemplo 4.9: Si de un recipiente con 10 esferas blancas y 5 negras se eligen dos sin reposición ¿cuál es la probabilidad de elegir una blanca y una negra?

$$\begin{aligned} P(B \text{ y } N) &= P\big[(B_1 \cap N_2) \cup (N_1 \cap B_2)\big] \\ &= P(B_1)\ P(N_2/B_1) + P(N_1)\ P(B_2/N_1) \\ &= 10/15 \cdot 5/14 + 5/15 \cdot 10/14 = 100/210 \end{aligned}$$

Observando los ejemplos anteriores es posible concluir que si la realización repetida de un experimento genera sucesos independientes, sus probabilidades se mantendrán constantes a lo largo de toda la serie de realizaciones. En cambio, si los sucesos generados en un experimento resultan condicionales, sus probabilidades variarán de realización en realización.

4.3. Probabilidad Total

Sea $E_1, E_2, \ldots, E_n$ un *sistema completo de sucesos* tales que la probabilidad de cada uno de ellos es distinta de cero. Es decir,

$$\begin{aligned} P(E_1 \cup E_2 \cup \ldots \cup E_n) = P(\Omega) &= 1; \\ P(E_i \cap E_j) &= 0 \quad \text{con } i \neq j; \\ P(E_i) &> 0, \quad i = 1 \ldots n. \end{aligned}$$

Un evento A, asociado a cada uno de los sucesos E_i puede obtenerse observando que:

$$A = (E_1 \cap A) \cup (E_2 \cap A) \cup \ldots \cup (E_n \cap A)$$

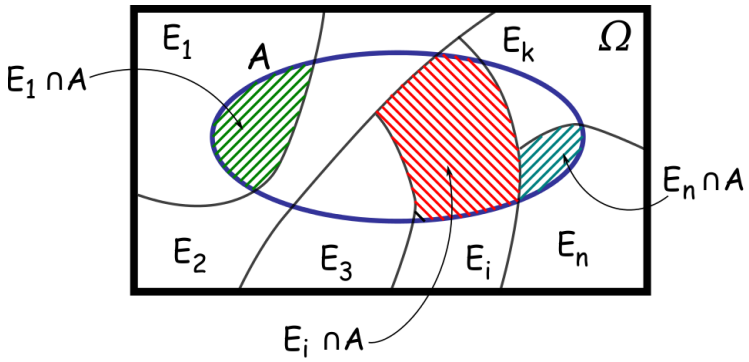


Figura 4.3: Sistema completo de sucesos E_i y evento asociado A.

por lo cual su probabilidad de ocurrencia puede calcularse como:

$$\begin{aligned} P(A) &= P[(E_1 \cap A) \cup (E_2 \cap A) \cup \ldots \cup (E_n \cap A)] \\ &= P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + \ldots + P(E_n)P(A/E_n) \end{aligned}$$

donde en la última expresión se utilizó la regla de la suma para eventos mutuamente excluyentes. Finalmente, el cálculo de la probabilidad del evento A, conocidas las probabilidades condicionales $P(A/E_i)$, queda definido como:

$$P(A) = \sum_{i=1}^{n} P(E_i)\ P(A/E_i)$$

ecuación conocida como "probabilidad total".

Teorema de Bayes

La definición de probabilidad condicional dada por la ecuación (4.11) es válida para cualquier par de sucesos, siendo posible verificar además que:

$$P(A \cap B) = P(A)\ P(B/A) = P(B)\ P(A/B) = P(B \cap A)$$

En particular, para un evento cualquiera E_k como los definidos anteriormente, puede escribirse:

$$P(A)\ P(E_k/A) = P(E_k)\ P(A/E_k)$$

A partir de pasajes de término y reemplazando la expresión de $P(A)$ utilizando probabilidad total, se obtiene:

$$P(E_k/A) = \frac{P(E_k)\ P(A/E_k)}{P(A)} = \frac{P(E_k)\ P(A/E_k)}{\sum_{i=1}^{n} P(E_i) \cdot P(A/E_i)}$$

conocida como *Teorema de Bayes* o teorema de las causas.

Ejemplo 4.10: Dos compañías A y B proveen materiales de construcción de un cierto tipo. La compañía A entrega 600 cargas por día, de las cuales el 3 % no satisface las especificaciones. La compañía B entrega 400 cargas por día, y sólo el 1 % no satisface las especificaciones.

a) ¿Cuál es la probabilidad de que una carga elegida aleatoriamente provenga de A?

b) ¿Cuál es la probabilidad de que una carga elegida aleatoriamente no pase las especificaciones?

c) Si una carga no pasa las especificaciones, ¿cuál es la probabilidad de que provenga de la empresa B?

a) En total se tienen 1000 cargas de las cuales 600 provienen de la compañía A, por lo tanto la probabilidad de que eligiendo una carga aleatoriamente, ésta provenga de la compañía A será:

$$P(A) = 600/1000 = 0{,}6$$

b) Si una carga no pasa las especificaciones, puede provenir tanto de la compañía A como de la B. Utilizando la teoría de la probabilidad total para calcular la probabilidad de N, "que no pase las especificaciones", se tiene:

$$\begin{aligned} P(N) &= P(N/A)\ P(A) + P(N/B)\ P(B) \\ &= 0,03 \cdot 0,60 + 0,01 \cdot 0,40 = 0,018 + 0,004 = 0,022 \end{aligned}$$

c) Para calcular la probabilidad de que si una carga no pasa las especificaciones ésta sea de la compañía B utilizamos el teorema de Bayes:

$$P(B/N) = \frac{P(B)\ P(N/B)}{P(B)\ P(N/B) + P(A)\ P(N/A)} = \frac{0,004}{0,022} = 0,182$$

Este último cálculo también se lo puede realizar si decimos que de la 600 cargas de la empresa A, 18 son defectuosas (3 %) y de las 400 de la empresa B solamente 4 lo son (1 %), entonces la probabilidad de que eligiendo una carga, y esta sea defectuosa, provenga de la compañía B será $4/22 = 0,182$.

4.4. Otras definiciones de probabilidad

Definición frecuencial estadística de la probabilidad

En la definición clásica se requiere conocer cuales son los valores correspondientes tanto a los casos favorables como a los casos posibles (espacio muestral). Sin embargo, a menudo ocurre que alguno de los datos, o ambos, resultan o completamente desconocidos o muy difíciles de conocer. Por ejemplo como calcularíamos:

a) La probabilidad que llueva un día determinado del año.

b) La probabilidad de que el equipo de fútbol A le gane al equipo de fútbol B.

c) La probabilidad que una creciente inunde la ciudad de Resistencia.

Para estos casos, los valores requeridos para aplicar la relación "pascaliana" son desconocidos, por lo que se necesita definir a la probabilidad de otra manera. Para encontrar un procedimiento adecuado, pensemos en el siguiente ejemplo: un recipiente que contiene un conjunto de esferas de colores, donde se desconoce la cantidad total de esferas ni cuántas hay de cada color. Tal como se mencionó anteriormente, el desconocimiento de esos datos hace impracticable el cálculo de probabilidades mediante la definición clásica. Imaginemos ahora el siguiente experimento: extraeremos esferas del recipiente en forma sucesiva, de a una y con reposición, observando en cada extracción el color de la esfera. La cantidad de esferas a extraer, n_i, puede ser determinada a voluntad por quien realiza el experimento, pero se supone que se harán una cantidad de extracciones lo suficientemente numerosa como para que n_i pueda imaginarse tendiendo a infinito. Esto sólo es posible, insistimos, si el experimento se realiza con reposición, ya que de esa manera generamos una población infinita de extracciones.

La experiencia así realizada permite construir una tabla en la cual se podrán volcar los siguientes datos:

- en la primer columna, n_i, el número que corresponde a cada una de las extracciones que se va realizando, partiendo de uno y finalizando en n.
- en la segunda columna se coloca la frecuencia de aparición f_i, es decir el número de esferas de un determinado color que vayan apareciendo, de modo que si en una extracción en particular se presenta el color deseado, f_i aumenta en una unidad mientras que si no se presenta el color buscado, f_i mantiene el valor anterior.
- en la tercer columna se coloca el resultado de efectuar el cociente entre los valores de f_i y n_i, denominada **relación frecuencial** o **estadística**.

Así planteado el experimento, imaginemos que en la primer extracción se presenta el color deseado pero no en la segunda, volviendo a presentarse el color deseado en la tercer extracción. Es decir:

N° de extracciones n_i	frecuencia de aparición f_i	relación frecuencial f_i/n_i
1	1	$1/1 = 1$
2	1	$1/2 = 0{,}5$
3	2	$2/3 = 0{,}66$
...	...	...
n	f	f/n

Es posible representar los resultados obtenidos mediante un gráfico de coordenadas, en el cual en el eje de abscisas se indique el número de extracciones y en ordenadas el valor de la relación frecuencial, figura (4.4).

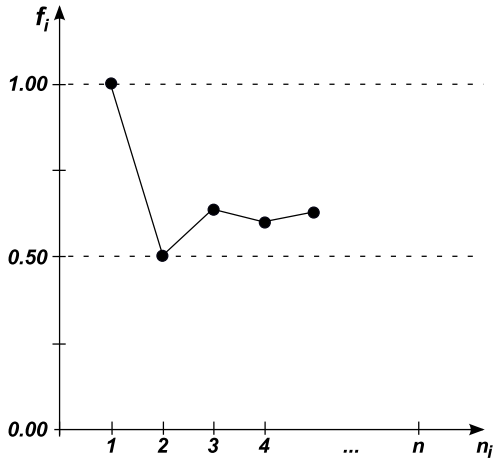


Figura 4.4: Relación frecuencial del experimento

A medida que aumenta el número de extracciones n_i se va ampliando el gráfico, originando un trazo poligonal con variaciones cada vez menos notorias, hasta el punto en que el trazado se acercará al verdadero valor de la probabilidad de ocurrencia del suceso motivo del experimento, a medida que $n_i \to \infty$.

Luego, la probabilidad estimada, o probabilidad empírica, de un suceso se toma como la *frecuencia relativa de ocurrencia* del suceso cuando el número de observaciones es muy grande. Esto quiere decir que la probabilidad de un suceso cualquiera A es el límite de la relación frecuencial f_i/n_i:

$$P(A) = \lim_{n_i \to \infty} \frac{f_i}{n_i} \qquad \text{o mejor aún,} \qquad \frac{f_i}{n_i} \to P(A) \text{ cuando } n_i \to \infty$$

Es decir, la relación frecuencial **converge** al valor de la probabilidad del suceso A cuando el número de realizaciones del experimento crece indefinidamente.

Ejemplo 4.11: ¿Cuál es la probabilidad de que llueva el 21 de septiembre de este año?

Para calcular esta probabilidad no se puede aplicar la definición clásica. Una solución posible para contestar esta pregunta es consultar en los diarios locales de los últimos años (la cantidad de años puede ser definida por el investigador y establecida, por ejemplo en 10,

20, 30 o más años, ya sea según su propio deseo o la disponibilidad del archivo periodístico) que ocurrió el 21 de septiembre. Si en el lapso de 20 años, en cuatro de ellos llovió el día 21 de septiembre, la probabilidad de que llueva en esa fecha puede calcularse haciendo:

$$P(\text{llueva el 21 de septiembre} = \frac{4}{20} = 0,20$$

Definición axiomática de la probabilidad

La definición axiomática de la probabilidad surgió en la década del 30 como consecuencia del aporte de un grupo de matemáticos, quienes opinaban que el concepto de probabilidad no debía estar asociado ni con juegos de azar ni con experiencias estadísticas previas, ya que constituía de por sí un tema con entidad propia y particular. De esa manera, formularon la siguiente serie de tres axiomas a partir de los cuales enunciaron la definición de probabilidad:

1) La probabilidad de ocurrencia de un suceso es un número real, mayor o igual que cero. Es decir que $P(A) \geq 0$.

2) La probabilidad de un sistema completo de sucesos es igual a uno. Si $S = (A_1 \cup A_2 \cup \ldots \cup A_n) \Rightarrow P(S) = 1$

3) Si dos sucesos A y B son mutuamente excluyentes, $P(A \cup B) = P(A) + P(B)$.

A partir de estos axiomas se deducen todas las reglas del campo teórico de la probabilidad.

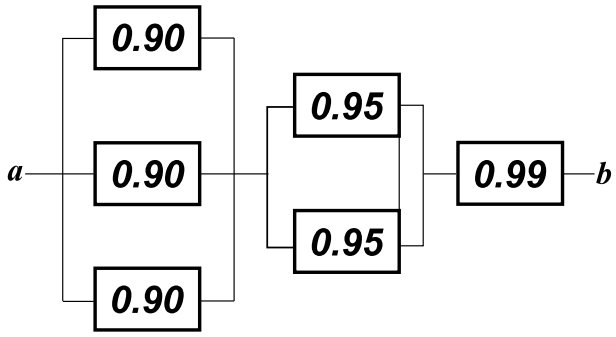
Ejercicio 4.1) Utilice las leyes de De Morgan para demostrar que:

a) $\overline{A \cap (B \cup C)} = (\overline{A} \cup \overline{B}) \cap (\overline{A} \cup \overline{C})$

b) $\overline{(A \cap B \cap C)} = \overline{A} \cup \overline{B} \cup \overline{C}$

En cada caso comprobar los resultados analizando un diagrama de Venn.

Ejercicio 4.2) El circuito siguiente trabaja sólo si existe una trayectoria de dispositivos en funcionamiento, de izquierda a derecha. La probabilidad de que cada dispositivo funcione se indica en la figura. Suponiendo que los dispositivos fallan de manera independiente, ¿cuál es la probabilidad de que el circuito funcione?



Ejercicio 4.3) Si $P(A/B) = 0{,}4$, $P(B) = 0{,}8$ y $P(A) = 0{,}6$, ¿puede decirse que los eventos A y B son independientes?

Capítulo 5
Distribuciones de Probabilidad

5.1. Variables aleatorias

Es posible definir el resultado de cualquier experimento aleatorio de manera genérica como "$\boldsymbol{X}$". De esta manera, al arrojar un dado, X podría tomar los valores 1, 2, 3, 4, 5 o 6; es decir, cualquiera de los puntos muestrales del espacio $\boldsymbol{\Omega}$ correspondiente a ese experimento, pero no puede predecirse cuál será el valor que tomará X antes de realizar la experiencia, en consecuencia X es *aleatoria*. Análogamente, si se desea extraer cinco tornillos de un lote y medir sus diámetros, no puede predecirse cuántos serán defectuosos (no cumplirán ciertos requisitos) de donde, una vez más, $X =$ *"número de defectuosos"* es un valor aleatorio.

También es posible obtener resultados no numéricos de las observaciones, como por ejemplo la condición de una máquina al cabo de 6 meses ($X =$ mala, regular, buena), el color de una esfera extraída al azar ($X =$ azul, rojo, verde, ...); étc. Sin embargo, estos últimos eventos pueden ser identificados numéricamente si se les asignan números a los posibles resultados. En otras palabras, los posibles resultados de un fenómeno aleatorio pueden identificarse numéricamente, ya sea natural o artificialmente.

En forma genérica diremos lo siguiente: si se efectúa un experimento aleatorio y ocurre el evento correspondiente a un número a, entonces se dice que en ese ensayo, la variable aleatoria X ha tomado el valor a, o bien que se ha observado el valor $X = a$. En lugar de "el evento correspondiente a un número a", se dice más brevemente "el evento $X = a$" y la probabilidad asociada se denota por $P(X = a)$. Las distintas formas de expresar la probabilidad de ocurrencia de los eventos se ejemplifican en la tabla siguiente:

X *toma cualquier valor en el intervalo* $a < X < b$	$P(a < X < b)$
$X \leq c$ (X *toma cualquier valor menor o igual a* c)	$P(X \leq c)$
$X > c$ (X *toma cualquier valor mayor que* c)	$P(X > c)$

Variable aleatoria

Una **variable aleatoria** puede considerarse como una **función** que asigna un número real a cada resultado contenido en el espacio muestral de un experimento aleatorio.

En la figura (5.1) los puntos muestrales contenidos en los eventos E_1 y E_2 son transformados desde el espacio muestral Ω a la recta de los números reales mediante la variable aleatoria X. Supongamos que E_1 y E_2 son definidos como:

$$E_1 = (a < X \leq b) \qquad \text{y} \qquad E_2 = (c < X \leq d)$$

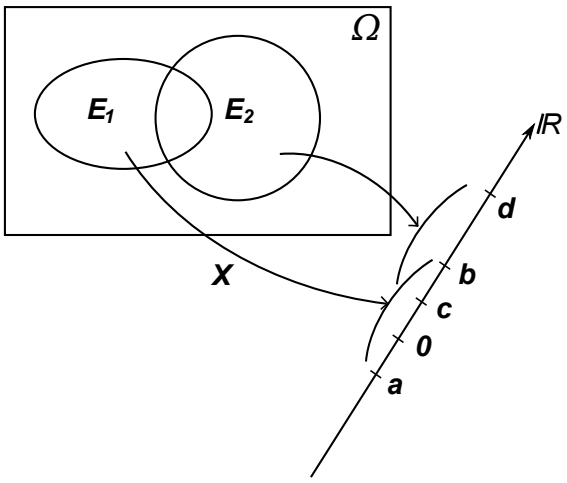


Figura 5.1

Vemos que, por ejemplo, las operaciones $E_1 \cap E_2$ y $\overline{(E_1 \cup E_2)}$ resultan:

$$E_1 \cap E_2 = (c < X \leq b) \qquad \text{y} \qquad \overline{(E_1 \cup E_2)} = (X \leq a) \cup (X > d)$$

Como hemos visto, las variables aleatorias se indican con una letra mayúscula, por ejemplo X, mientras que un valor posible de ella se indica en minúsculas, por ejemplo x, de tal manera que la probabilidad de obtener un resultado particular x queda indicada como $P(X = x)$. Al conjunto de los posibles valores de la variable aleatoria X se lo denomina **rango de X**.

Ejemplo 5.1: El sistema de comunicación de una empresa posee 48 líneas externas. En un determinado momento, se observa el sistema y algunas líneas están ocupadas. Sea X la variable aleatoria que denota el número de líneas en uso. Luego, X puede tomar cualquier valor entero de cero a 48.

Ejemplo 5.2: El análisis de un muestra de aire puede resumirse en el conteo de moléculas raras presentes. Sea X la variable aleatoria que denota el número de moléculas raras en la muestra. Los valores de X son enteros desde cero hasta algún número grande, el cual representa el número máximo de moléculas raras que pueden encontrarse en una de las muestras de aire. Si este número máximo es muy grande, entonces puede suponerse que el rango de X es el conjunto de enteros desde cero hasta infinito.

Las variables aleatorias cuyos posibles valores pueden escribirse como una secuencia finita $x_1, x_2, \ldots, x_N$, o como una secuencia infinita $x_1, x_2, \ldots$, son denominadas **discretas**. Es decir, si la variable aleatoria posee un rango finito o infinito contable se denomina discreta. Tal es el caso de los dos ejemplos anteriores.

Sin embargo, también existen variables aleatorias que pueden tomar valores en un intervalo continuo y se conocen como variables aleatorias **continuas**. Vale decir que si el

rango de una variable aleatoria es un intervalo (finito o infinito) perteneciente al conjunto de números reales, luego la variable aleatoria es continua. Un ejemplo de esto es la variable aleatoria continua que representa la tensión de rotura de una barra de acero, donde se asume que la tensión de rotura puede tomar cualquier valor dentro de un rango (a, b) perteneciente al conjunto de los números reales.

El propósito de identificar a los eventos en términos numéricos radica en que permite una descripción analítica conveniente, así como una descripción gráfica de los eventos y sus probabilidades. La vinculación entre variables aleatorias y probabilidad se realiza mediante lo que se conoce como "distribuciones de probabilidad".

Distribuciones de probabilidad

Variable discreta

Definición

Una distribución de probabilidad de variable discreta está constituida por el conjunto de todos los valores de la variable aleatoria X $(x_1, x_2, \ldots x_N)$ asociados con sus correspondientes probabilidades p_i $(p_1, p_2, \ldots p_N)$, tales que debe cumplirse la siguiente condición, llamada **condición de cierre**: "la suma de las probabilidades a lo largo del campo de variación de la variable aleatoria sea igual a la unidad".

La condición de cierre se traduce en:

$$\sum_{i=1}^{N} p_i = 1$$

Por consiguiente, una distribución de probabilidad constituye un **sistema completo de sucesos**.

Ejemplo 5.3: Dado el experimento "arrojar sucesivamente un dado", determinar:

a) la función de probabilidad y de distribución de probabilidad acumulada de la variable aleatoria X.

b) Calcular la probabilidad de obtener en una tirada *1)* $X \leq 4$; *2)* $X < 3$ o $X \geq 4$; *3)* $X < 3$ y $X \geq 4$; *4)* $X < 5$ y $X \geq 4$

a) si X es el número obtenido al lanzar un dado legal, obtendremos: $p(X = 1) = 1/6$, $p(X = 2) = 1/6$, ..., $p(X = 6) = 1/6$, con lo cual podemos definir la función de probabilidad. Esta función en particular tiene la característica de asignar a cada una de las variables el mismo valor de probabilidad y se conoce como **función de distribución uniforme**.

Por otro lado podemos observar que:

$$P(1 < X < 2) = 0$$
$$P(1 \leq X \leq 2) = P[(X = 1) \cup (X = 2)] = p(X = 1) + p(X = 2) = 1/6 + 1/6 = 1/3$$
$$P(0 \leq X \leq 3,2) = p(X = 0) + p(X = 1) + p(X = 2) + p(X = 3)$$
$$= 1/6 + 1/6 + 1/6 = 1/2$$
$$P(X > 4) = P[(X = 5) \cup (X = 6)] = 1/6 + 1/6 = 1/3$$
$$P(X \leq 0,5) = 0$$

valores que se reflejan en el siguiente gráfico:

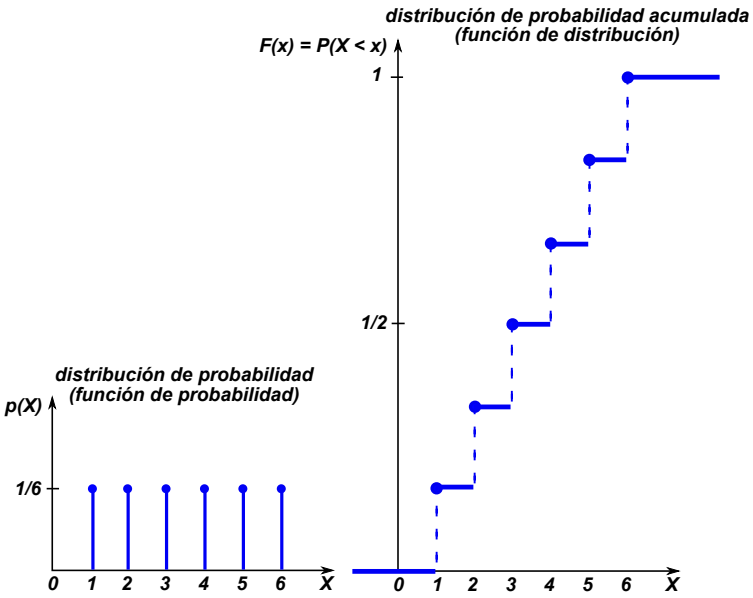


Figura 5.2: Función y distibución de probabilidad.

b.1)

$$P(X \leq 4) = P[(X = 1) \cup (X = 2) \cup (X = 3) \cup (X = 4)]$$
$$= p(X = 1) + p(X = 2) + p(X = 3) + p(X = 4)$$
$$= 1/6 + 1/6 + 1/6 + 1/6 = 2/3$$

También podemos calcular como:

$$P(X \leq 4) = 1 - P(X > 4) = 1 - P[(X = 5) \cup (X = 6)]$$
$$= 1 - (1/6 + 1/6) = 2/3$$

b.2) $X < 3$ o $X \geq 4$

$$P[(X < 3) \cup (X \geq 4)] =$$
$$= [p(X = 1) + p(X = 2)] \cup [p(X = 4) + p(X = 5) + p(X = 6)]$$
$$= 5/6$$

b.3) $X < 3$ y $X \geq 4$

Este suceso no tiene sentido, ya que ambos eventos son mutuamente excluyentes.

b.4) $X < 5$ y $X \geq 4$

$$P[(X < 5) \wedge (X \geq 4)] = p(X = 4) = 1/6$$

Es posible observar que si la variable aleatoria (*VA*) es discreta, su distribución de probabilidad cuenta con las siguientes características:

- La *VA* sólo puede tomar algunos valores dentro de un intervalo definido.
- Las probabilidades se representan con p_i o $p(x)$.
- En un punto cualquiera de la variable X la probabilidad tienen sentido y puede valer p o cero, según ese punto coincida o no con algún valor específico de la variable. Si $X = x_u$, la $p(X = x_u) = p_u$ mientras que $p(X \neq x_u) = 0$ (para X entre x_1 y x_N).
- El gráfico de la distribución de probabilidad se denomina **gráfico de bastones**, por la particular forma que adopta al afectar la probabilidad sólo a algunos valores de X.
- Las probabilidades se calculan mediante la aplicación tanto de las conocidas reglas provenientes de la teoría clásica de la probabilidad como de fórmulas específicas, cuya deducción está reservada a los siguientes capítulos.
- La condición de cierre se verifica probando que $\sum_{i=1}^{N} p_i = 1$.
- La distribución de probabilidad en el caso de una *VA* discreta se denomina genéricamente **función de probabilidad**.

Nótese que la distribución de probabilidad de una *VA* es análoga a la distribución de frecuencias relativas, con valores de probabilidad en lugar de frecuencias relativas. De manera que podemos pensar en las distribuciones de probabilidad como formas teóricas o ideales en el límite, de distribuciones de frecuencia relativa cuando el número de observaciones es muy grande. Por eso podemos asociar a las distribuciones de probabilidad con distribuciones de *poblaciones*, mientras que las distribuciones de frecuencia relativa son distribuciones correspondientes a una *muestra* de esa población.

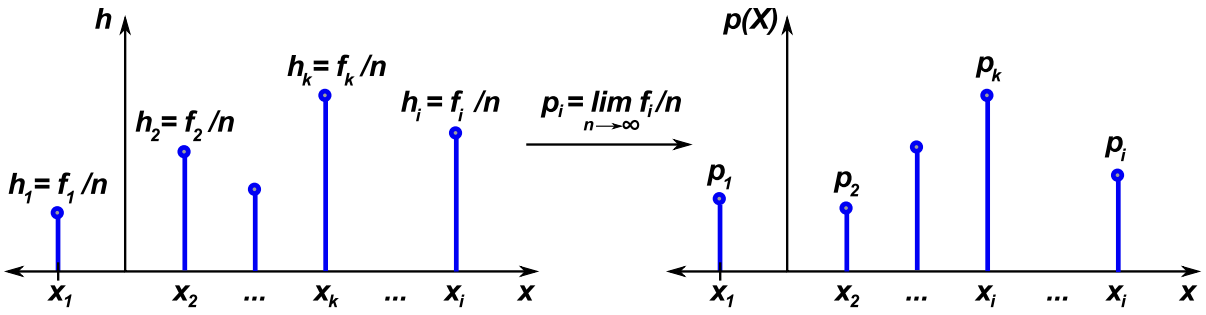


Figura 5.3: Relación de la distribución de frecuencias relativas con la función de probabilidad (VA discreta).

Acumulando probabilidades se obtiene la **función de distribución de probabilidad acumulada** o simplemente **función de distribución**, análoga a la distribución de frecuencia relativa acumulada. Esta función representa la probabilidad de que la variable aleatoria X sea menor o igual que un valor particular x_u. Luego,

$$F(x_u) = P(X \leq x_u) \tag{5.1}$$

Para VA discretas, la función de distribución se obtiene sumando sucesivamente las probabilidades correspondientes a los valores de la VA, a partir del primer valor hasta el u-ésimo:

$$F(x_u) = P(X \leq x_u) = \sum_{i=1}^{u} p_i = p_1 + p_2 + \ldots + p_u$$

Variable continua

Las ideas anteriores se extienden a variables que pueden tomar un *conjunto continuo* de valores. Dada una variable continua x, si dividimos su rango en intervalos discretos Δx, denominados intervalos de clase, contabilizando las observaciones u ocurrencias que caen en cada uno de ellos y finalmente el resultado se grafica en forma de diagrama de barras, obtenemos lo que anteriormente denominamos histograma de frecuencias absolutas. Si ahora al número de ocurrencias f_i en el intervalo i, que cubre el rango $[x_i - \Delta x, x_i]$, lo dividimos por el número total de observaciones n, el resultado lo conocemos como frecuencia relativa h_i:

$$h_i = \frac{f_i}{n} \tag{5.2}$$

La cual es una estimación de $P(x_i - \Delta x \leq X \leq x_i)$, la probabilidad de que la variable aleatoria X se encuentre en el intervalo $[x_i - \Delta x, x_i]$. La función correspondiente a la población se aproxima como el límite a medida que $n \to \infty$ y $\Delta x \to 0$. Es decir, en el límite, la ecuación (5.2) dividida por el intervalo de longitud Δx se convierte en la **función de densidad de probabilidad** $f(x)$:

$$f(x) = \lim_{\substack{n \to \infty \\ \Delta x \to 0}} \frac{f_i}{\Delta x \; n} \tag{5.3}$$

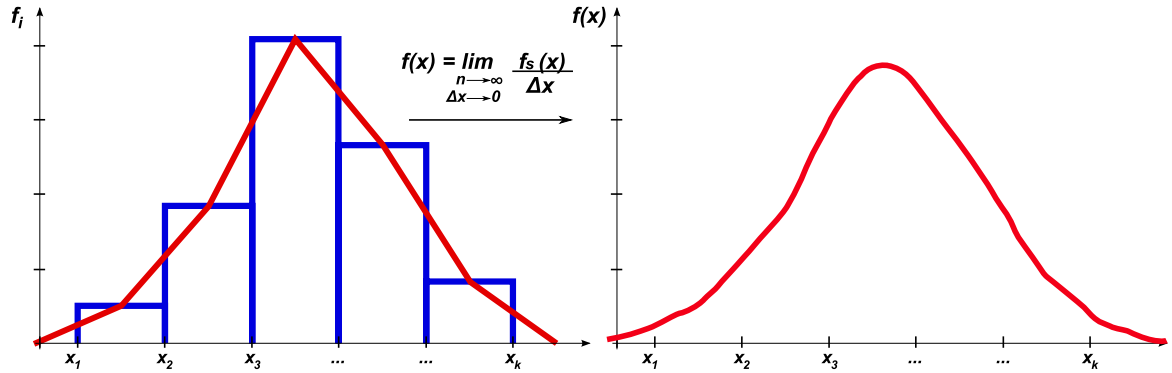


Figura 5.4: Relación de la distribución de frecuencias relativas con la función de probabilidad (VA continua).

La suma de los valores de las frecuencias relativas hasta un punto dado es la frecuencia relativa acumulada H_i:

$$H_i = \sum_{j=1}^{i} h_j$$

la cual es una estimación de $P(X \leq x_i)$, la probabilidad acumulada de x_i. La función de frecuencia acumulada se convierte en la función de distribución de probabilidad acumulada $F(x)$ haciendo:

$$F(x) = \lim_{\substack{n\to\infty \\ \Delta x\to 0}} \frac{H_i}{\Delta x} \tag{5.4}$$

Para un valor dado de la variable continua X, $F(x)$ es la probabilidad acumulada $P(X \leq x)$, y se expresa como la integral de la función de densidad de probabilidad sobre el rango $X \leq x$:

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)du \tag{5.5}$$

donde u es una variable de integración auxiliar. Es posible observar que la derivada de esta función es la función de densidad de probabilidad:

$$f(x) = \frac{dF(x)}{dx} \tag{5.6}$$

Gráficamente hemos visto que el polígono de frecuencias relativas de una muestra se convierte, en el caso teórico o límite de una población, en una curva continua. El área total entre la curva y el eje de las abscisas es uno (condición de cierre), y el área entre x_a y x_b, sombreada en la figura (5.5), indica la probabilidad de que X se encuentre entre x_a y x_b, que se denota por $P(x_a \leq X \leq x_b)$.

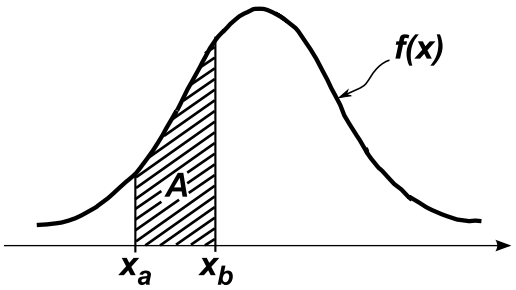


Figura 5.5: Representación gráfica de la probabilidad para variables continuas.

Definición

Dada una variable aleatoria X continua, una función cualquiera $f : \Re \to \Re$ es una distribución de probabilidad tal que

$$P(a \leq x \leq b) = \int_a^b f(x)\ dx$$

si f satisface que $f(x) > 0 \quad \forall\ x$ y además

$$\int_{-\infty}^{+\infty} f(x)\ dx = 1$$

La función de densidad *no tiene sentido puntualmente*, pero sí en un intervalo particular de la variable. Es decir, si el intervalo se torna cada vez más pequeño, la probablidad tenderá a cero. Esto se demuestra fácilmente considerando un número positivo cualquiera ε:

$$P(a - \varepsilon \leq x \leq a + \varepsilon) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\ dx$$

con lo cual,

$$\lim_{\varepsilon \to 0} P(a - \varepsilon \leq x \leq a + \varepsilon) = P(X = a) = 0 \quad \forall\ a$$

Vemos entonces que para una VA continua la función de distribución presenta las siguientes características:

- La *VA* puede tomar cualquier valor en un determinado campo de variación.
- La densidad de probabilidad se representa simbólicamente como $f(x)$.
- En un punto la probabilidad no tiene sentido, o sea $P(X = x_a)$ no es posible de calcular. Sólo tiene sentido en un intervalo particular de la variable aleatoria X, por más pequeño que sea éste. Es decir que, simbólicamente, podemos indicar a la probabilidad con la expresión $P(x_a \leq X \leq x_b) = A$.
- Gráficamente, la distribución de probabilidad se representa como una función continua $f(x)$ y la probabilidad en sí misma, denominada A, se representa como el área entre los puntos x_a y x_b.

- La probabilidad se obtiene calculando la integral definida, según el criterio de Riemann, de la función $f(x)$, entre los puntos x_a y x_b. Es que

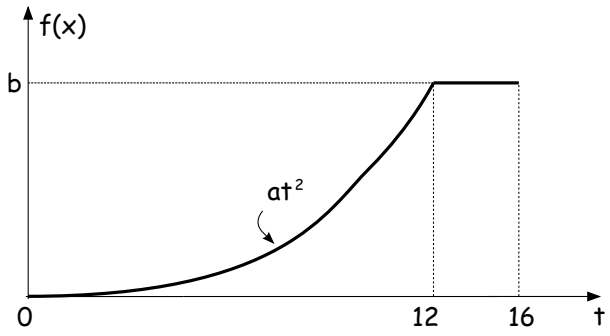
$$P(x_a \leq X \leq x_b) = \int_{x_a}^{x_b} f(x)dx = A$$

- La condición de cierre se verifica efectuando la integral de la función en todo el campo de variación de la VA, es decir,

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

- La denominación genérica de la distribución de probabilidad en el caso continuo es "función de densidad".

Ejemplo 5.4. La duración de la acción de una fuerza sobre una estructura es una variable aleatoria continua cuya densidad de probabilidad está representada en la figura:



a) Determinar los valores apropiados de a y b.

b) Calcular la probabilidad de que t sea mayor o igual a 6 segundos.

En el cuadro siguiente se resumen las características de las distribuciones de probabilidad para *VA* discretas y continuas.

Tipo de variable	Discreta	Continua
Simbología de la probabilidad	p_i o $p(x)$	f_i o $f(x)$
Concepto de probabilidad	puntual	en intervalo (en un punto no tiene sentido)
Valor de la probabilidad	vale p_i (bastón) o cero	vale A (área)
Gráfico	de bastones	de áreas
Cálculo de la probabilidad	$P(X = x_u) = p_u$ $P(X \neq x_u) = 0$	$P(x_a \leq X \leq x_b) = \int_{x_a}^{x_b} f(x)dx = A$
Condición de cierre	$\sum_{i=1}^{n} p_i = 1$	$\int_{-\infty}^{+\infty} f(x)dx = 1$
Denominación genérica	Función de probabilidad	Función de densidad

Esperanza Matemática de la variable aleatoria

La esperanza matemática de la variable aleatoria, que se define de una manera técnica, es decir, mediante su fórmula de cálculo, se obtiene efectuando la sumatoria de los productos de los valores de las variables por sus respectivas probabilidades, siempre que se cumpla la condición de cierre. La expresión de la fórmula de la esperanza matemática se presenta más abajo, según se trate de una variable aleatoria discreta o continua:

$$\begin{array}{cc} \text{Caso discreto } \textit{(a)} & \text{Caso continuo } \textit{(b)} \\ E(X) = \sum_{i=1}^{N} x_i \cdot p_i & E(X) = \int_{-\infty}^{\infty} x\ f(x)\ dx \\ (\text{si se cumple } \sum p_i = 1) & (\text{si se cumple } \int_{-\infty}^{\infty} f(x)\ dx = 1) \end{array} \tag{5.7}$$

Si las probabilidades p_i en la expresión (a) se sustituyen por las frecuencias relativas f_i/n, la esperanza matemática se reduce a $\frac{\sum f_i \cdot x_i}{n}$, que es la media aritmética $\overline{x}$ de una muestra de tamaño n en la que $x_1, x_2, \ldots, x_k$ aparecen con estas frecuencias relativas. Al crecer n las frecuencias relativas se acercan a las probabilidades p_i. De esta manera interpretamos $E(X)$ como la media de la población cuyo muestreo se consideraba, y se indica con la letra griega μ_x (o simplemente μ). La esperanza matemática también se denomina **media poblacional** mientras que la media aritmética $\overline{x}$ es una media muestral. Esta última se diferencia de la primera en que es empírica y real. Otra forma adecuada de expresar a la esperanza matemática, es señalarla con la expresión **número esperado**, lo cual en algunos casos suele resultar más sencillo y comprensible.

Variancia de la variable aleatoria

La variancia de la variable aleatoria constituye un concepto similar al de variancia definido en capítulos anteriores, ya que se trata de una medida de dispersión, cuya fórmula,

a similitud de aquella, es:

$$\begin{array}{cc} \text{Caso discreto } (a) & \text{Caso continuo } (b) \\ V(X) = \sum_{i=1}^{N} \left[x_i - E(X)\right]^2 \cdot p_i & V(X) = \int_{-\infty}^{\infty} \left[x - E(X)\right]^2 f(x)\ dx \end{array} \tag{5.8}$$

Puede resultar útil la fórmula de trabajo de la variancia para el caso discreto:

$$V(X) = \sum_{i=1}^{N} x_i^2 \cdot p_i - E(X)^2 \tag{5.9}$$

Así como la esperanza matemática se considera similar a la media poblacional, la variancia de la variable aleatoria se considera conceptualmente una variancia poblacional, y se simboliza del siguiente modo:

$$V(X) = \sigma_x^2$$

Pudiéndose obtener el desvío estándar mediante la aplicación de la raíz cuadrada en las en las expresiones indicadas precedentemente, es decir

$$DS(X) = \sigma_x$$

Ejercicio 5.1) Considerando la función densidad del ejemplo 5.4, calcular el valor esperado y la mediana de la variable t.

5.2. Distribuciones de probabilidad para variables aleatorias discretas

Distribución Binomial

La Distribución Binomial, estudiada originalmente por el matemático Bernoulli, por lo cual también se la conoce como la Distribución o Fórmula de Bernoulli, surge a partir de las siguientes condiciones:

1) se realiza un experimento aleatorio que sólo puede dar lugar a la ocurrencia de dos sucesos posibles: el suceso A o su opuesto $\overline{A}$ (se denominan resultados *dicotómicos*).

2) el experimento aleatorio se realiza n veces, manteniendo la independencia entre las distintas realizaciones.

3) en cualquier realización del experimento, la probabilidad de A es igual a p y, por consiguiente la probabilidad de $\overline{A}$ es $(1 - p) = q$. En función de la condición *2)*, las probabilidades p y q se mantienen constantes:

$$P(A) = p \qquad \text{y} \qquad P(\overline{A}) = 1 - p = q$$

Objetivo

Mediante el planteo de esta distribución es posible calcular **cuál es la probabilidad de que, en las n realizaciones del experimento, se presenten x veces el suceso A.**

Deducción de la fórmula de la Distribución Binomial

Si el experimento aleatorio se realiza n veces de acuerdo a las condiciones indicadas anteriormente y en x ocasiones se presenta el suceso A, es evidente que en las restantes $(n-x)$ ocasiones se presentará el suceso opuesto $\overline{A}$. A un conjunto de esas n realizaciones del experimento lo denominaremos **secuencia**. La secuencia i-ésima se simbolizará con S_i.

Supongamos ahora que la primer secuencia (S_1) sea aquella en la que, realizadas las n observaciones del experimento, se presenta en primer lugar x veces A y, a continuación, los $(n-x)$ sucesos $\overline{A}$. Por consiguiente, esa secuencia S_1 será:

$$S_1 = A_1 \cap A_2 \cap \ldots \cap A_x \cap \overline{A}_{x+1} \cap \ldots \cap \overline{A}_n$$

por lo que, aplicando el concepto de probabilidad, tendremos que

$$P(S_1) = P(A_1 \cap A_2 \cap \ldots \cap A_x \cap \overline{A}_{x+1} \cap \ldots \cap \overline{A}_n)$$

Como las n realizaciones del experimento son independientes, el cálculo de esa probabilidad se resuelve mediante la aplicación de la regla de la multiplicación para ese tipo de sucesos, dando como resultado:

$$P(S_1) = P(A_1)P(A_2)\ldots P(A_x)P(\overline{A}_{x+1})\ldots P(\overline{A}_n)$$

y recordando que $P(A) = p$ y $P(\overline{A}) = q$:

$$P(S_1) = \overbrace{\underbrace{p\ p \ldots p}_{x}\ \underbrace{q \ldots q}_{(n-x)}}^{n} = p^x\ q^{n-x} \tag{5.10}$$

con lo cual se obtiene la probabilidad de ocurrencia de la secuencia S_1.

Cualquier otra secuencia S_i (para $i \neq 1$) estará también compuesta por x sucesos $\overline{A}$ pero con un orden diferente de aparición. Por ejemplo, denominemos con S_2 a la secuencia que modifica la posición de los sucesos A_x y $\overline{A}_{x+1}$, de modo tal que la nueva secuencia tendrá la siguiente disposición:

$$S_2 = A_1 \cap A_2 \cap \ldots \cap \overline{A}_x \cap A_{x+1} \cap \overline{A}_{x+2} \cap \ldots \cap \overline{A}_n$$

Mediante un procedimiento similar al aplicado en el caso de la secuencia S_1, la probabilidad de la secuencia será:

$$\begin{aligned} P(S_2) &= P(A_1 \cap A_2 \cap \ldots \cap \overline{A}_x \cap A_{x+1} \cap \overline{A}_{x+2} \cap \ldots \cap \overline{A}_n) \\ &= P(A_1)P(A_2)\ldots P(\overline{A}_x)P(A_{x+1})P(\overline{A}_{x+2})\ldots P(\overline{A}_n) \end{aligned}$$

luego,

$$P(S_2) = p\ p\ \ldots q\ p\ \ldots q = p^x\ q^{n-x} \tag{5.11}$$

La expresión de la probabilidad de ocurrencia de la secuencia S_2 contiene, como puede verse si se la estudia con cierto cuidado, x factores iguales a p y $(n-x)$ factores iguales a q, lo que en definitiva da como resultado $p^x\ q^{(n-x)}$, igual a la probabilidad hallada para el caso de la secuencia S_1. Esto se verifica para cualquier secuencia S_i dado que todas tienen idéntica composición. Es decir, la probabilidad de cualquier secuencia S_i es igual a $p^x\ q^{(n-x)}$.

Ya se ha mencionado que cualquier secuencia S_i está compuesta por x repeticiones del suceso A y $(n-x)$ repeticiones del suceso $\overline{A}$, y que existen varias secuencias posibles (de las cuales S_1 y S_2 son sólo un par de ejemplos). Ahora indicaremos con j al total de las secuencias posibles diferentes entre sí, las que a su vez son excluyentes dado que la presentación de una secuencia cualquiera de ellas excluye a las restantes.

Por consiguiente, la probabilidad buscada puede enunciarse del siguiente modo:

$$\begin{aligned} P(\text{en } n \text{ realizaciones se presente } x \text{ veces } A) &= P(S_1 \cup S_2 \cup \ldots \cup S_j) \\ &= P(S_1) + P(S_2) + \ldots + P(S_j) \end{aligned}$$

y recordando el valor dado por las ecuaciones (5.10) y (5.11) y el cual se mantiene constante para todas las secuencias posibles (las secuencias son equiprobables), por la aplicación de la regla de la adición para sucesos excluyentes obtenemos:

$$\begin{aligned} P(\text{en } n \text{ realizaciones se presente } x \text{ veces } A) &= p^x\ q^{n-x} + p^x\ q^{n-x} + \cdots + p^x\ q^{n-x} \\ &= j\ p^x\ q^{n-x} \end{aligned} \tag{5.12}$$

Queda ahora por definir cuál es el valor de j, para lo cual recordaremos que ese valor se obtiene calculando las permutaciones con repetición de n elementos, entre los cuales hay un conjunto de x repetidos y $(n-x)$ repetidos, que se resuelve mediante la fórmula de combinaciones simples de n elementos tomados de a x, es decir:

$$j = \binom{n}{x} = \frac{n!}{x!\ (n-x)!} \tag{5.13}$$

Por consiguiente, la fórmula final de la probabilidad binomial se obtiene reemplazando adecuadamente la ecuación (5.13) en la ecuación (5.12):

$$P(\text{en } n \text{ realizaciones se presente } x \text{ veces } A) = \binom{n}{x}\ p^x\ q^{n-x} \tag{5.14}$$

Ejemplo 5.5. Se arrojan al aire tres monedas. Hallar la probabilidad de que se presenten dos caras.

Como se cumplen las condiciones exigidas para la aplicación del esquema binomial, ya que las realizaciones del experimento son independientes, el cálculo de la probabilidad buscada se realiza determinando los siguientes valores:

n : número de realizaciones del experimento = 3.

x : número de presentaciones del suceso A (cara) = 2.

p : probabilidad del suceso $A = 1/2$.

q : probabilidad del suceso $\overline{A} = 1/2$.

Luego,

$$P(\text{en 3 tiradas aparezcan 3 caras}) = \binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1 = 3\ \frac{1}{4}\ \frac{1}{2} = \frac{3}{8}$$

Condición de cierre en la Distribución Binomial

Según lo enunciado anteriormente, la condición de cierre establece que toda función de probabilidad debe verificar que $\sum_{i=1}^{N} p_i = 1$. Por consiguiente, el cumplimiento de esa condición en el caso de la Distribución Binomial se demuestra sumando las probabilidades a lo largo de todo el campo de variación de la variable, es decir, haciendo $\sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x}$.

Desarrollando esta última expresión entre 0 y n se obtiene:

$$\sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = \binom{n}{0} p^0 q^n + \binom{n}{1} p^1 q^{n-1} + \ldots + \binom{n}{n} p^n q^0$$

que es equivalente a la fórmula de Newton para el desarrollo de un binomio, es decir,

$$\sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (p+q)^n$$

Ahora, de acuerdo con las condiciones de la Distribución Binomial, como $(p+q) = 1$ también será $(p+q)^n = 1$, verificando así la condición de cierre de la distribución.

Asimismo queda en evidencia que el nombre *Binomial* proviene del hecho de que cada término del binomio desarrollado precedentemente dá como resultado la probabilidad para los diferentes valores de la variable aleatoria X.

Ejemplo 5.6. Verificar el cumplmiento de la condición de cierre para el caso planteado en el ejemplo 15.

N° de caras (x)	Probabilidad p(x)	Resultado numérico
0	$p(x=0)=\binom{3}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^3$	1/8
1	$p(x=1)=\binom{3}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^2$	3/8
2	$p(x=2)=\binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1$	3/8
3	$p(x=3)=\binom{3}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0$	1/8
		$\sum p_i = 1$

Parámetros de la Distribución Binomial - Esperanza matemática

De acuerdo a lo visto anteriormente, la Esperanza Matemática se calcula a partir de la ecuación $E(x)=\sum x_i p_i$. En el caso de la Distribución Binomial debe utilizarse la expresión matemática que le corresponde:

$$E(x)=\sum_{x=0}^{n} x\binom{n}{x}p^x q^{n-x}=0\binom{n}{0}p^0q^n+\sum_{x=1}^{n} x\binom{n}{x}p^x q^{n-x}=\sum_{x=1}^{n} x\binom{n}{x}p^x q^{n-x}$$

Haciendo ahora $n!=n(n-1)!$; $x!=x(x-1)!$ y verificando que $(n-x)=(n-1)-(x-1)$, es posible escribir

$$E(x)=\sum_{x=1}^{n} x\ \frac{n(n-1)!}{x(x-1)!\ [(n-1)-(x-1)]!}\ p\ p^{x-1}q^{(n-1)-(x-1)}$$

del cual simplificamos x en numerador y denominador y dado que tanto n como p son constantes respecto de la sumatoria, es posible extraerlos de la misma. Luego,

$$E(x)=n\ p\sum_{x=1}^{n}\frac{(n-1)!}{(x-1)!\ [(n-1)-(x-1)]!}\ p^{x-1}q^{(n-1)-(x-1)}$$

Realizando la siguiente sustitución: $N=n-1$, $X=x-1$ y observando que si $x=1\Rightarrow X=0$ y que para $x=n\Rightarrow X=N$,

$$E(x)=n\ p\sum_{X=0}^{N}\frac{N!}{X!\ (N-X)!}\ p^X q^{N-X}=n\ p$$

dado que la sumatoria en la última ecuación es igual a la unidad por la condición de cierre de la distribución binomial. Por lo tanto, queda demostrado que

$$E(x) = n\ p$$

Ejemplo 5.7. Se arrojan al aire cuatro monedas. Encontrar:

a) la distribución de probabilidad del "número de caras que pueden presentarse".

b) la esperanza matemática de la variable aleatoria, es decir, el número esperado de caras que pueden presentarse cuando se arrojan cuatro monedas.

a) En el siguiente cuadro se presenta la distribución de probabilidad requerida:

N° de caras (x)	Probabilidad $p(x)$	Resultado numérico	$x_i\ p_i$
0	$p(x=0) = \binom{4}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4$	1/16	0
1	$p(x=1) = \binom{4}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3$	4/16	4/16
2	$p(x=2) = \binom{4}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2$	6/16	12/16
3	$p(x=3) = \binom{4}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^1$	4/16	2/16
4	$p(x=4) = \binom{4}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0$	1/16	4/16
		$\sum p_i = 1$	$E(x) = 32/16 = 2$

b) En la cuarta columna del cuadro precedente se efectuó el producto $x_i\ p_i$, valores con cuya sumatoria se obtiene la esperanza matemática. Sin embargo, de acuerdo a lo demostrado previamente, la esperanza matemática también puede calcularse efectuando el producto np:

$$E(x) = n\ p = 4\ \frac{1}{2} = 2$$

Parámetros de la Distribución Binomial - Variancia

Recordando que la variancia de la variable aleatoria se obtiene mediante la fórmula de trabajo $V(x) = \sum_{i=1}^{N} x_i^2 p_i - E(x)^2$:

$$V(x) = \sum_{x=0}^{n} \left[x^2 \binom{n}{x} p^x q^{n-x} \right] - (np)^2 = \sum_{x=0}^{n} \left[x^2 \frac{n!}{x!\ (n-x)!} p^x q^{n-x} \right] - (np)^2$$

en la cual, haciendo $x^2 = x(x-1) + x$:

$$V(x) = \sum_{x=0}^{n} \left\{ [x(x-1) + x]\ \frac{n!}{x!\ (n-x)!} p^x q^{n-x} \right\} - (np)^2$$

Distribuyendo los términos entre corchetes,

$$V(x) = \sum_{x=0}^{n} \left[x(x-1)\ \frac{n!}{x!\ (n-x)!} p^x q^{n-x} \right] + \sum_{x=0}^{n} \left[x\ \frac{n!}{x!\ (n-x)!} p^x q^{n-x} \right] - (np)^2$$

donde se verifica que la segunda sumatoria es la esperanza matemática de la distribución binomial (np) y que la primer sumatoria puede desarrollarse tomando para la variable x los valores 0 y 1,

$$V(x) = \left[0\ (-1)\ \frac{n!}{0!\ n!} p^0\ q^n - 0\ \frac{n!}{1!\ n!} p^1\ q^{n-1} + \sum_{x=2}^{n} x(x-1)\ \frac{n!}{x!\ (n-x)!} p^x q^{n-x} \right] + np - (np)^2$$

Luego,

$$V(x) = \left[\sum_{x=2}^{n} x(x-1)\ \frac{n!}{x!\ (n-x)!} p^x\ q^{n-x} \right] + np - (np)^2 \tag{5.15}$$

Considerando ahora las siguientes sustituciones:

$$\begin{aligned} n! &= n(n-1)(n-2)! \\ x! &= x(x-1)(x-2)! \\ p^x &= p^2 p^{x-2} \\ (n-x) &= (n-2) - (x-2) \end{aligned}$$

y reemplazando en la ecuación (5.15):

$$V(x) = \left[\sum_{x=2}^{n} x(x-1)\ \frac{n(n-1)(n-2)!}{x(x-1)(x-2)!\ [(n-2)-(x-2)]!}\ p^2 p^{x-2}\ q^{(n-2)-(x-2)} \right] + np - (np)^2$$

Simplificando los términos $x(x-1)$ del numerador y denominador y considerando que tanto $n(n-1)$ como p^2 son constantes:

$$V(x) = n(n-1)\ p^2 \left[\sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!\ [(n-2)-(x-2)]!}\ p^{x-2}\ q^{(n-2)-(x-2)} \right] + np - (np)^2$$

Haciendo ahora la siguiente sustitución:

$$\begin{aligned} &(n-2) = N \\ &(x-2) = X \Rightarrow \text{ si } x = 2, X = 0 \text{ y si } x = n, X = N \end{aligned}$$

$$V(x) = n(n-1)\ p^2 \Big[\underbrace{\sum_{X=0}^{N} \frac{N!}{X!\ (N-X)!}\ p^X\ q^{N-X}}_{=1} \Big] + np - (np)^2$$

donde se utilizó la condición de cierre para el término entre corchetes. Luego,

$$\begin{aligned} V(x) &= n(n-1)\ p^2 + np - (np)^2 = n^2\ p^2 - n\ p^2 + np - n^2\ p^2 \\ &= np - n\ p^2 = np(1-p) = npq \end{aligned}$$

Con lo cual,

$$V(x) = npq \tag{5.16}$$

y

$$DS(x) = \sqrt{npq} \tag{5.17}$$

Ejemplo 5.8: Dado que no todos los pasajeros de una aerolínea abordan el vuelo para el que han reservado un lugar, la aerolínea vende 125 boletos para un vuelo con capacidad de 120 pasajeros. Asumiendo que la probabilidad de que un pasajero no aborde el vuelo (por no presentarse o por cambiar la fecha de viaje) es de 0.10 y que el comportamiento de los pasajeros es independiente,

a) ¿cuál es la probabilidad de que el día del vuelo cada pasajero que se presente pueda abordarlo?

b) cuál es la probabilidad de que el vuelo parta con asientos vacíos (no necesariamente todos vacíos).

Distribución de Poisson

Algunas de las condiciones requeridas para la aplicación de la Distribución de Poisson son similares a las exigidas para la implementación de la distribución binomial:

1) en cada una de las realizaciones del experimento aleatorio se presentan resultados dicotómicos.

2) el experimento aleatorio se realiza n veces en condiciones de independencia, por lo cual los posibles resultados mantienen sus probabilidades constantes a lo largo de las realizaciones.

3) Las n realizaciones del experimento crecen notoriamente, lo cual equivale a decir que $n \to \infty$.

4) La probabilidad p de ocurrencia del suceso A es notoriamente pequeña, es decir que $p \to 0$. Esta condición en particular determina que se denomine a esta distribución "distribución de sucesos raros", debido a que la probabilidad del suceso A es muy pequeña.

5) Las realizaciones del experimento se cumplen en un intervalo de tiempo (o espacio) continuo y no, como en el caso de la binomial, en momentos fijos o determinados.

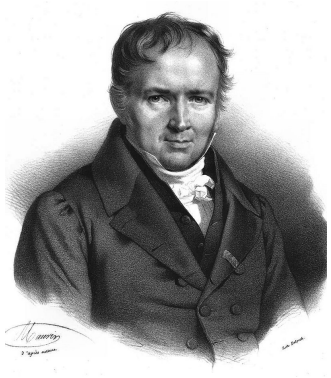


Figura 5.6: Siméon Denis Poisson (1781-1840).

Deducción de la fórmula de la Distribución de Poisson

El análisis de las condiciones requeridas para la aplicación de la distribución de Poisson demuestran que guardan cierta similitud con las exigidas para la aplicación de la distribución binomial. Pero cuando se analizan las condiciones 3) y 4), la aplicación de la fórmula binomial se vuelve sumamente laboriosa, lo cual se verifica tan sólo si se piensa en calcular el número combinatorio $\binom{n}{x}$ para $n \to \infty$ y además, los resultados obtenidos serían sumamente imprecisos.

A raíz de esta circunstancia es que Poisson obtuvo la expresión de la fórmula que lleva su nombre, haciendo en primer lugar las siguientes transformaciones:

- $\lambda = np \Rightarrow p = \dfrac{\lambda}{n}$

- luego, $q = 1 - p = 1 - \dfrac{\lambda}{n}$

Luego, aplicando el límite a la distribución binomial para $n \to \infty$, se tiene

$$\lim_{n\to\infty} \binom{n}{x}\, p^x\, q^{n-x} = \lim_{n\to\infty} \frac{n!}{x!\,(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} =$$
$$= \lim_{n\to\infty} \frac{n(n-1)(n-2)\ldots(n-x+1)(n-x)!}{(n-x)!}\, \frac{1}{n^x} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

donde se verifica que al desarrollar $n!$ hasta $(n-x)!$ es posible simplificar el término $(n-x)!$ en el numerador y en el denominador de la expresión, con lo cual en el numerador quedarán x términos, desde n hasta $(n-x+1)$, que pueden ser divididos por los x términos correspondientes a n^x. Es decir que la expresión anterior puede ser escrita de la siguiente manera:

$$= \lim_{n\to\infty} \frac{\overbrace{n(n-1)(n-2)\ldots(n-x+1)}^{x \text{ términos}}}{\underbrace{n^x}_{x \text{ términos}}} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

Dividiendo cada uno de los x términos del numerador por los x términos iguales a n del denominador,

$$= \lim_{n\to\infty} \frac{n}{n} \frac{(n-1)}{n} \cdots \frac{(n-x+1)}{n} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$
$$= \lim_{n\to\infty} 1 \left(1 - \frac{1}{n}\right) \ldots \left[1 - \frac{(x-1)}{n}\right] \left(1 - \frac{\lambda}{n}\right)^{-x} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n$$

expresión esta en la cual, al aplicar límite para $n \to \infty$, los términos encerrados entre paréntesis desde el primero hasta $\left(1 - \dfrac{\lambda}{n}\right)^{-x}$ se igualan a la unidad, mientras que $\dfrac{\lambda^x}{x!}$ está compuesto por factores constantes, de modo que es posible escribir

$$\begin{aligned} \lim_{n\to\infty} \binom{n}{x}\, p^x\, q^{n-x} &= \frac{\lambda^x}{x!} \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n \\ &= \frac{\lambda^x}{x!}\, e^{-\lambda} \end{aligned} \tag{5.18}$$

que resulta ser la fórmula de la distribución de Poisson.

Ejemplo 5.9. La probabilidad de que ocurra una falla diaria en el equipo de una empresa es de 0,02. Suponiendo que el equipo funciona durante 450 días, ¿cuál es la probabilidad de que ocurran fallas:

a) en exactamente 5 días diferentes?

b) en un día como mínimo?

a) $\lambda = np = 450 \cdot 0{,}02 = 9$

$$P(\text{fallas en 5 días}) = \frac{e^{-9}\ 9^5}{5!} = \frac{e^{-9}\ 9^5}{5!} = \frac{0{,}0001234 \cdot 59049}{120} = 0{,}06$$

b) $P(\text{fallas en un día como mínimo}) = 1 - P(\text{fallas en ningún día})$

$$P(\text{fallas en un día como mínimo}) = 1 - \frac{e^{-9}\ 9^0}{0!} = 1 - 0{,}0001234 = 0{,}9998$$

Condición de cierre en la Distribución de Poisson

Para demostrar que se cumple la condición de cierre se aplica a la distribución de poisson la sumatoria a lo largo de todo el campo de variación de la variable aleatoria:

$$\sum_{x=0}^{\infty} \frac{\lambda^x}{x!}\ e^{-\lambda} = e^{-\lambda}\left[\overbrace{\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots}^{\text{desarrollo en serie de } e^{\lambda}}\right]$$
$$= e^{-\lambda}\ e^{\lambda} = 1$$

Parámetros de la Distribución de Poisson - Esperanza matemática

$$E(x) = \sum_{x=0}^{\infty} x\ p(x) = 0 + \sum_{x=1}^{\infty} x\ p(x) = \sum_{x=1}^{\infty} x\ \frac{e^{-\lambda}\ \lambda^x}{x!} = \sum_{x=1}^{\infty} x\ \frac{e^{-\lambda}\ \lambda\ \lambda^{x-1}}{x\ (x-1)!}$$

extrayendo λ fuera de la sumatoria y simplificando x en el numerador y denominador, se obtiene la siguiente expresión:

$$E(x) = \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\ \lambda^{x-1}}{(x-1)!}$$

haciendo ahora $X = (x-1)$, y verificando que para $x = 1$ se obtiene $X = 0$ así como para $x \to \infty$ tenemos $X \to \infty$,

$$E(x) = \lambda \underbrace{\sum_{X=0}^{\infty} \frac{e^{-\lambda}\ \lambda^X}{X!}}_{=1\ \text{por condición de cierre}} = \lambda \tag{5.19}$$

Se demuestra entonces que la esperanza matemática en la Distribución de Poisson es igual a λ.

Parámetros de la Distribución de Poisson - Variancia

$$V(x) = \sum_{x=0}^{\infty} x^2 \; p(x) - E(x)^2 = \sum_{x=0}^{\infty} x^2 \; \frac{e^{-\lambda} \; \lambda^x}{x!} - \lambda^2$$

reemplazando ahora x^2 por $x(x-1)+x$:

$$V(x) = \sum_{x=0}^{\infty} \big[x(x-1)+x\big] \; \frac{e^{-\lambda} \; \lambda^x}{x!} - \lambda^2 = \sum_{x=0}^{\infty} x(x-1) \; \frac{e^{-\lambda} \; \lambda^x}{x!} + \underbrace{\sum_{x=0}^{\infty} x \; \frac{e^{-\lambda} \; \lambda^x}{x!}}_{=\lambda} - \lambda^2$$

$$= 0 + 0 + \sum_{x=2}^{\infty} x(x-1) \; \frac{e^{-\lambda} \; \lambda^x}{x!} + \lambda - \lambda^2$$

desarrollando $x! = x(x-1)(x-2)!$ y simplificando:

$$V(x) = \sum_{x=2}^{\infty} x(x-1) \; \frac{e^{-\lambda} \; \lambda^2 \; \lambda^{x-2}}{x(x-1)(x-2)!} + \lambda - \lambda^2 = \lambda^2 \sum_{x=2}^{\infty} \frac{e^{-\lambda} \; \lambda^{x-2}}{(x-2)!} + \lambda - \lambda^2$$

Por último, haciendo $X = (x-2)$ y verificando que para $x = 2 \rightarrow X = 0$:

$$V(x) = \lambda^2 \underbrace{\sum_{X=0}^{\infty} \frac{e^{-\lambda} \; \lambda^X}{X!}}_{=1 \text{ por condición de cierre}} + \lambda - \lambda^2 = \lambda \tag{5.20}$$

Verificando entonces que la Distribución de Poisson tiene su esperanza matemática y variancia iguales a λ.

Proporcionalidad de la Distribución de Poisson

Cuando se enumeraron las condiciones para la aplicación de la distribución de Poisson se estableció que las realizaciones del experimento aleatorio deben cumplirse dentro de un intervalo de tiempo o espacio continuo. Esto permite la aplicación del principio de proporcionalidad que establece que el parámetro $E(X)$ del proceso bajo estudio, representado en este caso por λ, es proporcional a la extensión total del tiempo de duración del proceso. En función de esto, una distribución de Poisson con un valor particular de λ correspondiente a un proceso que tiene una duración determinada de tiempo, puede aplicarse sin inconvenientes a una duración menor, modificando proporcionalmente el valor de λ.

Ejemplo 5.10. La oficina de un servicio de reparaciones de equipos de acondicionadores de aire recibe durante el verano, en promedio, cinco pedidos de reparación por hora. ¿Cuál es la probabilidad de que reciban tres pedidos en media hora?

Para comprender mejor el problema se calculará primeramente la probabilidad de que se presenten tres pedidos para el tiempo total, equivalente a una hora, teniendo presente que $\lambda = 5$:

$$P(\text{en 1 hora, 3 pedidos}) = \frac{e^{-5}\ 5^3}{3!} = \frac{0{,}000674 \cdot 125}{6} = 0{,}1404$$

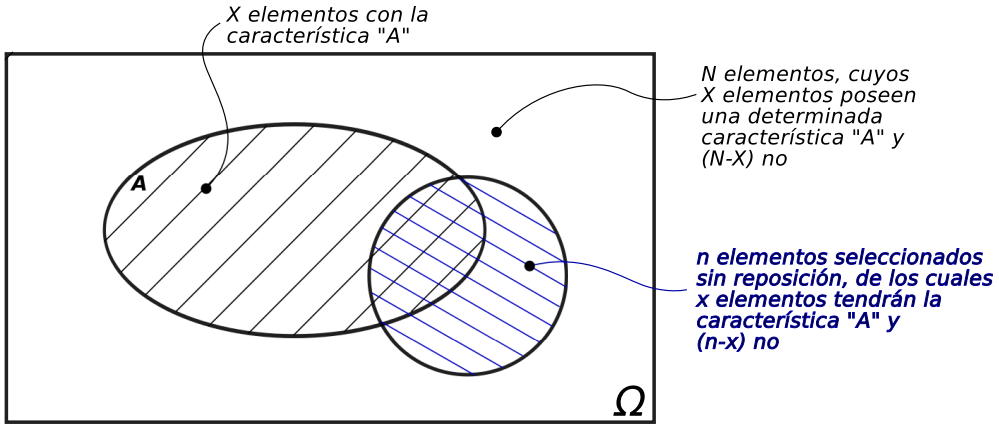
Ahora obtengamos el valor proporcional de λ para media hora. Como para una hora $\lambda = 5$, para media hora deberá ser $\lambda = 2{,}5$, por lo que la probabilidad pedida se calcula haciendo:

$$P(\text{en 1/2 hora, 3 pedidos}) = \frac{e^{-2{,}5}\ 2{,}5^3}{3!} = \frac{0{,}082 \cdot 15{,}625}{6} = 0{,}2138$$

Distribución Hipergeométrica

La distribución Hipergeométrica se aplica cuando las realizaciones del experimento aleatorio se realizan **sin reposición** o se generan sucesos que son condicionales entre sí, lo cual marca la principal diferencia con las distribuciones Binomial y de Poisson, ambas para realizaciones con reposición o de eventos independientes entre sí.

El siguiente ejemplo permite presentar el tema: sea un conjunto de N elementos cuyos X elementos poseen una determinada característica A, mientras que el resto de los $(N-X)$ elementos no la poseen. Se selecciona una cantidad n de elementos, efectuando la elección sin reposición y se desea calcular la probabilidad de que, en esos n elementos, hayan x elementos con la característica A.



Con estas condiciones planteadas, es claro que debe cumplirse que:

$$N > 0$$

$$0 \leq X \leq N$$

$$1 \leq n \leq N$$

$$0 \leq x \leq n \quad \text{y} \quad x \leq X$$

lo que equivale a decir que la probabilidad de que se presente el suceso A en la primer realización del experimento es $P(A_1) = \dfrac{X}{N}$.

Deducción de la fórmula de la Distribución Hipergeométrica

La deducción de la fórmula de la distribución Hipergeométrica se plantea a partir de un análisis semejante al realizado en la deducción de la fórmula de la distribución Binomial. El experimento aleatorio se realiza n veces sin reposición y se desea que en x ocasiones se presente el suceso A, por lo que en las restantes $(n-x)$ ocasiones deberá presentarse el suceso opuesto $\overline{A}$. A un conjunto de esas n realizaciones del experimento lo denominaremos secuencia i-ésima, simbolizada como S_i.

Supongamos ahora que la primer secuencia, S_1, sea aquella en la que luego de realizar los n experimentos, en primer lugar se presentan los x sucesos A y a continuación los $(n-x)$ sucesos $\overline{A}$:

$$S_1 = A_1 \cap A_2/A_1 \cap A_3/(A_1 \cap A_2) \cap \ldots \cap \overline{A}_{x+1}/(A_1 \cap A_2 \cap \ldots \cap A_x) \cap \\ \cap \ldots \cap \overline{A}_n/(A_1 \cap A_2 \cap \ldots \cap \overline{A}_{n-1})$$

Si en la igualdad precedente se aplica el concepto de probabilidad, tendremos que,

$$P(S_1) = P\big[A_1 \cap A_2/A_1 \cap A_3/(A_1 \cap A_2) \cap \ldots \cap \overline{A}_{x+1}/(A_1 \cap A_2 \cap \ldots \cap A_x) \cap \\ \cap \ldots \cap \overline{A}_n/(A_1 \cap A_2 \cap \ldots \cap \overline{A}_{n-1})\big]$$

Como las n realizaciones del experimento son condicionales, el cálculo de la probabilidad se resuelve mediante la aplicación de la regla de la multiplicación para ese tipo de sucesos:

$$P(S_1) = P(A_1)\ P(A_2/A_1)\ P\big[A_3/(A_1A_2)\big] \ \ldots\ P\big[\ \overline{A}_{x+1}/(A_1A_2 \ldots A_x)\big] \ \ldots \\ \ldots\ P\big[\ \overline{A}_n/(A_1A_2 \ldots \overline{A}_{n-1})\big]$$

Tomando en consideración, como ya se ha indicado, que la $P(A_1) = \dfrac{X}{N}$ y que con cada realización del experimento el valor de la probabilidad se calcula modificando en una unidad menos las cantidades tanto del numerador X como del denominador N, la expresión anterior se convierte en

$$P(S_1) = \\ = \overbrace{\underbrace{\frac{X}{N}\ \frac{X-1}{N-1}\ \frac{X-2}{N-2} \cdots \frac{(X-x)+1}{(N-x)+1}}_{x \text{ factores}}\ \underbrace{\frac{(N-X)}{(N-x)}\ \frac{(N-X)-1}{(N-x)-1} \cdots \frac{(N-X)-(n-x)+1}{(N-n)+1}}_{(n-x) \text{ factores}}}^{n \text{ factores}} \tag{5.21}$$

Ahora bien, en el numerador se cumple que $X(X-1)(X-2)\ldots[(X-x)+1] = \dfrac{X!}{(X-x)!}$ y que $(N-X)[(N-X)-1]\ldots[(N-X)-(n-x)+1] = \dfrac{(N-X)!}{[(N-X)-(n-x)]!}$; mientras

que en el denominador se cumple que $N(N-1)(N-2)\dots[(N-n)+1] = \dfrac{N!}{(N-n)!}$, por lo cual la ecuación (5.21) puede escribirse:

$$P(S_1) = \frac{\dfrac{X!}{(X-x)!}\ \dfrac{(N-X)!}{[(N-X)-(n-x)]!}}{\dfrac{N!}{(N-n)!}}$$

que resulta ser la probabilidad de ocurrencia de la secuencia S_1.

Cualquier otra secuencia S_i (para $i \neq 1$) estará compuesta por la misma cantidad de sucesos A y $\overline{A}$ que la secuencia S_1 pero con un orden de aparición diferente. Por ejemplo, denominaremos S_2 a la secuencia que modifica la posición de los sucesos A_x y $\overline{A}_{x+1}$

$$S_2 = A_1 \cap A_2/A_1 \cap \dots \cap \overline{A}_x/(A_1 \cap A_2 \cap \dots \cap A_{x-1}) \cap A_{x+1}/(A_1 \cap A_2 \cap \dots \cap \overline{A}_x) \cap \\ \cap \dots \cap \overline{A}_n/(A_1 \cap A_2 \cap \dots \cap \overline{A}_{n-1})$$

Es posible verificar que la probabilidad de ocurrencia de la secuencia S_2 es exactamente igual a la de la secuencia S_1 y en general a cualquier otra secuencia S_i. Por lo tanto, si todas las secuencias tienen igual probabilidad, el resultado buscado se obtiene multiplicando la probabilidad de una secuencia por la cantidad existente de secuencias, dada por el número combinatorio $\binom{n}{x}$. Es decir que, finalmente, la probabilidad de realizar n experimentos condicionales y obtener x veces un suceso A está dada por:

$$\begin{aligned} P(n,x) &= \frac{n!}{x!\ (n-x)!}\ \frac{\dfrac{X!}{(X-x)!}\ \dfrac{(N-X)!}{[(N-X)-(n-x)]!}}{\dfrac{N!}{(N-n)!}} \\ &= \frac{n!\ X!\ (N-X)!(N-n)!}{x!\ (n-x)!\ (X-x)!\ [(N-X)-(n-x)]!\ N!} \\ &= \frac{\dfrac{X!}{x!\ (X-x)!}\ \dfrac{(N-X)!}{(n-x)!\ [(N-X)-(n-x)]!}}{\dfrac{N!}{n!\ (N-n)!}} = \frac{\binom{X}{x}\ \binom{N-X}{n-x}}{\binom{N}{n}} \end{aligned}$$

Luego, la probabilidad en la distribución Hipergeométrica se obtiene calculando

$$P(\text{en } n \text{ realizaciones, } x \text{ apariciones}) = \frac{\binom{X}{x}\ \binom{N-X}{n-x}}{\binom{N}{n}} \tag{5.22}$$

Ejemplo 5.11. Se reciben pequeños motores eléctricos en lotes de 50. Antes de aceptarlos un inspector elige 5 motores y los inspecciona. Si ninguno de ellos es defectuoso, el lote se acepta. Si se encuentra que uno o más son defectuosos, se inspecciona el lote

completo. Supongamos que en realidad hay 3 motores defectuosos en el lote. ¿ Cuál es la probabilidad de que se requiera una inspección del 100 %?

Si hacemos que x sea el número de motores defectuosos encontrados, se necesitará una inspección del 100 % si y sólo si $x \geq 1$. Luego,

$$P(x \geq 1) = 1 - P(x = 0) = 1 - \frac{\binom{3}{0}\binom{50-3}{5-0}}{\binom{50}{5}} = 0{,}276$$

Condición de cierre

Para demostrar la condición de cierre partiremos del desarrollo del Binomio de Newton para la función $(1+x)^{m+n}$, siendo m y n dos números enteros positivos cualquiera; es decir:

$$(1+x)^{m+n} = 1 + \binom{m+n}{1} x + \binom{m+n}{2} x^2 + \ldots + \binom{m+n}{r} x^r + \ldots + x^{m+n} \quad (5.23)$$

Además, es posible escribir la misma función anterior como $(1+x)^m\ (1+x)^n$,

$$(1+x)^{m+n} = (1+x)^m\ (1+x)^n = \left[1 + \binom{m}{1} x + \binom{m}{2} x^2 + \ldots + \binom{m}{r} x^r + \ldots + x^m\right] \times \left[1 + \binom{n}{1} x + \binom{n}{2} x^2 + \ldots + \binom{n}{r} x^r + \ldots + x^n\right] \quad (5.24)$$

Ahora bien, igualando los coeficientes que multiplican a x^r en las ecuaciones (5.23) y (5.24), obtenemos:

$$\binom{m+n}{r} = \binom{m}{r} + \binom{n}{1}\binom{m}{r-1} + \binom{n}{2}\binom{m}{r-2} + \ldots + \binom{n}{s}\binom{m}{r-s} + \ldots + \binom{n}{r} \quad (5.25)$$

igualdad que se satisface para todos los enteros m y n tales que $m+n \geq r$. La ecuación (5.25) puede escribirse como

$$\binom{m+n}{r} = \sum_{s=0}^{r} \binom{n}{s}\binom{m}{r-s} \quad (5.26)$$

la cual se conoce como **teorema de Vandermonde**. Utilizando (5.26) se observa que si $m = N - X$, $n = X$, $r = n$, y $s = x$:

$$\binom{N}{n} = \sum_{x=0}^{n} \binom{X}{x}\binom{N-X}{n-x} \quad (5.27)$$

Por lo cual, recordando la ecuación (5.22), la suma de las probabilidades hipergeométricas es igual a la unidad.

Parámetros de la Distribución Hipergeométrica - Esperanza matemática

De la misma manera que en las distribuciones anteriores, la esperanza matemática se obtiene haciendo:

$$E(x) = \sum_{x=0}^{n} x \, \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}} = \sum_{x=1}^{n} x \, \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}} \tag{5.28}$$

dado que para $x = 0$ el término correspondiente de la sumatoria se anula. En lo siguiente utilizaremos el lema presentado a continuación:

Lema 1: $\binom{a}{b} = \frac{a}{b}\binom{a-1}{b-1}$.

Demostración: $\binom{a}{b} = \frac{a!}{b!\,(a-b)!} = \frac{a}{b}\,\frac{(a-1)!}{(b-1)!\,\big[(a-1)-(b-1)\big]!} = \frac{a}{b}\binom{a-1}{b-1}$.

En consecuencia, podemos ver que

$$\binom{N}{n} = \frac{N}{n}\binom{N-1}{n-1} \qquad \text{y} \qquad x\binom{X}{x} = X\binom{X-1}{x-1}$$

Con lo cual, reemplazando ambas ecuaciones en (5.28):

$$E(x) = \sum_{x=1}^{n} \frac{X\binom{X-1}{x-1}\binom{N-X}{n-x}}{\frac{N}{n}\binom{N-1}{n-1}} = n\,\frac{X}{N}\sum_{x=1}^{n} \frac{\binom{X-1}{x-1}\binom{N-X}{n-x}}{\binom{N-1}{n-1}}$$

Haciendo, $z = x-1, Z = X-1, M = N-1$ y $m = n-1$:

$$E(x) = n\,\frac{X}{N}\underbrace{\sum_{z=0}^{m} \frac{\binom{Z}{z}\binom{M-Z}{m-z}}{\binom{M}{m}}}_{\text{cond. de cierre}} = n\,\frac{X}{N}$$

Finalmente, haciendo $P(A_1) = \dfrac{X}{N} = p$:

$$E(x) = n\,\frac{X}{N} = np \tag{5.29}$$

Parámetros de la Distribución Hipergeométrica - Variancia

Nuevamente, partimos de la definición de variancia, reemplazando por la expresión correspondiente a la distribución hipergeométrica,

$$V(x) = \sum_{x=0}^{n} x^2 \, \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}} - E(x)^2 \tag{5.30}$$

Considerando que $x^2 = x(x-1) + x$:

$$\begin{aligned} V(x) &= \sum_{x=0}^{n} \big[x(x-1) + x\big] \, \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}} - \left(n \, \frac{X}{N}\right)^2 \\ &= \sum_{x=0}^{n} x(x-1) \, \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}} + \underbrace{\sum_{x=0}^{n} x \, \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}}}_{E(x)} - \left(n \, \frac{X}{N}\right)^2 \\ &= \sum_{x=2}^{n} x(x-1) \, \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}} + n \, \frac{X}{N} - \left(n \, \frac{X}{N}\right)^2 \end{aligned} \tag{5.31}$$

donde en la última sumatoria se han descartado los dos primeros términos dado que se anulan para $x = 0$ y $x = 1$. Por otro lado, utilizando el *lema 1* se observa que

$$\binom{N}{n} = \frac{N\ (N-1)}{n\ (n-1)} \binom{N-2}{n-2} \qquad \text{y} \qquad x\ (x-1) \binom{X}{x} = X\ (X-1) \binom{X-2}{x-2}$$

por lo cual la expresión (5.31) queda:

$$\begin{aligned} V(x) &= \sum_{x=2}^{n} X(X-1) \, \frac{\binom{X-2}{x-2}\binom{N-X}{n-x}}{\frac{N\ (N-1)}{n\ (n-1)}\binom{N-2}{n-2}} + n \, \frac{X}{N} - \left(n \, \frac{X}{N}\right)^2 \\ &= \frac{X\ (X-1)}{N\ (N-1)}\ n\ (n-1) \sum_{x=2}^{n} \frac{\binom{X-2}{x-2}\binom{N-X}{n-x}}{\binom{N-2}{n-2}} + n \, \frac{X}{N} - n^2 \, \frac{X^2}{N^2} \end{aligned}$$

Realizando las siguientes sustituciones, $z = x-2, Z = X-2, M = N-2$ y $m = n-2$:

$$V(x) = \frac{X\ (X-1)}{N\ (N-1)}\ n\ (n-1)\ \underbrace{\sum_{z=0}^{m} \frac{\binom{Z}{z}\ \binom{M-Z}{m-z}}{\binom{M}{m}}}_{\text{cond. de cierre}} + n\ \frac{X}{N} - n^2\ \frac{X^2}{N^2}$$

$$= \frac{X\ (X-1)}{N\ (N-1)}\ n\ (n-1)\ + n\ \frac{X}{N} - n^2\ \frac{X^2}{N^2}$$

tomando como común denominador $N^2\ (N-1)$ y factor común nX,

$$V(x) = \frac{nX}{N^2\ (N-1)}\ \Big[N(X-1)(n-1) + N(N-1) - nX(N-1)\Big]$$

$$= \frac{nX}{N^2\ (N-1)}\ \Big[nXN - XN - nN + N + N^2 - N - nXN + nX\Big]$$

simplificando y reordenando,

$$V(x) = \frac{nX}{N^2\ (N-1)}\ \Big[-XN - nN + N^2 + nX\Big] = \frac{nX}{N^2\ (N-1)}\ \Big[N(N-X) - n(N-X)\Big]$$

$$= n\ \frac{X}{N}\frac{(N-X)}{N}\ \frac{(N-n)}{(N-1)} = npq\ \frac{(N-n)}{(N-1)}$$

con lo cual,

$$V(x) = npq\ \frac{N-n}{N-1} \tag{5.32}$$

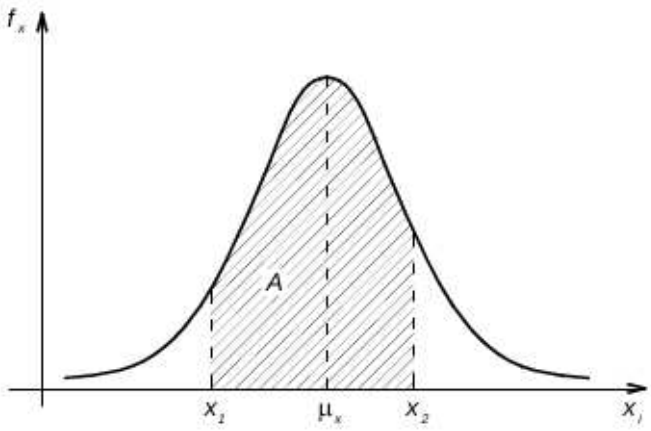
5.3. Distribuciones de probabilidad para variables aleatorias continuas

Distribución Normal o de Gauss-Laplace

La función de densidad Normal fue propuesta por los matemáticos Carl Friedrich Gauss y Pierre Simon Laplace, quienes llegaron a ella prácticamente en forma simultánea y estudiando la distribución de los errores en mediciones. Su expresión matemática es:

$$f(x) = \frac{1}{\sigma\ \sqrt{2\ \pi}}\ e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{para} -\infty < x < +\infty \tag{5.33}$$

donde $E(x) = \mu$ y $V(x) = \sigma^2$. La forma típica de la distribución Normal es como la que se muestra a continuación:



indicando que una variable tiene Distribución Normal del siguiente modo:

$$x \sim N(\mu; \sigma^2)$$

Figura 5.7: Pierre Simon Laplace (1749-1827).

La probabilidad de que la variable X se encuentre entre dos valores arbitrarios x_1 y x_2 se obtiene calculando la integral de la función $f(X)$ entre los dos valores indicados, y como en toda función de densidad, se representa gráficamente por un área. En la práctica no se realiza el cálculo de las probabilidades, ya que se utiliza una tabla a partir de una transformación de variables que permite hallar la variable z denominada **variable estandarizada**:

$$z = \frac{x - \mu_x}{\sigma_x} \tag{5.34}$$

Es posible observar que:

$$E(z) = E\left(\frac{x - \mu_x}{\sigma_x}\right) = \frac{1}{\sigma_x}\big[E(x) - \mu_x\big] = 0$$

dado que $E(x) = \mu_x$. Además,

$$V(z) = E\big[\big(z - E(z)\big)^2\big] = E(z^2) = E\left[\frac{(x-\mu_x)^2}{\sigma_x^2}\right] = \frac{1}{\sigma_x^2} E\Big[\big(x - E(x)\big)^2\Big] = 1.$$

Es decir, $E(z) = 0$ y $V(z) = 1$, de modo que $DS(z) = 1$, con lo cual la expresión de la función de densidad en el caso de la variable estandarizada es

$$f(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}$$

Luego, $z \sim N(0;1)$ (se lee: *la variable aleatoria z se encuentra distribuida normalmente, con media igual a* 0 *y variancia igual a* 1).

La solución práctica para obtener las probabilidades consiste en utilizar la tabla de probabilidades, apropiada para calcular cualquier probabilidad en el caso normal, sin que importe cuáles son los valores particulares de la variable aleatoria ni los parámetros de la distribución. Esta tabla se utiliza en cualquier circunstancia en que la variable aleatoria se distribuya normalmente, pero para su aplicación se requiere transformar la variable x_i bajo estudio en una variable estandarizada z_i, con lo cual se consigue que la media y la variancia de z_i sean iguales a cero y a uno, respectivamente. Efectuada la transformación se utiliza la tabla, reconociéndose dos casos diferentes de búsqueda:

Caso de búsqueda directa: se dispone de determinados valores de la variable aleatoria y se obtienen en la tabla las probabilidades que les corresponden.

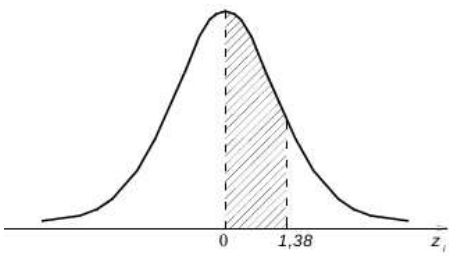
Caso de búsqueda inversa: se dispone de ciertos valores de probabilidad y se desea encontrar a qué valores de la variable aleatoria corresponden.

Manejo de la tabla

Caso directo: para ejemplificar el manejo de la tabla, se considerará que ya se ha efectuado la transformación de la variable x_i en la variable estandarizada z_i, y se presentarán las siguientes alternativas:

1) Hallar la probabilidad de que la variable z_i se encuentre entre los valores 0 y 1,38, es decir, $P(0 \leq z_i \leq 1{,}38)$.

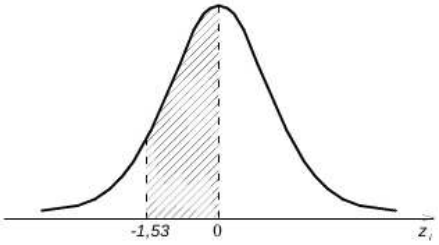
Se puede verificar en el gráfico de la derecha que el área sombreada es la probabilidad requerida, y que ella se puede obtener, en este caso, directamente de la tabla. Luego



$$P(0 \leq z_i \leq 1{,}38) = 0{,}4162$$

2) Hallar la probabilidad de que la variable z_i se encuentre entre los valores $-1{,}53$ y 0, es decir, $P(-1{,}53 \leq z_i \leq 0)$.

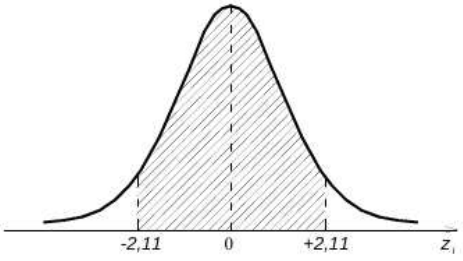
El área correspondiente a la probabilidad pedida se presenta sobre el semieje negativo de las abscisas. Como la curva es simétrica, esa área es equivalente a aquella extendida sobre el semieje positivo, alternativa que fuera planteada en el caso anterior,



$$P(-1{,}53 \leq z_i \leq 0) = P(0 \leq z_i \leq 1{,}53) = 0{,}4370$$

3) Hallar la probabilidad de que la variable z_i se encuentre entre los valores $-2{,}11$ y $2{,}11$, es decir, $P(-2{,}11 \leq z_i \leq +2{,}11)$.

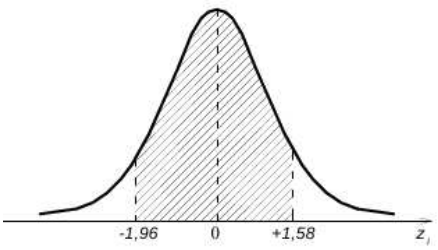
El área correspondiente a la probabilidad pedida es simétrica respecto del valor cero, por lo que se obtiene sumando dos áreas equivalentes extendidas sobre el semieje positivo, es decir que



$$\begin{aligned} P(-2{,}11 \leq z_i \leq 2{,}11) &= P(0 \leq z_i \leq 2{,}11) \\ &+ P(0 \leq z_i \leq 2{,}11) \\ &= 2\ P(0 \leq z_i \leq 2{,}11) \\ &= 2\ (0{,}4826) = 0{,}9652 \end{aligned}$$

4) Hallar la probabilidad de que la variable z_i se encuentre entre los valores $-1{,}96$ y $1{,}58$, es decir, $P(-1{,}96 \leq z_i \leq +1{,}58)$.

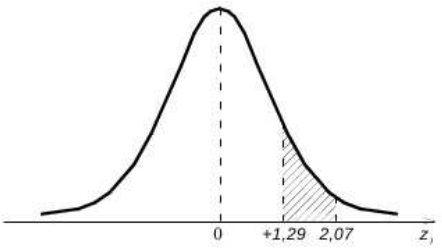
El área que corresponde a la probabilidad pedida se puede obtener sumando los sectores de la izquierda y de la derecha, que este caso no son iguales entre sí. Como la curva es simétrica,



$$\begin{aligned} P(-1{,}96 \leq z_i \leq 1{,}58) &= P(0 \leq z_i \leq 1{,}96) \\ &+ P(0 \leq z_i \leq 1{,}58) \\ &= 0{,}4750 + 0{,}4429 = 0{,}9179 \end{aligned}$$

5) Hallar la probabilidad de que la variable aleatoria z_i se encuentre entre los valores $1,29$ y $2,07$; es decir: $P(1,29 \leq z_i \leq 2,07)$.

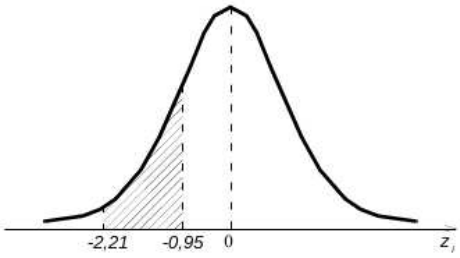
El área que corresponde a la probabilidad pedida se obtiene restando a la superficie bajo la curva entre los valores 0 y 2,07, la superficie entre los valores 0 y 1,29, es decir que



$$\begin{aligned} P(1{,}29 \leq z_i \leq 2{,}07) &= P(0 \leq z_i \leq 2{,}07) \\ &\quad - P(0 \leq z_i \leq 1{,}29) \\ &= 0{,}4808 - 0{,}4015 = 0{,}0793 \end{aligned}$$

6) Hallar la probabilidad de que la variable aleatoria z_i se encuentre entre los valores $-2,21$ y $-0,95$, es decir $P(-2,21 \leq z_i \leq -0,95)$.

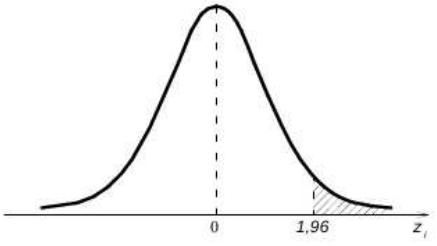
La probabilidad requerida puede obtenerse calculándola como si los valores de la variable fueran positivos, es decir, del mismo modo en que fue calculada en el caso anterior,



$$\begin{aligned} P(-2{,}21 \leq z_i \leq -0{,}95) &= P(0{,}95 \leq z_i \leq 2{,}21) \\ &= P(0 \leq z_i \leq 2{,}21) \\ &\quad - P(0 \leq z_i \leq 0{,}95) \\ &= 0{,}4864 - 0{,}3289 = 0{,}1575 \end{aligned}$$

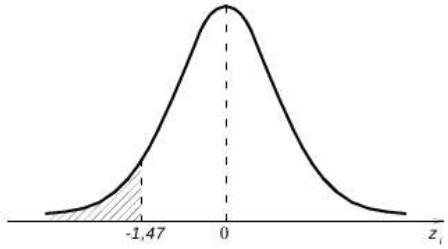
7) Hallar la probabilidad de que la variable aleatoria z_i sea mayor o igual que el valor 1,96 es decir, $P(z_i \geq 1{,}96)$.

La probabilidad requerida se obtiene tomando toda el área a la derecha del valor 0, equivalente a 0,50, restándole luego el área entre 0 y 1,96,



$$\begin{aligned} P(z_i \geq 1{,}96) &= 0{,}50 - P(0 \leq z_i \leq 1{,}96) \\ &= 0{,}50 - 0{,}4750 = 0{,}025 \end{aligned}$$

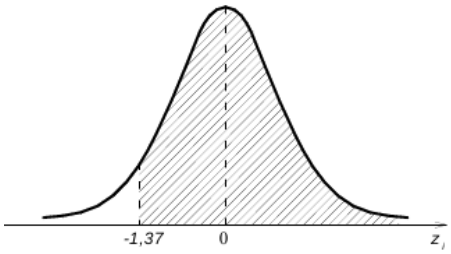
8) Hallar la probabilidad de que la variable aleatoria z_i sea menor o igual que el valor $-1{,}47$, es decir, $P(z_i \leq 1{,}47)$.



La probabilidad requerida se calcula aplicando el procedimiento explicado en el punto anterior, considerando el valor absoluto de 1,47,

$$\begin{aligned} P(z_i \geq -1{,}47) &= P(z_i \geq 1{,}47) \\ &= 0{,}50 - P(0 \leq z_i \leq 1{,}47) \\ &= 0{,}50 - 0{,}4292 = 0{,}0708 \end{aligned}$$

9) Hallar la probabilidad de que la variable aleatoria z_i sea mayor o igual que $-1{,}37$, es decir, $P(z_i \geq 1{,}37)$.

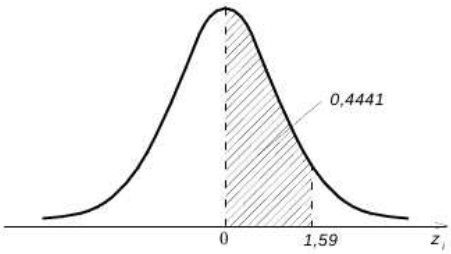


La probabilidad buscada se obtiene luego de considerar que el área total requerida se extiende entre el valor $-1{,}37$ y $+\infty$, lo cual equivale a tomar el área entre 0 y $+\infty$, por un lado, y sumarle el área entre $-1{,}37$ y 0 por el otro. Es decir,

$$\begin{aligned} P(z_i \geq -1{,}37) &= P(-1{,}37 \leq z_i \leq 0) + 0{,}50 \\ &= P(0 \leq z_i \leq 1{,}37) + 0{,}50 \\ &= 0{,}4147 + 0{,}50 = 0{,}9147 \end{aligned}$$

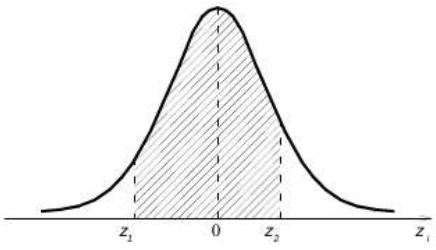
Caso inverso: se reconocen tres casos fundamentales.

1) Hallar el valor z_I tal que la probabilidad de que la variable z_i esté entre 0 y z_I sea igual a 0,4441.



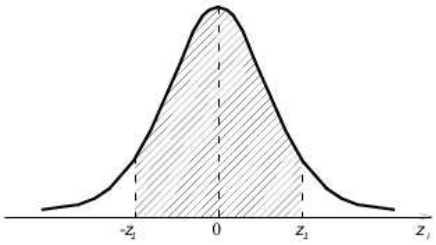
En el área de las probabilidades de la tabla, buscamos el valor 0,4441 y encontramos que ese valor está presente exactamente en la tabla y que se corresponde con un z_I igual a $1,59$. Por consiguiente $P(0 \leq z_i \leq z_I = 1,59) = 0,4441 \Rightarrow z_I = 1,59$.
Si el valor de la probabilidad no estuviera presente exactamente en la tabla, se puede efectuar una interpolación aritmética entre aquellos dos que lo encierran o, en caso de no requerir demasiada precisión, tomar el valor más cercano que figura en la tabla.

2) Hallar los valores z_1 y z_2 tales que se tenga una probabilidad igual a 0,80 de que la variable z_i se encuentre entre ambos.
En este caso se buscan dos valores de la variable z_i, z_1 y z_2, que encierren una probabilidad igual a 0,80, es decir $P(z_1 \leq z_i \leq z_2) = 0{,}80$.



Esto no tiene solución única, ya que hay infinitos pares de valores $(z_1; z_2)$ que encierran un área igual a 0,80 y sólo puede resolverse si se amplía la información disponible.

3) Hallar el valor de z_1 tal que la probabilidad de la variable z_i se encuentre entre $-z_1$ y $+z_1$ sea igual a 0,90.



Para resolver este caso, se busca aquel valor absoluto z_1 de la variable z_i que encierra una probabilidad igual a 0,90, es decir que $P(-z_1 \leq z_i \leq +z_1) = 0{,}90$, el cual sí tiene solución única, que se consigue reduciéndolo al ejemplo dado en el primer caso, es decir, correspondería calcular la $P(0 \leq z_i \leq z_1) = 0{,}45$ en la cual z_1 es igual a 1,645.

Ejemplo 5.12. Según los registros disponibles, el total de lluvia caída anualmente en una zona tiene una distribución $N(60, 225)$.

a) ¿cuál es la probabilidad de lluvias entre 40 mm y 70 mm?

b) ¿cuál es la probabilidad de lluvias de al menos 30 mm?

c) ¿cuál es la probabilidad de que si el registro supera los 40 mm, la lluvia sea menor que 70 mm?

a) sabemos que $E(x) = \mu = 60$ mm y $V(x) = \sigma^2 = 225$ mm^2. Por lo tanto,

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{40 - 60}{15} = -1{,}333; \qquad z_2 = \frac{x_2 - \mu}{\sigma} = \frac{70 - 60}{15} = 0{,}667$$

Luego, podemos obtener las probabilidades:

$$P(-1{,}333 \leq z \leq 0) = P(0 \leq z \leq 1{,}333) = 0{,}4082$$

$$P(0 \leq z \leq 0{,}667) = 0{,}2486$$

Entonces la probabilidad buscada es: $0{,}4082 + 0{,}2486 = 0{,}6568$.

b) La probabilidad de que se produzca una precipitación de *al menos* 30 *mm* debe interpretarse como la probabilidad de ocurrencia de cualquier valor igual o superior a 30 mm:

$$z = \frac{30 - 60}{15} = -2{,}00$$

$$P(0 \leq z \leq 2) = 0{,}4772 \rightarrow P(-2 \leq z \leq +\infty) = 0{,}4772 + 0{,}50 = 0{,}9772$$
$$P(\text{al menos 30 mm}) = 0{,}9772$$

c) En este caso debemos resolver la probabilidad condicional $P[(x < 70mm)/(x > 40mm)]$:

$$P(x < 70/x > 40) = \frac{P(40 < x < 70)}{P(x > 40)}$$

El valor del numerador lo tenemos calculado del punto *a*). Para determinar el denominador, procedemos:

$$P(x > 40) = P(-1{,}333 \leq z \leq 0) + 0{,}50 = 0{,}9082$$

Luego,

$$P(x < 70/x > 40) = \frac{0{,}6568}{0{,}9082} = 0{,}7232$$

Distribución Chi-cuadrado

Si la sucesión $Z_1, Z_2, \ldots, Z_n$ está compuesta por variables aleatorias independientes con distribución normal estándar, es decir, $Z_i \sim N(0, 1)$; luego la variable aleatoria Y, definida como

$$Y = Z_1^2 + Z_2^2 + \ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2} \tag{5.35}$$

posee una distribución *chi-cuadrado con n grados de libertad*. Se utilizará la notación

$$Y \sim \chi_n^2$$

para indicar que Y posee una distribución chi-cuadrado con n grados de libertad.

La distribución chi-cuadrado tiene la propiedad aditiva tal que si Y_1 e Y_2 son variables aleatorias independientes de distribución chi-cuadrado con n_1 y n_2 grados de libertad, respectivamente, entonces la suma $Y_1 + Y_2$ es una variable aleatoria chi-cuadrado con $n_1 + n_2$ grados de libertad:

$$\text{Si} \quad Y_1 \sim \chi_{n_1}^2 \quad \text{e} \quad Y_2 \sim \chi_{n_2}^2 \Rightarrow (Y_1 + Y_2) \sim \chi_{n_1+n_2}^2$$

Si Y es una variable aleatoria chi-cuadrado con n grados de libertad, entonces para cualquier $\alpha \in (0,1)$, el valor $\chi^2_{\alpha,n}$ es tal que

$$P(Y \leq \chi^2_{\alpha,n}) = \alpha$$

tal como se puede observar en la figura (5.8).

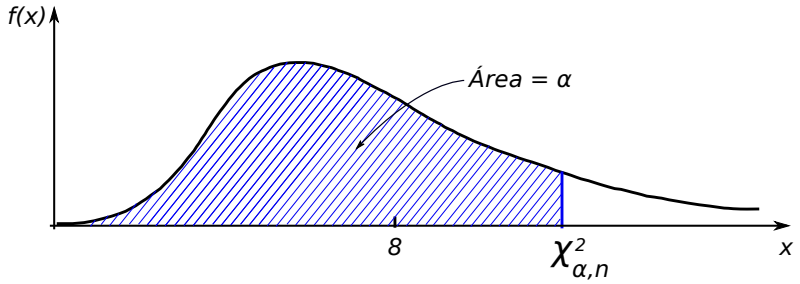


Figura 5.8: Densidad de probabilidad chi-cuadrado con 8 grados de libertad.

Ejemplo 5.13. La determinación de la posición de un punto fijo en el espacio tridimensional se realiza mediante un instrumento que proporciona las tres coordenadas con errores que se consideran variables aleatorias independientes distribuidas normalmente con media 0 y desvío estándar 2. Hallar la probabilidad de que la distancia entre el punto fijo y el punto calculado supere los 3 metros.

Si D es la distancia entre los puntos, entonces

$$D^2 = X_1^2 + X_2^2 + X_3^2$$

donde X_i es el error en la coordenada i-ésima. Como se precisa determinar el evento $D > 3$ m, luego la probabilidad buscada es

$$P(D^2 > 9) = P(X_1^2 + X_2^2 + X_3^2 > 9)$$

Estandarizando la variable de la forma $Z_i = (X_i - \mu_X)/\sigma_X = X_i/2$ y reemplazando:

$$P(D^2 > 9) = P(Z_1^2 + Z_2^2 + Z_3^2 > 9/4)$$

en donde es posible apreciar que la sumatoria de las variables Z_i^2 cumple con las condiciones enunciadas anteriormente, con lo cual es posible suponer que tiene una distribución chi-cuadrado con 3 grados de libertad. Entonces,

$$P(D^2 > 9) = P(\chi_3^2 > 9/4) = 1 - P(\chi_3^2 < 9/4) \approx 1 - 0{,}47 = 0{,}53$$

donde el valor aproximado se obtiene interpolando los valores de la tabla.

En la figura (5.9) se grafica la función densidad de acuerdo a los grados de libertad. Se puede observar que los gráficos se desarrollan sobre el semieje positivo de la variable y que la forma no es simétrica.

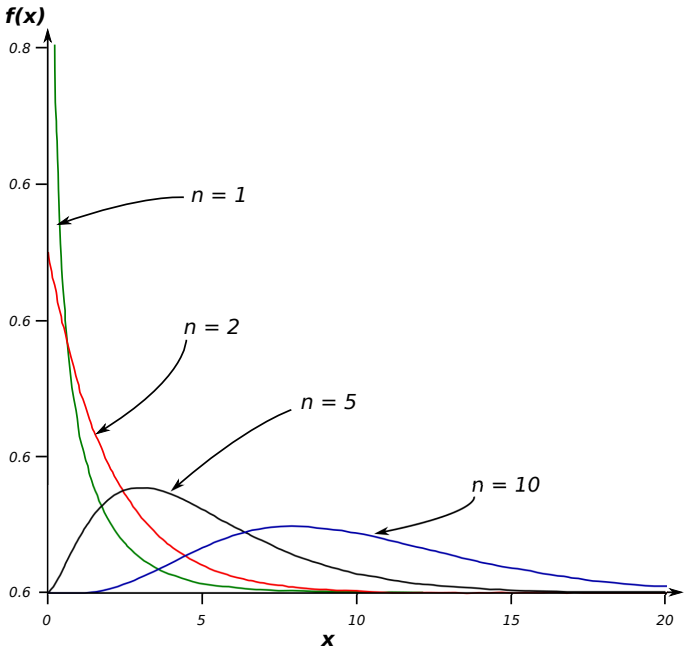


Figura 5.9: Densidad de probabilidad chi-cuadrado para distintos grados de libertad.

Variable aleatoria χ^2 para μ desconocida

Supongamos que tenemos una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una población distribuida normalmente con media μ (desconocida) y variancia σ^2. Luego, la variable aleatoria

$$\chi^2_\nu = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} = \frac{n\ S^2_X}{\sigma^2} = \frac{(n-1)\ S^2_c}{\sigma^2} \tag{5.36}$$

donde $S^2_c = \dfrac{n}{n-1} S^2_X$ es la "variancia corregida", también tiene una distribución *chi-cuadrado* (χ^2_ν) pero con ν *grados de libertad*. Los grados de libertad están dados de acuerdo a

$$\nu = n - 1$$

es decir, un grado de libertad menos que la variable aleatoria dada según la ecuación (5.35). El procedimiento para el cálculo de las probabilidades en este caso es el mismo que el planteado anteriormente.

Distribución t de Student

Si Z y χ^2_n son variables aleatorias independientes, con Z distribuida normalmente y χ^2_n con distribución chi-cuadrado con n grados de libertad, es posible definir una nueva variable aleatoria T_n según la siguiente ecuación:

$$T_n = \frac{Z}{\sqrt{\chi^2_n / n}} \tag{5.37}$$

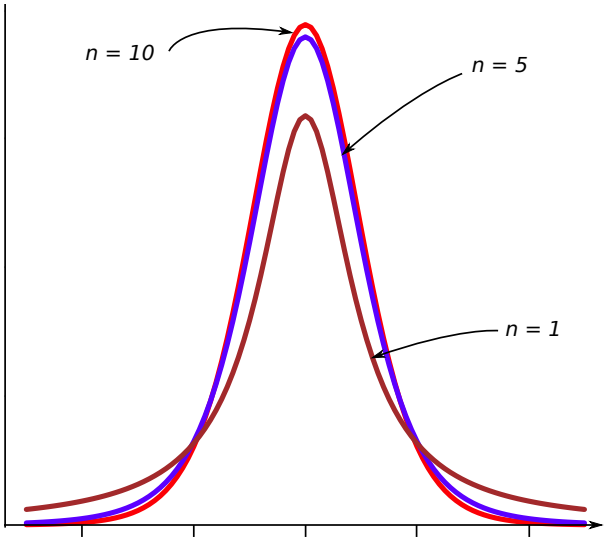


Figura 5.10: Densidad de probabilidad t de Student para distintos grados de libertad.

Esta nueva variable posee una distribución denominada "t de Student" con n grados de libertad. En la figura (5.10) se puede observar la forma de la función densidad para distintos grados de libertad.

Al igual que la función de densidad normal estandarizada, la densidad t de Student es simétrica con respecto al eje de ordenadas para $t = 0$. Además, a medida que n crece la función se acerca cada vez más a la función normal estándar.

Variable aleatoria t para σ^2 desconocida

Consideremos ahora la variable aleatoria:

$$t = \frac{\overline{X} - \mu_{\overline{X}}}{\sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n\ (n-1)}}} = \frac{\overline{X} - \mu}{\dfrac{S}{\sqrt{n-1}}} = \frac{\overline{X} - \mu}{\dfrac{S_c}{\sqrt{n}}} \tag{5.38}$$

Para determinar la distribución de esta variable aleatoria supondremos que la muestra aleatoria $X_1, X_2, \ldots, X_n$ utilizada para calcular $\overline{X}$ y S se tomó de una distribución normal con media μ y variancia σ^2. Dividiendo numerador y denominador de la expresión (5.38) por σ se obtiene

$$t = \frac{\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{S_c^2/\sigma^2}} \tag{5.39}$$

En el numerador de la ecuación (5.39) tenemos una variable aleatoria $(\overline{X} - \mu)/(\sigma/\sqrt{n})$ distribuida en forma normal estándar y en el denominador el cociente S_c^2/σ^2, el cual es una

variable aleatoria con distribución chi-cuadrado de $\nu = n - 1$ grados de libertad. Es decir, la expresión (5.39) puede escribirse de la siguiente manera:

$$t = \frac{Z}{\sqrt{\chi^2_\nu / \nu}}$$

Además, las dos variables aleatorias que aparecen en la ecuación, Z y χ^2, son independientes. Por lo tanto, de acuerdo al planteo de la ecuación (5.37), t es una variable aleatoria con distribución t de Student con ν grados de libertad.

Capítulo 6
Inferencia Estadística

6.1. Teoría de las muestras

En el capítulo inicial del presente apunte se mencionó la existencia de la teoría del muestreo, mediante la cual es posible analizar y desarrollar métodos apropiados para la selección de muestras que permitan llevar a cabo investigaciones estadísticas muestrales en reemplazo de las investigaciones censales que requieren trabajar con toda la población. Luego se ha estudiado el concepto de variable aleatoria y su distribución de probabilidad, los diferentes modelos de distribución tanto para el caso discreto como para el continuo y sus características básicas (media, variancia, etc.).

Si lo que interesa conocer ahora son los *parámetros poblacionales*, como por ejemplo la media poblacional μ_X y la variancia poblacional σ_X^2 de una determinada variable X, se deberán conocer (a través de un censo) todos los valores que pueda tomar la variable en la población. Pero según lo visto anteriormente, lo que podemos obtener son muestras de la población, de las cuales sí podemos conocer su media muestral ($\overline{x}$) y variancia muestral (S_x^2), los cuales son *estadísticos* y como tales, su valor puede variar de muestra en muestra.

También anteriormente hemos mencionado que en todo muestreo existía un cierto margen de error o incertidumbre debido a que no trabajamos con todos los elementos de la población. Pero si pudiéramos tomar una gran cantidad de muestras de una misma población, ¿podríamos conocer exactamente y sin errores los parámetros poblacionales?

En este capítulo nos basaremos en la información que aporta una muestra para inferir sobre distintos aspectos de la población. En particular, nos centraremos en la media y variancia muestrales.

Distribución de muestreo de la media muestral

Consideremos una población de elementos, cada uno de los cuales es representado por un valor numérico. Por ejemplo, la población puede estar referida a un conjunto de personas de una determinada región y su valor representativo puede ser su edad, su ingreso anual, su estatura, etc. Sean $X_1, X_2, \ldots, X_n$ una muestra de esa población. La media muestral estará definida por:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Dado que el valor de la media muestral $\overline{X}$ es determinado por los valores de las variables aleatorias en la muestra, luego $\overline{X}$ también es una variable aleatoria. Su valor esperado y variancia pueden ser determinados de la siguiente manera:

$$\begin{aligned} E(\overline{X}) &= E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] \\ &= \frac{1}{n}\big[E(X_1) + E(X_2) + \ldots + E(X_n)\big] \\ &= \frac{1}{n} n\ \mu_X = \mu_X \end{aligned}$$

y

$$\begin{aligned} V(\overline{X}) &= V\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] \\ &= \frac{1}{n^2}\big[V(X_1) + V(X_2) + \ldots + V(X_n)\big] \\ &= \frac{1}{n^2} n\ \sigma_X^2 = \frac{\sigma_X^2}{n} \end{aligned}$$

donde esta última ecuación utiliza la propiedad de la variancia de la suma de **variables aleatorias independientes**.

Por lo tanto, el valor esperado de la variable aleatoria *media muestral* es igual al valor esperado de la variable aleatoria X, mientras que la variancia de la media muestral es igual a la variancia de X dividido n, es decir:

$$\mu_{\overline{X}} = \mu_X \tag{6.1}$$

$$\sigma_{\overline{X}}^2 = \frac{\sigma_X^2}{n} \quad \Rightarrow \quad \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \tag{6.2}$$

Esto se traduce en que la variable aleatoria $\overline{X}$ también tiene su centro en la media poblacional μ_X pero su dispersión disminuye a medida que aumenta n.

Ejemplo: La población sobre la cual se realiza el muestreo cuando deseamos obtener la distribución de muestreo de la media muestral está compuesta por todas las medias muestrales que se podrían obtener mediante muestreos aleatorios simples. Supongamos la existencia de una población de una variable $\boldsymbol{X}$ de tamaño $N = 5$ con los siguientes valores:

$X_1 = 2; X_2 = 3; X_3 = 6; X_4 = 8; X_5 = 11.$

Con estos datos podemos calcular la media y variancia poblacionales:

$$\mu_X = \frac{\sum_{i=1}^{N} X_i}{N} = 6$$

$$\sigma_X^2 = \frac{\sum_{i=1}^{N} (X_i - \mu_X)^2}{N} = 10{,}8$$

Efectuamos ahora la selección de todas las muestras posibles de tamaño $n = 2$ que podemos obtener de la población de la variable $\boldsymbol{X}$ realizando muestreos aleatorios simples

con reposición. Como la selección de muestras con reposición impide que el conjunto de elementos muestreados se agote, esto resulta equivalente a trabajar con una población de tamaño infinito.

La siguiente tabla contiene el total de muestras que pueden generarse. Es decir, es la población de muestras de tamaño $n = 2$ extraídas con reposición de una población de tamaño $N = 5$:

2-2	2-3	2-6	2-8	2-11
3-2	3-3	3-6	3-8	3-11
6-2	6-3	6-6	6-8	6-11
8-2	8-3	8-6	8-8	8-11
11-2	11-3	11-6	11-8	11-11

Se han obtenido 25 muestras diferentes de tamaño n con reposición, es decir que el número total de muestras de tamaño n que pueden ser seleccionadas con reposición es $M = N^n$. Si el tamaño de la muestra hubiera sido de 3 elementos, se hubieran podido construir 125 muestras diferentes; en cambio, si el tamaño de la población hubiera sido igual a 6 y se extraían muestras de tamaño $n = 2$, el total de muestras hubiera sido 36; etcétera.

La población de medias muestrales, denominada "distribución de muestreo de la media muestral", resulta entonces:

2	2.5	4	5	6.5
2.5	3	4.5	5.5	7
4	4.5	6	7	8.5
5	5.5	7	8	9.5
6.5	7	8.5	9.5	11

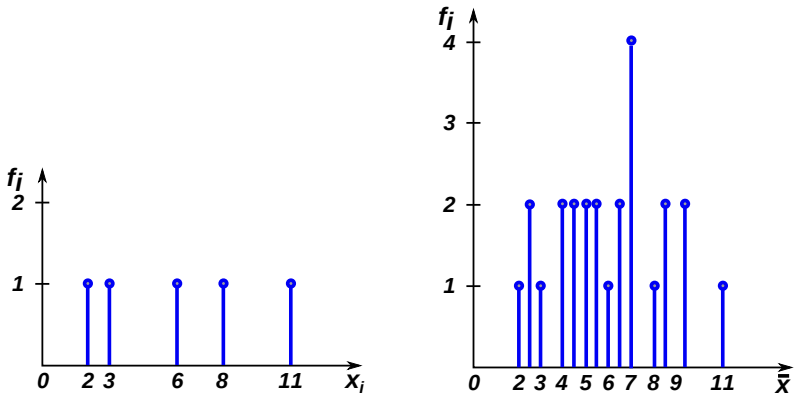
Se observa claramente que hay varios resultados para el valor de la media muestral, y que ellos dependen de cómo se encuentren conformadas las M muestras diferentes de tamaño $n = 2$. Cualquier cálculo muestral (es decir cualquier otra medida de posición e incluso de dispersión que se obtenga a partir de los datos muestrales) puede considerarse como variable. Atendiendo al hecho de que la media muestral resulta una variable y que el cuadro precedente constituye la población de medias muestrales, se pueden obtener tanto la media poblacional como la variancia poblacional de la variable media muestral, es decir que pueden calcularse $\mu_{\overline{x}}$ y $\sigma^2_{\overline{x}}$:

$$\mu_{\overline{x}} = \frac{\sum_{i=1}^{M} \overline{x}_i}{M} = \frac{150}{25} = 6$$

$$\sigma^2_{\overline{x}} = \frac{1}{M} \sum_{i=1}^{M} (\overline{x}_i - \mu_{\overline{x}})^2 = \frac{135}{25} = 5{,}4$$

Con lo cual vemos que, tal como se ha demostrado anteriormente, $\mu_{\overline{x}} = \mu_x = 6$ y $\sigma^2_{\overline{x}} = \frac{\sigma^2_x}{n} = \frac{10{,}8}{2} = 5{,}4$.

En la figura (6.1) podemos comparar gráficamente la distribución de la variable original $\boldsymbol{X}$ con la distribución muestral de las medias,



(a) Distribución de $\boldsymbol{X}$. (b) Distribución de $\overline{\boldsymbol{X}}$.

Figura 6.1: Comparación de las distribuciones.

en la cual se observa que:

a) los valores extremos de la variable media muestral son coincidentes con los de la variable $\boldsymbol{X}$ (en este caso son 2 y 11).

b) a medida que el tamaño n de la muestra crece, como la cantidad de muestras posibles aumentará considerablemente (N^n), aparecerán para la media muestral nuevos valores que siempre oscilarán entre los valores extremos ya determinados. Es decir que, en ese caso, la gráfica de bastones que se observa más arriba presentará nuevos valores y una mayor cantidad de bastones. En el límite, cuando n crezca indefinidamente, la variable media muestral se convertirá en continua y la gráfica de bastones **se transformará en un área**.

Teorema Central del Límite

El teorema central del límite es uno de los enunciados más importantes en el campo de la probabilidad. En forma breve, este teorema enuncia que la suma de un número grande de variables aleatorias independientes posee una función de distribución aproximadamente Normal. En consecuencia, este teorema no sólo proporciona un método sencillo para calcular de manera aproximada la probabilidad de sumas de variables aleatorias independientes, sino que permite explicar por qué las frecuencias empíricas de muchas poblaciones exhiben una curva en forma de campana (es decir, aproximadamente Normal). Este teorema se enuncia de la siguiente manera:

Definición

Sea $X_1, X_2, \ldots, X_n$ una secuencia de variables aleatorias independientes distribuidas idénticamente, cada una de las cuales tiene una media μ_X y una variancia σ_X^2. Luego, la suma de esa secuencia, $S_n = \sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n$ es una nueva variable aleatoria que se distribuye siguiendo aproximadamente una función Normal con media $E(S_n) = n\mu_X$ y variancia $V(S_n) = n\sigma_X^2$, para n lo suficientemente grande.

A partir de este enunciado es posible concluir que estandarizando la variable S_n de la siguiente manera:

$$Z = \frac{S_n - n\mu_X}{\sqrt{n\ \sigma_X^2}} = \frac{S_n - n\mu_X}{\sqrt{n}\ \sigma_X} \tag{6.3}$$

se obtiene una variable aleatoria Z tal que, para n grande,

$$P\left(\frac{S_n - n\mu_X}{\sqrt{n}\ \sigma_X} < z\right) = P(Z < z) = \mathbf{\Phi}(z) \quad \forall\ z \in \mathbb{R}$$

donde $\mathbf{\Phi}$ es la función de distribución $N(0,1)$. Es decir, $Z \underset{n\to\infty}{\to} N(0,1)$.

Una consecuencia de este teorema puede verse si se analiza la ecuación (6.3). En efecto,

$$\frac{S_n - n\mu_X}{\sqrt{n\ \sigma_X^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu_X}{\frac{\sqrt{n}}{n}\ \sigma_X} = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}}$$

por lo que se puede concluir que la variable media muestral también se distribuye normalmente con parámetros $\mu_{\overline{X}} = \mu_X$ y $\sigma_{\overline{X}}^2 = \frac{\sigma_X^2}{n}$. Es decir que $\overline{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$ para n lo suficientemente grande. Obsérvese que en esta última conclusión no se hace mención alguna sobre el tipo de distribución de la variable X, lo cual es sumamente interesante.

En conclusión:

- La distribución de muestreo de $\overline{\boldsymbol{X}}$ es aproximadamente Normal si el tamaño muestral es lo suficientemente grande.
- A mayor tamaño muestral, mejor aproximación a la distribución Normal. En general, se considera que si $n > 30$ se obtiene una buena aproximación.
- La media de la distribución de $\overline{X}$ es igual a de la distribución de $\boldsymbol{X}$, es decir, μ_X.
- El desvío estándar de la distribución de $\overline{X}$, $\sigma_X/\sqrt{n}$, es menor que el correspondiente a la distribución de $\boldsymbol{X}$.
- La distribución de muestreo de $\frac{\overline{X}-\mu_X}{\sigma_X/\sqrt{n}}$ tiende a la distribución Normal Estándar a medida que $n \to \infty$.

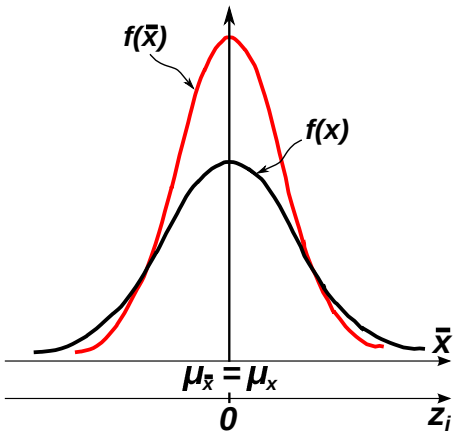


Figura 6.2: Comparación de la distribución de la variable $\overline{X}$ *vs* la distribución de la variable X.

Distribución para el caso de muestro sin reposición

Para el caso de muestro sin reposición o bien para poblaciones finitas, siguen siendo válidas las conclusiones alcanzadas en los puntos anteriores con algunas diferencias a destacar. En primer lugar, el total de muestras sin reposición de tamaño n que pueden extraerse de una población de tamaño N es $\binom{N}{n}$, lo cual difiere del caso con reposición.

La segunda diferencia es que en el caso sin reposición la variancia de la variable aleatoria media muestral, la variancia y el desvío resultan:

$$\sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}\frac{N-n}{N-1} \quad \Rightarrow \quad \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{6.4}$$

El factor de corrección $\frac{N-n}{N-1}$ es el mismo que aparece en la fórmula del cálculo de la variancia para la distribución hipergeométrica. Es posible observar que este factor es menor o igual a 1, siendo cercano a 1 cuando $N >>> n$ y cercano a 0 cuando $n \approx N$.

Ejemplo: Continuando con el ejemplo anterior para una población de tamaño $N = 5$, seleccionando ahora muestras de tamaño $n = 2$ sin reposición se obtienen las siguientes muestras:

2-3	2-6	2-8	2-11
	3-6	3-8	3-11
		6-8	6-11
			8-11

donde claramente vemos que el número de muestras es $\binom{5}{2} = 10$. Es decir, es el número de combinaciones que se pueden hacer con N elementos tomados de a n. El siguiente cuadro contiene las medias muestrales de cada una de las muestras que aparecen en el cuadro anterior:

2.5	4	5	6.5
	4.5	5.5	7
		7	8.5
			9.5

Siendo el valor de la media poblacional y variancia de la variable media muestral:

$$\mu_{\overline{x}} = \frac{\sum_{i=1}^{M} \overline{x}_i}{M} = \frac{60}{10} = 6$$

$$\sigma^2_{\overline{x}} = \frac{1}{M}\sum_{i=1}^{M}(\overline{x}_i - \mu_{\overline{x}})^2 = \frac{40{,}5}{10} = 4{,}05$$

Es decir, $\mu_{\overline{x}} = \mu_x$ y $\sigma^2_{\overline{x}} = \frac{\sigma^2_x}{n}\frac{N-n}{N-1} = 5{,}4\frac{3}{4} = 4{,}05$.

Finalmente es necesario destacar que también en el muestreo sin reposición, cuando $n \to \infty$, la variable media muestral se distribuye normalmente con media poblacional $\mu_{\overline{X}} = \mu_X$ pero con variancia poblacional $\sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}\frac{N-n}{N-1}$. Luego,

$$\overline{X} \sim N\left(\mu_{\overline{X}}, \frac{\sigma^2_X}{n}\frac{N-n}{N-1}\right) \quad \text{si } n \to \infty \tag{6.5}$$

Distribución de muestreo del desvío estándar

El desvío muestral (S_X) es también una variable que posee media y variancia poblacionales, las cuales son (sin demostración):

$$E(S_X) = \mu_{S_X} = \sigma_X$$

y

$$V(S_X) = \sigma^2_{S_X} = \frac{\sigma^2_X}{2n}$$

por lo cual podemos decir que la variable desvío estándar muestral, S_X, se distribuye normalmente con parámetros esperanza matemática σ_X y variancia $\frac{\sigma^2_X}{2n}$:

$$S_x \sim N\left(\sigma_X, \frac{\sigma^2_X}{2n}\right) \quad \text{si } n \to \infty \tag{6.6}$$

En el siguiente cuadro se resumen todas las distribuciones tratadas en este capítulo:

Variables	Media Poblacional	Variancia Poblacional	Desvío Estándar Poblacional
Media muestral $\overline{X}$	$E(\overline{X}) = \mu_{\overline{X}} = \mu_X$	$V(\overline{X}) = \sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}$	$DS(\overline{X}) = \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$
Desvío	$E(S_X) = \mu_{S_X} = \sigma_X$	$V(S_X) = \sigma^2_{S_X} = \frac{\sigma^2_X}{2n}$	$DS(S_X) = \sigma_{S_X} = \frac{\sigma_X}{\sqrt{2n}}$

Variables	Media Poblacional	Variancia Poblacional	Desvío Estándar Poblacional
Media	$E(\overline{X}) = \mu_{\overline{X}} = \mu_X$	$V(\overline{X}) = \sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}\frac{N-n}{N-1}$	$DS(\overline{X}) = \sigma_{\overline{X}} = \frac{\sigma^2_X}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

6.2. Teoría de la estimación estadística

La "teoría de la estimación estadística" es la parte de la inferencia estadística que trata acerca de los procedimientos específicos que posibilitan **inferir** o **estimar**, sobre la base de resultados muestrales conocidos denominados *estadísticas* (por ejem. media muestral, variancia y desvío estándar muestrales), cuáles son los valores poblacionales desconocidos correspondientes (media poblacional, variancia y desvío estándar poblacionales) denominados *parámetros*. En estos procedimientos los valores muestrales conocidos o estadísticas se transforman en estimadores de los valores poblacionales desconocidos o parámetros.

	Muestra	**Población**
Denominación	Estadísticas	Parámetros
Simbología	$\overline{x}$, M_e, S^2_x, S_x	μ_x, σ^2_x, σ_x
Función	son estimadores	deben ser estimados
Características	conocidos variables	desconocidos fijos

Existen dos tipos fundamentales de estimaciones:

- Estimación puntual. Es un procedimiento de estimación en el que se estima el parámetro mediante un sólo valor muestral.
- Estimación por intervalos. Es un procedimiento que permite, a partir de un estimador puntual, obtener dos valores que limitan un intervalo denominado *intervalo de confianza*, dentro del cual se encuentra el parámetro a estimar con una cierta probabilidad conocida, denominada **nivel de confianza.**

Estimación puntual

Los estimadores puntuales están constituidos por las estadísticas, que como se indicó anteriormente, son cálculos realizados sobre la muestra que permiten estimar a los correspondientes valores desconocidos de la población, a los que se denomina parámetros. Las estadísticas a las que se hace referencia son, por ejemplo, la media muestral ($\overline{x}$), la variancia muestral (S^2_x), el desvío estándar muestral (S_x), los cuales estiman al correspondiente parámetro, es decir, a la media poblacional (μ_x), a la variancia poblacional (σ^2_x) y al desvío estándar poblacional (σ_x).

Siendo las estadísticas valores variables (dependen de la muestra) y los parámetros valores constantes (determinados por la población), no es correcto considerar que una estadística sea igual a un parámetro. Por consiguiente, no es factible por ejemplo aceptar la igualdad:

$$\overline{x} = \mu_x$$

En su lugar, para indicar que una estadística estima a un parámetro se utiliza la siguiente simbología:

$$\overline{x} = \hat{\mu}_x$$

lo cual se lee "μ estimado" o "estimador de μ". Idéntico criterio se utiliza para indicar que las siguientes estadísticas estiman a los correspondientes parámetros:

$$M_e = \hat{\mu}_x$$
$$S_x^2 = \hat{\sigma}_x^2$$
$$S_x = \hat{\sigma}_x$$

En todos los casos la simbología utilizada indica que cada una de las estadísticas estima al parámetro, el cual lleva precisamente el símbolo del "sombrerito" para indicar que está siendo estimado.

Estimador insesgado o no viciado

Se denomina estimador insesgado (o no viciado) a aquel estimador cuya esperanza matemática dá como resultado el parámetro a estimar.

- Caso del estimador $\overline{x}$. El estimador media muestral es un estimador insesgado o no viciado. Efectivamente,

  $$E(\overline{x}) = \mu_x$$

  como ya se ha demostrado cuando estudiamos la distribución de muestreo de la media muestral. Cabe aclarar que, a diferencia del estimador $\overline{x}$, las estadísticas M_e y M_o son estimadores viciados.

  Ventajas de la media aritmética como estimador: A lo largo de los capítulos anteriores se han desarrollado una serie de demostraciones que han permitido verificar la existencia de numerosas ventajas y propiedades de la media aritmética:

  a) La suma de los desvíos respecto a ella es igual a cero.

  b) La suma de los cuadrados de los desvíos respecto a ella es un mínimo.

  c) Puede ser considerada una variable en el campo de la teoría de las muestras.

  d) Su distribución tiende a ser normal cuando el tamaño de la muestra $n \to \infty$.

  e) La media poblacional de su distribución muestral es igual a la media poblacional de la variable x_i, es decir que es no viciada.

  f) La dispersión de su distribución muestral es menor que correspondiente a la distribución de la variable x_i, es decir que es $\sigma_{\overline{x}}^2 \leq \sigma_x^2$.

- Caso del estimador S_x^2. Partiendo de la igualdad:

$$x_i - \overline{x} = x_i - \overline{x} + \mu_x - \mu_x$$

agrupando, elevando al cuadrado y sumando para todo i:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}\big[(x_i - \mu_x) - (\overline{x} - \mu_x)\big]^2$$

dividiendo por n se obtiene la expresión de la variancia muestral:

$$\begin{aligned} S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}\big[(x_i - \mu_x) - (\overline{x} - \mu_x)\big]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\big[(x_i - \mu_x)^2 - 2(x_i - \mu_x)(\overline{x} - \mu_x) + (\overline{x} - \mu_x)^2\big] \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 - \frac{2}{n}(\overline{x} - \mu_x)\sum_{i=1}^{n}(x_i - \mu_x) + \frac{1}{n}n(\overline{x} - \mu_x)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 - 2(\overline{x} - \mu_x)\left(\sum_{i=1}^{n}\frac{x_i}{n} - \frac{1}{n}n\mu_x\right) + (\overline{x} - \mu_x)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 - 2(\overline{x} - \mu_x)^2 + (\overline{x} - \mu_x)^2 \end{aligned}$$

por lo tanto,

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 - (\overline{x} - \mu_x)^2$$

Si se calcula la esperanza matemática de esta última igualdad:

$$\begin{aligned} E(S_x^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 - (\overline{x} - \mu_x)^2\right] \\ = \frac{1}{n}\sum_{i=1}^{n}E\big[(x_i - \mu_x)^2\big] &- E\big[(\overline{x} - \mu_x)^2\big] \end{aligned}$$

como $E\big[(x_i - \mu_x)^2\big] = \sigma_x^2$ y $E\big[(\overline{x} - \mu_x)^2\big] = \sigma_{\overline{x}}^2 = \dfrac{\sigma_x^2}{n}$,

$$E(S_x^2) = \frac{1}{n}\sum_{i=1}^{n}\sigma_x^2 - \frac{\sigma_x^2}{n} = \sigma_x^2 - \frac{\sigma_x^2}{n} = \sigma_x^2\left(\frac{n-1}{n}\right) \tag{6.7}$$

con lo cual se verifica que el estimador S_x^2 es un estimador viciado o sesgado: su esperanza matemática no da un resultado igual al parámetro a estimar, dado que

se obtuvo el parámetro a estimar acompañado por un coeficiente que, precisamente, convierte al estimador en viciado.

Se verifica también que el estimador S_x^2 estima al parámetro σ_x^2 por defecto, ya que el coeficiente que acompaña al estimador da un resultado menor que 1. Ahora bien, dado que S_x^2 es un estimador viciado ¿puede corregirse el vicio? Analizaremos la posibilidad de realizar tal corrección. Partiendo de la ecuación (6.7),

$$E(S_x^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right] = \sigma_x^2\left(\frac{n-1}{n}\right)$$

luego,

$$\left(\frac{n}{n-1}\right)E(S_x^2) = E\left[\left(\frac{n}{n-1}\right)\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right] = \sigma_x^2$$

con lo cual se verifica que

$$E\left[\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(x_i - \overline{x})^2\right] = \sigma_x^2$$

es la esperanza matemática de una nueva estadística, $\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(x_i - \overline{x})^2$, cuyo resultado es el parámetro variancia poblacional, es decir, es una estadística *no viciada*.

Se comprueba que esta nueva estadística tiene forma de variancia, sólo que en lugar de estar dividida por n lo está por $n-1$. Debido a esto se la denomina *variancia corregida* y se la simboliza con S_c^2. Por consiguiente, se ha comprobado que S_c^2 es insesgada o no viciada, por cuanto su esperanza es el parámetro a estimar, es decir que

$$E(S_C^2) = E\left[\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(x_i - \overline{x})^2\right] = \sigma_x^2$$

Cabe mencionar que no siempre es necesario efectuar la corrección del vicio en el caso del estimador S_x^2 ya que cuando n crece indefinidamente, el término $\frac{n}{n-1}$ tiende a la unidad. Es decir

$$\text{si} \quad n \to \infty \Rightarrow \frac{n-1}{n} \to 1$$

Empíricamente se considera que si el tamaño de la muestra n es menor o igual a 30 se está trabajando con las denominadas "muestras pequeñas", en cuyo caso debe efectuarse la corrección, transformando S_x^2 en S_c^2. En cambio, si $n > 30$ se está trabajando con "muestras grandes", en cuyo caso no es necesario corregir el vicio. En este último caso además, todas las distribuciones de las estadísticas tienden a la distribución normal, según el teorema central del límite.

En todos los ejemplos planteados a lo largo de este texto, cuando se debió calcular una medida de dispersión se obtuvo el S_x^2. En caso de resultar necesario el cálculo de

la variancia corregida S_c^2, es posible obtener una expresión matemática que permita relacionar las ecuaciones

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{y} \quad S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

En efecto, vemos que

$$nS_x^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{y} \quad (n-1)S_c^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

donde sus segundos miembros son iguales por lo que podemos igualar

$$nS_x^2 = (n-1)S_c^2$$

obteniendo la estadística S_c^2 a partir de S_x^2 de acuerdo a

$$S_c^2 = \frac{n}{n-1}S_x^2$$

Estimación por intervalos de confianza

La estimación por intervalos de confianza permite, a partir de un estimador puntual, encontrar dos valores que limitan un intervalo, $l \leq \theta \leq u$, denominado *intervalo de confianza*, dentro del cual se encuentra el parámetro θ a estimar con una cierta probabilidad conocida denominada *nivel de confianza*. Es decir,

$$P(l \leq \theta \leq u) = \text{NC} = 1 - \alpha$$

donde l y u son los límites de confianza inferior y superior, respectivamente, y $\text{NC} = 1 - \alpha$ es el nivel de confianza, siendo α conocido como "nivel de significación". La interpretación de un intervalo de confianza es que, si se obtiene un número muy grande (o infinito) de muestras aleatorias y se calcula un intervalo de confianza para un parámetro desconocido θ en cada una de las muestras, el $100(1-\alpha)$ por ciento de los intervalos contendrán el valor verdadero de θ. Por ejemplo, si se analizan los intervalos de confianza para la media poblacional μ de 15 muestras aleatorias, figura(6.3), puede darse el caso de que uno de los 15 intervalos no contenga al verdadero valor de μ. Si el nivel de confianza fuera del 95 %, esto significaría que en una cantidad de muestras muy grande, sólo el 5 % de los intervalos de confianza no contendría a μ.

Ahora, en la práctica, sólo se obtiene una muestra aleatoria, de la cual se construye un intervalo de confianza. Puesto que este intervalo puede o no contener el valor verdadero de θ, no es posible asociar un nivel de probabilidad a este evento específico. La proposición adecuada es que el intervalo observado $[l, u]$ contiene al verdadero valor de θ con una confianza de $(1-\alpha)$. Esto indica que la proposición dada tiene una interpretación "frecuencial", es decir, no se sabe si es correcta para la muestra en particular pero el método utilizado en la construcción del intervalo $[l, u]$ permite obtener proposiciones correctas el $100(1-\alpha)$ por ciento de las veces.

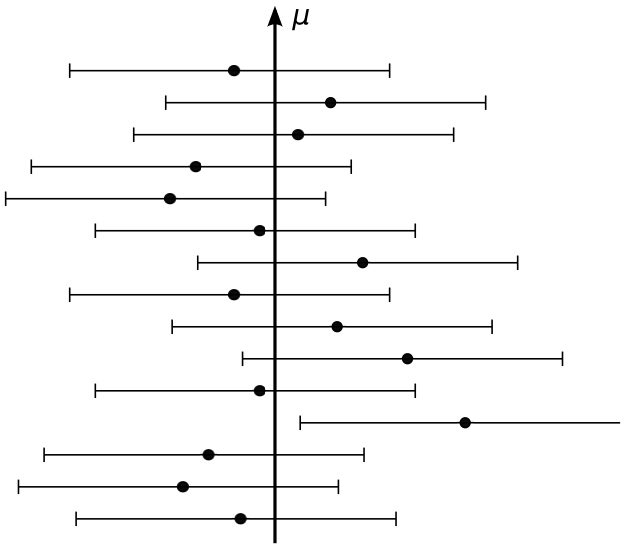


Figura 6.3: Sucesivos intervalos de confianza para la media poblacional. Los puntos medios de cada intervalo indican la estimación puntual de μ ($\overline{x}$).

Para su determinación deben plantearse dos situaciones completamente diferentes:

a) Estimación en el caso de muestras grandes.

b) Estimación en el caso de muestras pequeñas.

Algo que es común en la construcción de intervalos de confianza, cualquiera sea el tamaño de la muestra, es que en primer lugar debe fijarse el nivel de confianza, determinada por quien encarga el cálculo de la estimación y no por quien construye el intervalo. Los valores más comunes (aunque no los únicos) para el nivel de confianza son 0,99, 0,95 y 0,90.

Caso de muestras grandes

Las condiciones que se deben tener en cuenta en este caso son:

1) La muestra tiene un tamaño $n > 30$.

2) Todas las estadísticas son variables y, como tales, tienen una determinada distribución que tiende a ser normal cuando n es grande.

3) Como n es grande, no es necesario corregir el vicio del estimador S_x^2.

Media poblacional: como la variable que se utiliza para estimar la media poblacional es la media muestral $\overline{x}$ y ella tiene distribución normal, es posible construir el gráfico de la figura (6.4) donde se observa la distribución normal de la variable $\overline{x}$ con media poblacional $\mu_{\overline{x}} = \mu_x$ y desvío estándar $\sigma_{\overline{x}} = \sigma_x/\sqrt{n}$.

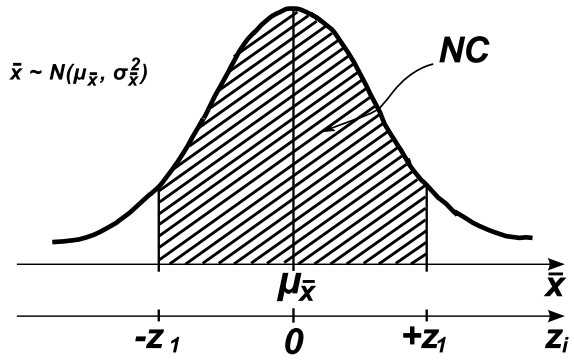


Figura 6.4: Intervalo de confianza para la media poblacional (muestras grandes).

Para poder trabajar con esta distribución normal, estandarizamos la variable $\overline{x}$:

$$z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}$$

El nivel de confianza, NC, se ubica en el centro de la figura (6.4) y una vez determinado su valor se verifica que existen dos valores de la variable estandarizada z_i, simétricos entre sí ($-z_1$ y $+z_1$) tales que la $P(-z_1 \leq z_i \leq +z_1) = \text{NC}$. Reemplazando en este término la variable z obtenemos la expresión:

$$P\left(-z_1 \leq \frac{\overline{x} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}} \leq +z_1 \right) = NC$$

la cual puede ser escrita como

$$P\left(-z_1 \frac{\sigma_x}{\sqrt{n}} \leq \overline{x} - \mu_x \leq +z_1 \frac{\sigma_x}{\sqrt{n}} \right) = \text{NC}$$
$$\Rightarrow P\left(-\overline{x} - z_1 \frac{\sigma_x}{\sqrt{n}} \leq -\mu_x \leq -\overline{x} + z_1 \frac{\sigma_x}{\sqrt{n}} \right) = \text{NC}$$

por lo cual,

$$P\left(\overline{x} + z_1 \frac{\sigma_x}{\sqrt{n}} \geq \mu_x \geq \overline{x} - z_1 \frac{\sigma_x}{\sqrt{n}} \right) = \text{NC}$$

Esta ecuación es una primera expresión para el intervalo de confianza, la cual está compuesta por los siguientes elementos:

- $\overline{x}$: se trata del estimador puntual media muestral, la cual se obtiene a partir de la muestra disponible.
- $\pm z_1$: se trata de dos valores simétricos que se obtienen a partir de la tabla de distribución normal estandarizada una vez fijado el valor de NC.
- n: tamaño de la muestra.

σ_x: desvío estándar poblacional, parámetro que al no ser conocido se lo reemplaza directamente por su estimador S_x sin efectuar corrección alguna por tratarse de muestras grandes.

En consecuencia, la expresión para la estimación de la media poblacional utilizando intervalos de confianza resulta:

$$P\left(\overline{x} - z_1 \frac{S_x}{\sqrt{n}} \leq \mu_x \leq \overline{x} + z_1 \frac{S_x}{\sqrt{n}}\right) = \text{NC} \tag{6.8}$$

Algunas características de los intervalos de confianza.

1) El intervalo de confianza tiene dos límites que se obtienen sumando y restando un mismo valor al estimador puntual media muestral $\overline{x}$. Estos límites se denominan límite superior e inferior del intervalo de confianza, respectivamente.

2) Si el nivel de confianza aumenta, su superficie en el gráfico será mayor y eso se corresponderá con mayores valores para los $\pm z_1$. Es decir, a mayor NC, mayor amplitud del intervalo de confianza. Sin embargo, una mayor amplitud en el intervalo de confianza implica que hay más valores posibles para estimar la media poblacional μ_x, lo que convierte a la estimación en algo menos precisa. Luego, **a mayor nivel de confianza, menor precisión en la estimación**.

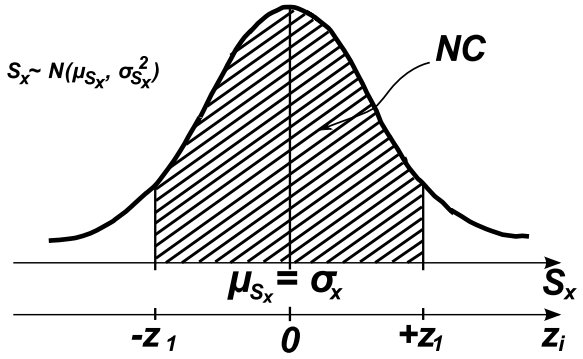
3) Si el nivel de confianza llegara a tomar el valor extremo máximo para una probabilidad, es decir igual a 1, el valor de los $\pm z_1 \to \pm\infty$. En ese caso no sería posible obtener resultados para los límites del intervalo de confianza porque darían un resultado indefinido.

4) La decisión de tomar el nivel de confianza entre dos valores simétricos de z_1 posibilita la búsqueda inversa en la tabla, dado que de lo contrario existirían infinitas combinaciones de valores de z para un mismo NC. Por otro lado, esto también conduce a obtener un *intervalo mínimo*, ya que el intervalo de confianza conseguido es más pequeño que cualquier otro que pueda obtenerse tomando los valores de z_1 de cualquier otra forma diferente.

5) El nivel de confianza es una probabilidad y como tal, según el planteo pascaliano, es el resultado de realizar un cociente entre el número de casos favorables sobre el número de casos posibles. Recordando este concepto, puede decirse entonces que de cada cien intervalos que se construyan, en una proporción de ellos igual a NC, el parámetro quedará encerrado en el intervalo construido. Esta es una forma de medir la confianza existente de que en un porcentaje de los casos se estime correctamente el parámetro desconocido.

Variancia poblacional: Para construir el intervalo de confianza para este parámetro partiremos del intervalo de confianza para el desvío estándar poblacional. El estimador puntual de este último parámetro es el desvío estándar muestral S_x el cual tiene una distribución normal con media poblacional $\mu_{S_x} = \sigma_x$ y variancia poblacional $\sigma^2_{S_x} = \frac{\sigma^2_x}{2n}$, con lo cual es posible construir la siguiente variable estandarizada:

$$z_i = \frac{S_x - \mu_{S_x}}{\sigma_{S_x}} = \frac{S_x - \sigma_x}{\frac{\sigma_x}{\sqrt{2n}}}$$



Luego, procediendo de la misma manera que en el caso anterior,

$$P\left(-z_1 \leq \frac{S_x - \sigma_x}{\frac{\sigma_x}{\sqrt{2n}}} \leq +z_1 \right) = NC$$

$$\Rightarrow P\left(S_x + z_1 \frac{\sigma_x}{\sqrt{2n}} \geq \sigma_x \geq S_x - z_1 \frac{\sigma_x}{\sqrt{2n}} \right) = NC$$

donde, reemplazamos en los límites del intervalo el valor del parámetro σ_x por su estimador S_x:

$$P\left(S_x + z_1 \frac{S_x}{\sqrt{2n}} \geq \sigma_x \geq S_x - z_1 \frac{S_x}{\sqrt{2n}} \right) = NC \tag{6.9}$$

Este intervalo se puede convertir en un intervalo para estimar la variancia poblacional simplemente elevando al cuadrado los términos de la desigualdad:

$$P\left[\left(S_x + z_1 \frac{S_x}{\sqrt{2n}} \right)^2 \geq \sigma_x^2 \geq \left(S_x - z_1 \frac{S_x}{\sqrt{2n}} \right)^2 \right] = NC \tag{6.10}$$

Caso de muestras pequeñas

Las condiciones que se deben tener en cuenta en este caso son:

1) La muestra tiene un tamaño $n \leq 30$.

2) Todas las estadísticas son variables y tienen una distribución que no necesariamente es normal.

3) Como n es pequeño, es necesario corregir el vicio del estimador S_x^2.

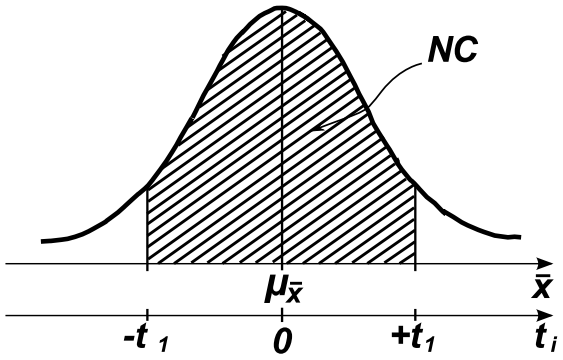
Media poblacional: Si el tamaño de la muestra es relativamente grande ($n > 30$) se ha visto que es posible estimar el valor de la variancia poblacional mediante el valor calculado

de la variancia muestral. Cuando el tamaño de la muestra es pequeño, es necesario realizar una hipótesis más fuerte con respecto a la población de interés. La hipótesis usual es que la población está distribuida de manera normal, por lo cual, para la construcción de este intervalo de confianza se utiliza la variable t de Student.

$$t = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu_x}{\frac{S_x}{\sqrt{n-1}}} = \frac{\overline{x} - \mu_x}{\frac{S_c}{\sqrt{n}}}$$

Fijado el nivel de confianza, existirán dos valores de la variable t, iguales en valor absoluto pero de distinto signo tales que:

$$P(-t_1 \leq t \leq +t_1) = \text{NC}$$



Reemplazando la variable t por una de las expresiones de la ecuación (6.2) donde interviene S_x:

$$P\left(-t_1 \leq \frac{\overline{x} - \mu_x}{\frac{S_x}{\sqrt{n-1}}} \leq +t_1 \right) = \text{NC}$$

Despejando la media poblacional que debe ser estimada:

$$P\left(\overline{x} + t_1 \frac{S_x}{\sqrt{n-1}} \geq \mu_x \geq \overline{x} - t_1 \frac{S_x}{\sqrt{n-1}} \right) = \text{NC} \tag{6.11}$$

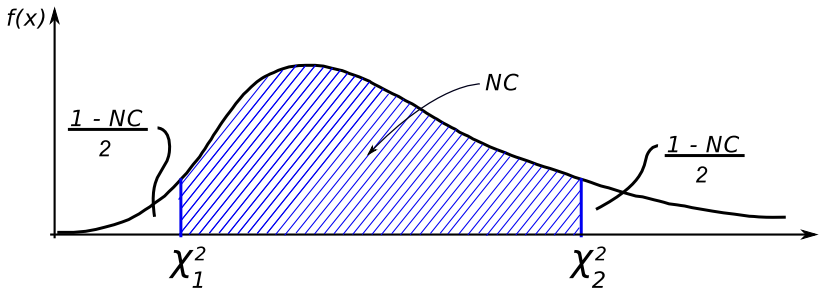
Expresión compuesta por un conjunto de elementos conocidos.

Variancia poblacional: Para la construcción de este intervalo de confianza se utiliza la estadística S_x^2 cuya distribución se obtiene recordando que

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \Rightarrow nS_x^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

Dividiendo ambos miembros de la última igualdad por un valor constante como σ_x^2, se obtiene la variable chi-cuadrado, χ^2 con ν grados de libertad.

$$\frac{nS_x^2}{\sigma_x^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sigma_x^2} = \chi_\nu^2$$



Fijado el NC, existen dos valores de la variable χ^2 tales que la

$$P\left(\chi_1^2 \leq \chi_\nu^2 \leq +\chi_2^2\right) = NC$$

con lo cual,

$$P\left(\chi_1^2 \leq \frac{nS_x^2}{\sigma_x^2} \leq +\chi_2^2\right) = NC$$

e invirtiendo la ecuación, se obtiene:

$$P\left(\frac{1}{\chi_1^2} \geq \frac{\sigma_x^2}{nS_x^2} \geq +\frac{1}{\chi_2^2}\right) = NC$$
$$\Rightarrow P\left(\frac{nS_x^2}{\chi_1^2} \geq \sigma_x^2 \geq +\frac{nS_x^2}{\chi_2^2}\right) = NC$$

Con lo cual se ha encerrado el parámetro a estimar, σ_x^2, entre dos límites con una cierta probabilidad NC, obteniéndose de ese modo el intervalo de confianza requerido. Este intervalo resultará mínimo si las dos superficies que quedan fuera del NC bajo la curva son iguales a $\frac{1-NC}{2}$.

Cálculo del tamaño de la muestra

Poblaciones infinitas o bajo muestreo con reposición

Se sabe que la media muestral no es igual a la media poblacional, es decir:

$$\overline{x} \neq \mu_x$$

y a la diferencia entre ambas la simbolizaremos con d, positiva o negativa, denominada *margen de error* o *tolerancia*,

$$\overline{x} - \mu_x = d$$

Dividiendo ambos miembros de esta igualdad por $\sigma_{\overline{x}}$:

$$\frac{\overline{x} - \mu_x}{\sigma_{\overline{x}}} = \frac{d}{\sigma_{\overline{x}}} = z_I$$

siendo que $\sigma_{\overline{x}} = \dfrac{\sigma_x}{\sqrt{n}}$, se tiene

$$\frac{\overline{x} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}} = \frac{d}{\frac{\sigma_x}{\sqrt{n}}} = z_I$$

tomando los dos últimos miembros de la igualdad y despejando n:

$$n = \frac{z_I^2 \ \sigma_x^2}{d^2} \tag{6.12}$$

Observando la expresión (6.12) se verá que el tamaño de la muestra depende de los siguientes factores:

a) de la variancia de la variable bajo estudio (σ_x^2) en forma directa, con lo cual a mayor variabilidad de la variable bajo estudio, mayor tamaño de la muestra.

b) del valor z_I, el cual indica el grado de confianza exigido en la estimación, también en forma directa, por lo cual a mayor grado de confianza exigido, mayor tamaño de la muestra. Si el grado de confianza fuera igual a uno, z_I tendería a infinito, por lo que el tamaño de la muestra $n \to \infty$.

c) del valor de la tolerancia d, en forma inversa: a mayor margen de error o tolerancia admitida, menor tamaño de la muestra. Si la tolerancia fuera cero estaríamos implicando que la diferencia entre la media muestral y poblacional debería ser cero, es decir iguales, por lo cual n deberá ser igual a N (infinito).

Poblaciones finitas o bajo muestreo sin reposición

El procedimiento en este caso es idéntico al anterior, sólo que al momento de reemplazar $\sigma_{\overline{x}}$ se recurre a la fórmula del desvío estándar para poblaciones finitas o sin reposición, con lo cual se obtiene:

$$\frac{d}{\frac{\sigma_x}{\sqrt{n}} \frac{\sqrt{N-n}}{\sqrt{N-1}}} = z_I$$

y despejando n:

$$n = \frac{z_I^2 \ \sigma_x^2 N}{d^2(N-1) + z_I^2 \sigma_x^2}$$

Puede verificarse fácilmente que si en esta última ecuación $N \to \infty$ se obtiene la ecuación (6.12).

Capítulo 7
Teoría de la decisión estadística

7.1. Introducción

En el desarrollo de la teoría de la estimación del capítulo anterior hemos visto cómo construir intervalos de confianza, con el objetivo de estimar parámetros (los cuales resultaban desconocidos) mediante la investigación muestral. Sin embargo, en muchos problemas de la ingeniería es necesario aceptar o rechazar un *enunciado* o *declaración* sobre algún parámetro en particular. Este enunciado o declaración se denomina **hipótesis** y al procedimiento de tomar decisiones frente a la hipótesis planteada se lo conoce como **prueba de hipótesis**. En ingeniería esta parte de la inferencia estadística es una de las más útiles dado que muchos problemas de toma de decisiones, ensayos o experimentos pueden formularse mediante pruebas de hipótesis. Además y tal como lo veremos más adelante, existe una conexión directa entre las pruebas de hipótesis y los intervalos de confianza.

Definición

La **teoría de la decisión estadística** constituye la parte de la inferencia estadística que permite decidir, sobre la base de resultados muestrales y con los procedimientos estadísticos y matemáticos apropiados, acerca de la validez de algún supuesto que se formule respecto del valor de un parámetro.

En la teoría de la estimación no se sabe cuál es el valor de un parámetro y se intenta estimarlo mediante una investigación muestral mientras que en la teoría de la decisión se supone un valor determinado para un parámetro y mediante una investigación muestral se trata de probar si ese supuesto es correcto.

Definición

Las **pruebas de hipótesis** son los procedimientos estadísticos apropiados que permiten probar la validez de cualquier supuesto formulado respecto del valor de un parámetro bajo estudio.

Las pruebas de hipótesis estadísticas junto con la estimación de parámetros poblacionales mediante intervalos de confianza son métodos fundamentales en la realización de

experimentos. En estos experimentos, por ejemplo, el ingeniero pueda estar interesado en contrastar la media poblacional con algún valor específico. Esta comparación entre experimentos es utilizada con mucha frecuencia en la práctica y sienta las bases de todo otro campo de la estadística conocido como "diseños experimentales".

Definición

Las **hipótesis estadísticas** son supuestos formulados respecto del valor de algún parámetro de una población.

Existen dos tipos de hipótesis estadísticas:

Hipótesis nula ($\boldsymbol{H_0}$): es la hipótesis concreta que se formula acerca del valor de un parámetro y que consiste en suponer que el parámetro toma un valor determinado. Se denomina así porque el propósito del estudio es anularla o rechazarla.

Hipótesis alternativa ($\boldsymbol{H_1}$), la cual constituye una hipótesis diferente a la hipótesis nula.

Por ejemplo podemos pensar que poseemos una máquina que elabora 400 productos por hora. Un nuevo modelo sale al mercado con, según informa el fabricante de la misma, una mejora tecnológica que permite aumentar la cantidad de productos elaborados. Ante ello es pertinente conocer cuál es la cantidad de productos que producirá la nueva máquina, lo cual puede llevar a dos situaciones posibles: *1)* el fabricante proporciona una información precisa y concreta diciendo, por ejemplo, que la mejora introducida ahora permitirá fabricar 500 o 600 productos por hora; o *2)* el fabricante no tiene definida la cantidad esperada de producción y sólo indica un aumento en la misma, sin poder establecer la magnitud de ese aumento.

Supongamos ahora que el vendedor de la máquina no puede informar exactamente cuál será el valor de la producción promedio del nuevo modelo y sólo asegura un aumento en la producción. Decidimos entonces efectuar una prueba de su funcionamiento, la cual consiste en hacer funcionar la nueva máquina durante 36 horas, arrojando los siguientes resultados:

$$\overline{x} = 490 \ \text{ productos/hora}$$

$$\sum_{i=1}^{36} x_i^2 = 10717200 \ \ (\text{productos/hora})^2$$

De acuerdo a estos datos, ¿conviene comprar la máquina? Cabe destacar que mediante esta pregunta estamos intentando resolver la cuestión de cuán verdadero es el supuesto de que la máquina nueva produce en mayor cantidad, es decir que intenta establecer si el nuevo modelo es diferente al anterior o sigue siendo el mismo producto. Ahora bien, es importante observar que el valor "400 productos por hora" que la máquina anterior producía constituye un *promedio poblacional*, $\boldsymbol{\mu}$, ya que se trata del promedio producido por la máquina en su versión anterior a lo largo de todo el tiempo en que la misma prestó servicios y que el valor de "490 productos por hora" que surge luego de la prueba de 36 horas, constituye una *media muestral*, $\boldsymbol{\overline{x}}$.

Pasos operativos de la prueba de hipótesis

Media poblacional (para muestras grandes y pequeñas)

En este caso precisamos realizar una prueba de hipótesis a un supuesto realizado sobre el parámetro μ. El desarrollo de los pasos del proceso operativo es el siguiente:

1^{ro} : Formulación de la hipótesis nula. La hipótesis nula se formula de la siguiente manera:

$$H_0)\ \mu = \mu_0$$

donde μ_0 constituye el valor concreto y específico que se asigna al parámetro. En el ejemplo analizado, si el vendedor de la máquina proporciona un valor tentativo para la producción de su nuevo modelo (500 productos por hora, por ejemplo) la hipótesis nula sería:

$$H_0)\ \mu = 500 \text{ prod/hora}$$

Como hemos supuesto que el vendedor no puede informarnos sobre cuál será la producción promedio por hora de la máquina en su nueva versión, la prueba de hipótesis debe realizarse para probar el valor del único parámetro conocido, es decir 400 productos por hora, correspondiente a la producción de la máquina en su versión anterior. La hipótesis nula se enunciaría entonces como:

$$H_0)\ \mu = 400 \text{ prod/hora}$$

Resulta oportuno indicar que así como se puede probar el supuesto de un aumento en la producción, también puede efectuarse una prueba para verificar una mejora en los tiempos de duración de un proceso productivo. En ese caso lo deseable sería que el tiempo de duración total del proceso *disminuya* en lugar de aumentar.

2^{do} : Formulación de la hipótesis alternativa. La hipótesis alternativa es cualquier hipótesis diferente a la nula. Puede adoptar una de las siguientes formas, todas excluyentes entre sí:

a) $\boldsymbol{H_1)\ \mu > \mu_0}$, "prueba unilateral a la derecha".

b) $\boldsymbol{H_1)\ \mu < \mu_0}$, "prueba unilateral a la izquierda".

c) $\boldsymbol{H_1)\ \mu \neq \mu_0}$, "prueba bilateral".

Las tres formas se aplican en diferentes circunstancias, resumidas en la siguiente tabla:

Si se desea probar	Parámetro nuevo desconocido	Parámetro nuevo conocido
Un aumento en la producción	$\mu > \mu_0$	$\mu < \mu_0$
Una disminución en los tiempos	$\mu < \mu_0$	$\mu > \mu_0$

Para una mejor comprensión de las alternativas planteadas, se efectuará una aplicación del cuadro anterior al ejemplo que se está desarrollando. Como se está probando el rendimiento de una nueva máquina cuyo vendedor la publicita diciendo que posee una producción mayor que la versión anterior pero sin dar a conocer el valor del nuevo parámetro, en la hipótesis nula se ha supuesto que esa nueva versión tiene la misma producción que la versión antigua. Por consiguiente, puede verse que estamos en el caso que se desea probar un aumento en la producción sin conocer el nuevo parámetro. Esto significa que la hipótesis alternativa tomará la forma del item *a)*, expresada como:

"se prueba la hipótesis de que la máquina produzca 400 productos por hora contra la alternativa de que la producción sea mayor".

Si en cambio se pudiera tener un valor concreto de producción (como sería en el caso de que fuera de 500 productos por hora) la prueba se realizaría teniendo un parámetro nuevo conocido, por lo que la hipótesis alternativa tomaría la forma *b)* y en ese caso se enunciaría:

"se prueba la hipótesis de que la máquina produzca 500 productos por hora contra la alternativa de que la producción sea menor".

Es decir, la hipótesis nula siempre se construye con el valor disponible del parámetro (sea nuevo o anterior) mientras que la hipótesis alternativa varía según cuál sea la forma adoptada por la nula y cuál sea el proceso que se desea probar: producción o tiempo, porque en el caso de probar una disminución en los tiempos de realización de un proceso las formas de las hipótesis alternativas se invierten, ya que en ese caso el tipo de prueba es opuesta a la de una mejora en la producción.

Por consiguiente, en el ejemplo planteado, la hipótesis alternativa tomará la forma

$$H_1)\ \mu > 400 \text{ prod/hora}$$

Finalmente, la prueba bilateral se utiliza cuando se desean probar hipótesis sobre parámetros que, a partir de un valor supuesto, pueden tener desvíos indistintamente hacia uno u otro lado. Este es el caso de la prueba de hipótesis sobre parámetros que no se refieren directamente a un aumento de producción o a una disminución de tiempos, sino a productos o procesos que deben cumplir con especificaciones concretas: una cierta longitud, un cierto espesor, peso, etc, y que serían rechazados en caso de tener desvíos en un sentido o en otro. Como ejemplo podemos imaginar la siguiente situación: se sabe que la temperatura corporal media de un adulto saludable es de $37\ ^\circ C$ pero queremos confirmar ese valor mediante una prueba de hipótesis. Para ello planteamos nuestra hipótesis nula, la cual tendrá la forma:

$$H_0)\ \mu = 37^\circ C$$

mientras que la hipótesis alternativa será:

$$H_1)\ \mu \neq 37°C$$

dado que no nos interesa probar si la temperatura media es mayor o menor, sino simplemente distinta.

3^{ro} : Determinación de los valores muestrales. Este paso está reservado a la efectiva realización de la investigación muestral, obteniendo los valores muestrales que permitirán construir la prueba de hipótesis. Para ello se decide cuál es el tamaño de la muestra (el cual en el ejemplo planteado es de 36 horas) y se pone a prueba la nueva máquina. Los datos recabados, como se mencionó en un principio, fueron los siguientes:

$$\begin{aligned} n &= 36 \text{ horas} \\ \overline{x} &= 490 \text{ prod/hora} \\ S_x^2 &= \frac{10717200}{36} - 490^2 = 57600 \text{ prod}^2/\text{hora}^2 \\ S_x &= 240 \text{ prod/hora} \end{aligned}$$

4^{to} : Elemento de comparación. Para verificar la validez del supuesto formulado compararemos el valor de la media poblacional μ_0 (que en este caso toma el valor de 400 prod/hora) con la media muestral $\overline{x}$ (que toma el valor de 490 prod/hora), tomando en cuenta que de esa comparación pueden surgir las siguientes alternativas:

a) Si la media muestral obtenida se encuentra "lejos" de la media poblacional supuesta, es muy posible que $\overline{x}$ no esté relacionada con μ_0, lo que equivale a decir que el nuevo proceso difiere del anterior y que se consiguió una verdadera mejora.

b) Si la media muestral se encuentra "cerca" de la media poblacional, es muy posible que $\overline{x}$ efectivamente esté relacionada con μ_0, lo cual implica que el nuevo proceso es muy similar al anterior y que no se logró mejorarlo.

El procedimiento para efectuar esta comparación consiste en construir una variable estandarizada z cuando $n > 30$, porque al estar trabajando con una muestra grande las estadísticas tienen distribución normal, en cuyo numerador aparece, precisamente, la diferencia entre $\overline{x}$ y μ_0:

$$z = \frac{\overline{x} - \mu_0}{\frac{S_x}{\sqrt{n}}}$$

o en su defecto la variable t de Student (si se tratara de $n \leq 30$ y trabajáramos con una muestra pequeña):

$$t = \frac{\overline{x} - \mu_0}{\frac{S_x}{\sqrt{n-1}}}$$

El cálculo de la variable estandarizada permite obtener un resultado numérico. En el caso del ejemplo planteado, como $n > 30$, se utiliza la variable z:

$$z = \frac{490 - 400}{\frac{240}{\sqrt{36}}} = 2{,}25$$

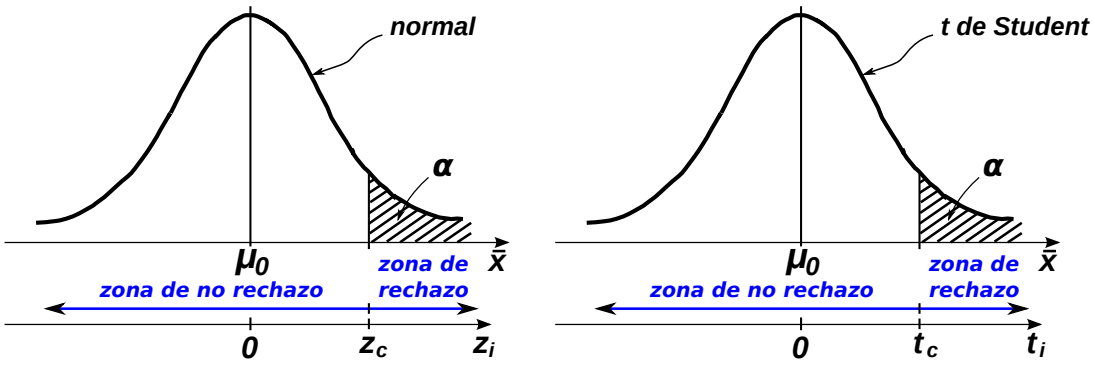
Al observar ese resultado se podrá verificar que el mismo será más grande o más pequeño según como sea la diferencia entre la media muestral y la media poblacional.

5^{to} : Criterio objetivo. Para poder establecer si el valor del estadístico de prueba (t o z) permite aceptar o rechazar la hipótesis nula, es preciso determinar objetivamente un valor límite. Para ello se debe elegir una probabilidad cercana a cero ($0{,}01, 0{,}05, 0{,}10$, entre los más comunes) denominada *"nivel de significación"*, simbolizado con α y que se representa como un área (como toda probabilidad en una función de densidad) que se ubica a la derecha, a la izquierda o a ambos lados (en ese caso con la mitad de α a cada lado) de la función densidad según como se haya definido la hipótesis alternativa.

Mediante esta probabilidad es posible encontrar en la tabla correspondiente (normal o t de Student, según sea el caso) un valor de la variable z o t denominado *"valor crítico"*, simbolizado con z_c o t_c, que divide el eje de las abcisas en dos zonas: la *"zona de rechazo"* y la *"zona de no rechazo"*, como se puede observar en la figura (7.1). Las ubicaciones de ambas zonas dependen de cómo se haya planteado la hipótesis alternativa.

Por otra parte, cabe señalar que el valor de α debe fijarse al comenzar la construcción de la prueba de hipótesis y quién lo fija debe ser quién encarga la prueba. De ese modo el valor crítico se transforma en el elemento objetivo que permite decidir respecto del criterio de "lejos" o "cerca".

Para el ejemplo analizado supondremos que se ha fijado $\alpha = 0{,}05$ con lo cual $z_c = 1{,}65$. En la figura (7.1) se ha considerado una prueba unilateral a la derecha tanto para el caso de muestras grandes utilizando la distribución normal como para muestras pequeñas, utilizando la distribución t de Student.



(a) Muestras grandes (b) Muestras pequeñas

Figura 7.1: Prueba unilateral a la derecha para distintos tamaños de muestras.

6^{to} : Regla de decisión. Para establecer la regla de decisión se compara el valor z calculado en el 4^{to} paso con z_c obtenido en el 5^{to} paso:

- si $|z| \geq |z_c| \Rightarrow z$ cae en la "zona de rechazo" y se considera que las diferencias entre $\overline{x}$ y μ_0 son significativas y por lo tanto *se rechaza la hipótesis nula.*
- si $|z| < |z_c| \Rightarrow z$ cae en la "zona de no rechazo" y se considera que las diferencias entre $\overline{x}$ y μ_0 no son significativas y por lo tanto *no se rechaza la hipótesis nula.*

Cabe aclarar que si se cuenta con una muestra pequeña la regla de decisión no se modifica, simplemente se trabaja con la distribución t de Student.

En el caso del ejemplo, como $z = 2{,}25 > z_c = 1{,}65 \Rightarrow z$ cae en la zona de rechazo y se debe rechazar la hipótesis nula, la cual fue formulada indicando que $\mu_0 = 400$ prod/hora. Ante esta decisión, ¿debe o no debe comprarse la nueva versión de la máquina? La respuesta es que, ante el análisis realizado, debería comprarse la nueva máquina.

Si se revisa la decisión tomada, se verá que rechazar la hipótesis nula implica que debe considerarse como válida la hipótesis alternativa, planteada haciendo $H_1)$ $\mu > 400$ prod/hora. Si esta hipótesis es considerada válida significa que la máquina en su nueva versión tiene un parámetro poblacional mayor que la versión anterior y que, en tal caso, conviene comprarla. En el lenguaje de la teoría de la decisión sólo resulta apropiado decir que una hipótesis nula se rechaza o no se rechaza, siendo incorrecto hablar de aceptar una hipótesis nula.

Análisis del nivel de significación

En los párrafos precedentes se ha mencionado que el nivel de significación es una probabilidad cercana a cero simbolizada con α pero, ¿qué representa esa probabilidad? Para responder adecuadamente esta pregunta conviene pensar que si bien la prueba realizada en el ejemplo de la máquina permitió rechazar la hipótesis nula que se postulaba, si ahora se supone por un instante que esa hipótesis era verdadera, queda claro que se ha cometido un error, dado que se ha rechazado un hipótesis nula que era verdadera. En otras palabras, si se supone que la nueva versión de la máquina no tiene un mejor nivel de producción pero la prueba realizada finaliza con una conclusión diferente y finalmente se compra la nueva máquina (como ha ocurrido en este caso), se comete un error.

Este es uno de los riesgos inevitables del muestreo, que en ciertas ocasiones, como consecuencia de errores normales propios de la investigación muestral, puede generar resultados equívocos. Por consiguiente, si una hipótesis nula que es verdadera se ha rechazado es debido a que el valor z o t calculado cayó en la zona de rechazo, lo cual induce al investigador a incurrir en el error de rechazar una hipótesis verdadera que realmente no debía ser rechazada.

Este tipo de error se denomina *error de tipo I* (e_I) de modo que α es, precisamente, la probabilidad de cometer ese tipo de error, es decir, es la probabilidad de rechazar una hipótesis que es verdadera,

$$\alpha = P(e_I)$$

Existe otro tipo de error que puede ser cometido por el investigador: no rechazar una hipótesis falsa, denominado *error de tipo II* (e_{II}) cuya probabilidad se simboliza con β. Como resulta imposible controlar simultáneamente ambos errores, en teoría de la decisión se determina que el único error que puede controlarse anticipadamente es el error de tipo I, por lo cual debe fijarse α con anterioridad a la realización de cualquier prueba de hipótesis. En conclusión, al realizar una prueba de hipótesis existen cuatro posibilidades, dos de las cuales conducen a una decisión errónea:

- No rechazar H_0 cuando H_0 es verdadera.
- Rechazar H_0 cuando H_0 es verdadera (error de tipo I).
- No rechazar H_0 cuando H_0 es falsa (error de tipo II).
- Rechazar H_0 cuando H_0 es falsa.

Prueba bilateral

Este tipo de prueba se utiliza cuando se desea probar el valor de aquellos parámetros que pueden tener desvíos tanto hacia la derecha como hacia la izquierda, lo que en general ocurre cuando los parámetros tienen que ver con el cumplimiento de determinadas especificaciones. Para comprender mejor el procedimiento se supone el siguiente caso: "se ha asegurado que el peso promedio de los alumnos de un colegio es de 54,4 Kg pero el profesor de gimnasia no cree que tal afirmación sea correcta y para verificarla selecciona una muestra aleatoria de 26 alumnos que da como resultado un peso promedio de 52 Kg y un desvío estándar de 5,4 Kg. Se desea probar la afirmación anunciada con un nivel de significación del 5 %".

La prueba debe realizarse de tal modo que cualquier desvío significativo, en más o en menos, constituya un motivo por el cual se deba rechazar la hipótesis planteada. Por consiguiente debe plantearse una prueba bilateral del siguiente modo:

$$\begin{aligned} &H_0)\ \mu_X = 54{,}4 \text{ Kg} \\ &H_I)\ \mu_X \neq 54{,}4 \text{ Kg} \\ &\quad n = 26 \Rightarrow \text{como se trata de una muestra pequeña} \\ &\qquad \text{se utilizará la distribución } t \text{ de Student} \\ &\quad \overline{x} = 52 \text{ Kg} \\ &\quad S_x = 5{,}4 \text{ Kg} \end{aligned}$$

obteniéndose t_I:

$$t_I = \frac{52 - 54{,}4}{\frac{5{,}4}{\sqrt{26-1}}} = \frac{-2{,}4}{\frac{5{,}4}{\sqrt{25}}} = \frac{-2{,}4}{1{,}08} = -2{,}22$$

Como la prueba es bilateral, el nivel de significación establecido $\alpha = 0{,}05$ debe dividirse por dos, de modo que corresponde ubicar $\alpha/2 = 0{,}025$ a cada lado de la distribución. En

este caso, el valor de t_c para un número de grados de libertad $\nu = 25$ es igual a 2,06, ver figura (7.2).

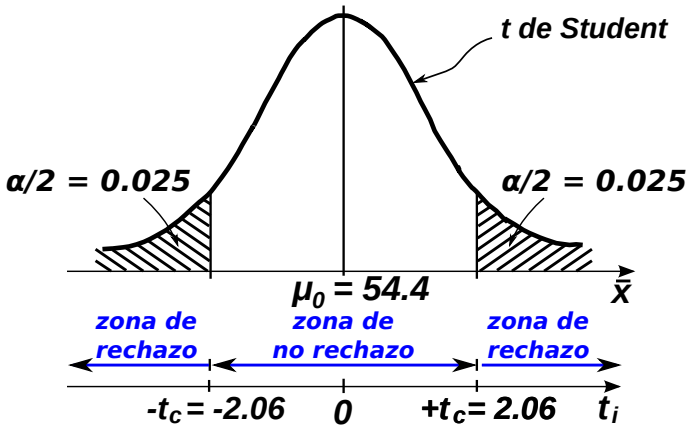


Figura 7.2: Prueba bilateral para muestras pequeñas.

Luego, comparando los valores de t disponibles, como $|t_I| > |t_c| \Rightarrow t_I$ cae en la zona de rechazo y en consecuencia se rechaza la hipótesis nula. Esto quiere decir que la prueba ha permitido encontrar diferencias significativas con un nivel de significación del 5 % y la postura que indicaba que los alumnos tenían un peso promedio de 54,4 kg debe modificarse.

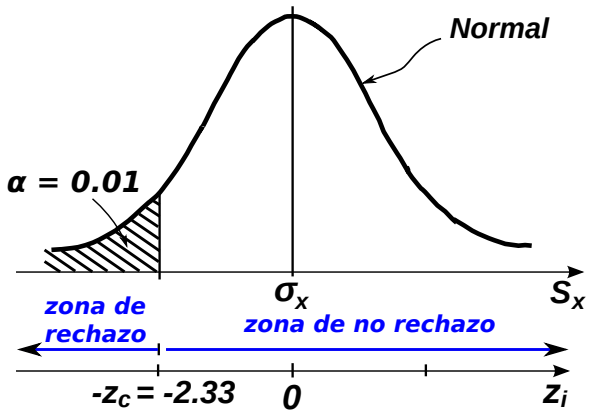
Prueba de hipótesis para la variancia poblacional

Muestras grandes (n>30)

El procedimiento que sigue la prueba de hipótesis para verificar la validez de un postulado respecto del valor de la variancia poblacional no tiene modificaciones fundamentales respecto de la estructura de la prueba que fue realizada para la media poblacional. En ambos casos deben cumplirse los seis pasos operativos y se utilizan aquellas funciones de densidad normal o χ^2 según estemos en presencia de muestras grandes o pequeñas, respectivamente.

Volviendo al problema de la máquina, si se desea probar ahora la hipótesis de que el desvío estándar sea igual a 300 prod/hora con un nivel de significación del 1 %, se procede de la siguiente manera:

$$
\begin{aligned}
H_0)\ & \sigma_X = 300 \text{ prod/hora} \\
H_I)\ & \sigma_X < 300 \text{ prod/hora} \\
& n = 36 \\
& S_x^2 = 57600 \text{ (prod/hora)}^2 \Rightarrow S_x = 240 \text{ prod/hora}
\end{aligned}
$$



Tratándose de una muestra de tamaño $n = 36$ (muestra grande), es posible considerar que las estadísticas tienen distribución normal, luego:

$$z_I = \frac{240 - 300}{\frac{240}{\sqrt{2 \cdot 36}}} = \frac{-60}{28{,}28} = -2{,}12$$

con lo cual, para un nivel de significación $\alpha = 0{,}01 \Rightarrow z_c = -2{,}33$, como $|z_I| < |z_c| \Rightarrow z_I$ cae en la zona de no rechazo, por lo cual no se rechaza H_0. Es decir, no se rechaza el supuesto de que el desvío estándar pueda tomar un valor igual a 300 prod/hora, con un nivel de significación del 1 %.

Muestras pequeñas (n≤30)

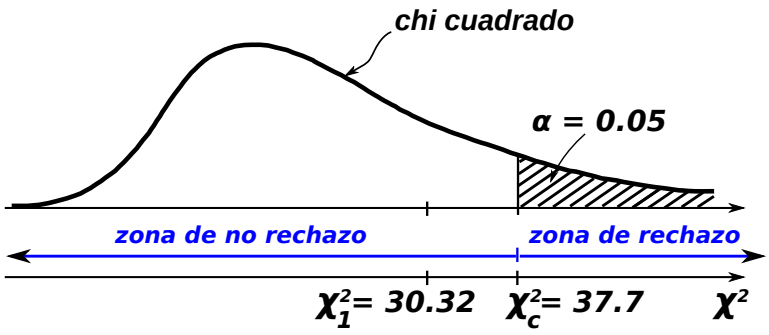
A partir de los datos del ejemplo del peso de los alumnos del colegio, se desea probar ahora que el valor de la variancia es igual a 25 Kg^2 con un nivel de significación del 5 %:

$$\begin{aligned} H_0)\ & \sigma_X^2 = 25\ \text{Kg}^2 \\ H_I)\ & \sigma_X^2 > 25\ \text{Kg}^2 \\ & n = 26 \\ & \overline{x} = 52\ \text{kg} \\ & S_x^2 = 5{,}4^2 = 29{,}16\ \text{Kg}^2 \end{aligned}$$

luego,

$$\chi_I^2 = \frac{n\ S_x^2}{\sigma_0^2} = \frac{26 \cdot 29{,}16^2}{25} = 30{,}32$$

Para un nivel de significación del 5 % en una prueba unilateral a la derecha y con los grados de libertad $\nu = 25$, el valor crítico es igual a $\chi_c^2 = 37{,}7$.



Como $\chi_I^2 < \chi_c^2$, significa que χ_I^2 cae en la zona de no rechazo, con lo cual no se rechaza la hipótesis nula. Por lo tanto, no se rechaza la hipótesis de que la variancia poblacional sea igual a 25 Kg^2 con un nivel de significación de 0,05.

.

Parte I

Apéndice 1

Apéndice A
Otras demostraciones

A.1. Media aritmética

En el Capítulo 3 se han enunciado las propiedades de la media aritmética y se demostró que la sumatoria de los desvíos al cuadrado (*error cuadrático*), entre los valores de la variable y un valor constante y arbitrario A, es mínimo si $A = \overline{x}$. Esta propiedad también es posible demostrarla de la siguiente manera:

Sea $k \neq \overline{x}$ un valor constante y arbitrario. Es posible demostrar que el error cuadrático de los valores x_i, $i = 1 \ldots n$, con respecto a la media aritmética $\overline{x}$ es siempre menor al error cuadrático de los valores x_i con respecto a k. Es decir:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 < \sum_{i=1}^{n}(x_i - k)^2 \qquad \text{si } k \neq \overline{x}$$

Demostración:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - k)^2 &= \sum_{i=1}^{n}[x_i - (k - \overline{x} + \overline{x})]^2 = \sum_{i=1}^{n}[(x_i - \overline{x}) - (k - \overline{x})]^2 \\ &= \sum_{i=1}^{n}(x_i - \overline{x})^2 - 2\sum_{i=1}^{n}(x_i - \overline{x})(k - \overline{x}) + \sum_{i=1}^{n}(k - \overline{x})^2 \\ &= \sum_{i=1}^{n}(x_i - \overline{x})^2 - 2(k - \overline{x})\underbrace{\sum_{i=1}^{n}(x_i - \overline{x})}_{=0} + n(k - \overline{x})^2\end{aligned}$$

Luego,

$$\sum_{i=1}^{n}(x_i - k)^2 > \sum_{i=1}^{n}(x_i - \overline{x})^2.$$

La última desigualdad es válida dado que $n(k - \overline{x})^2 > 0$ para cualquier valor de k excepto para $k = \overline{x}$, donde en ese caso $n(k - \overline{x})^2 = 0$ y se obtiene nuevamente $\sum_{i=1}^{n}(x_i - \overline{x})^2$.

A.2. Propiedades del operador sumatoria

a)

$$\sum_{i=1}^{n} x_i \; y_i = x_1 \; y_1 + x_2 \; y_2 + \ldots + x_n \; y_n$$

b)

$$\sum_{i=1}^{n} a \; x_i = a \; x_1 + a \; x_2 + \ldots + a \; x_n = a(x_1 + x_2 + \ldots + x_n) = a \sum_{i=1}^{n} x_i, \quad (a = \text{cte})$$

b) Si a, b y c son constantes:

$$\begin{aligned} \sum_{i=1}^{n} a \; x_i + b \; y_i + c \; z_i &= a \; x_1 + b \; y_1 + c \; z_1 + a \; x_2 + b \; y_2 + c \; z_2 + \ldots + a \; x_n + b \; y_n + c \; z_n \\ &= a(x_1 + x_2 + \ldots + x_n) + b(y_1 + y_2 + \ldots + y_n) + c(z_1 + z_2 + \ldots + z_n) \\ &= a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i + c \sum_{i=1}^{n} z_i \end{aligned}$$

Parte II

Apéndice 2

Apéndice B
Teoría de las muestras

B.1. Ejemplos

En el Capítulo 5 se realizado una introducción teórica a la Teoría de las Muestras. En esta sección el tema es presentado empíricamente tomando como modelo un ejemplo publicado en el texto de aplicaciones estadísticas prácticas titulado "Teoría y Problemas de Estadística" de Murray Spiegel. La resolución del ejemplo permitirá la obtención de ciertas conclusiones generales que serán indicadas en cada caso.

Distribución muestral de las medias

Caso con reposición o para poblaciones infinitas

Supongamos la existencia de una población de tamaño $N = 5$ que posee los siguientes valores de la variable aleatoria x_i:

$$x_1 = 2;\ x_2 = 3;\ x_3 = 6;\ x_4 = 8;\ x_5 = 11$$

Con los datos disponibles se calculan la media y la variancia poblacionales:

$$\mu_x = \frac{\sum x_i}{N} = \frac{2+3+6+8+11}{5} = 6$$

$$\sigma_x^2 = \frac{1}{N}\sum (x_i - \mu_x)^2 = \frac{1}{5}\left[(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2\right] = 10{,}8$$

y se construye el siguiente gráfico de la distribución, llamado "de bastones":

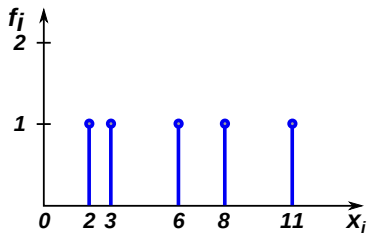


Figura B.1: Distribución de la variable x_i.

A continuación se procederá a efectuar la selección de muestras con reposición de tamaño $n = 2$. Como la selección de muestras con reposición impide que con la selec-

ción se agote el conjunto de elementos muestreados, el sistema es equivalente a trabajar con una población de tamaño infinito.

El siguiente cuadro contiene las muestras que se generan. Como se trata de todas las muestras posibles, puede considerarse que el cuadro contiene la población de muestras de tamaño $n = 2$ extraídas con reposición de una población de tamaño $N = 5$:

2-2	2-3	2-6	2-8	2-11
3-2	3-3	3-6	3-8	3-11
6-2	6-3	6-6	6-8	6-11
8-2	8-3	8-6	8-8	8-11
11-2	11-3	11-6	11-8	11-11

Como puede verificarse fácilmente, se han logrado construir 25 muestras diferentes de tamaño n con reposición. Si el tamaño de la muestra hubiera sido de 3 elementos, se hubieran podido construir 125 muestras diferentes; en cambio, si el tamaño de la población hubiera sido igual a 6 y se extraían muestras de tamaño $n = 2$, el total de muestras hubiera sido 36. Estas observaciones permiten obtener la:

1^{ra} conclusión: El total de muestras con reposición de tamaño n que pueden extraerse de una población de tamaño N es $M = N^n$.

Definido claramente que el cuadro anterior contiene todas las muestras posibles y que en total ellas son 25, puede plantearse la siguiente pregunta: ¿cuántas muestras de tamaño n deben extraerse en una aplicación real, cuando se desea realizar una investigación estadística mediante el muestreo con reposición? La respuesta a esta pregunta da lugar a la

2^{da} conclusión: Para realizar una investigación estadística por muestreo es suficiente una sola muestra de tamaño n.

Resulta oportuno aclarar, por supuesto, que la única muestra a la que se alude precedentemente debe estar bien seleccionada, para lo cual se han dado los lineamientos en el Capítulo 1 del presente apunte. A continuación se calcularán las medias muestrales de cada una de las M muestras seleccionadas. El siguiente cuadro contiene los 25 valores:

2	2.5	4	5	6.5
2.5	3	4.5	5.5	7
4	4.5	6	7	8.5
5	5.5	7	8	9.5
6.5	7	8.5	9.5	11

Se observa claramente que hay varios resultados para el valor de la media muestral, y que ellos dependen de cómo se encuentren conformadas las M muestras diferentes de tamaño $n = 2$. Los resultados presentados en el cuadro anterior constituyen el conjunto de medias muestrales posibles de ser calculadas a partir de la población de medias muestrales definidas previamente, lo que implica que el cuadro contiene una población de medias muestrales de

tamaño M, y ese conjunto se denomina "Distribución muestral de las medias". Esto da a la media muestral una característica no descubierta hasta este momento, que origina la:

3ra conclusión: la media muestral resulta ser una variable mientras la conformación de la muestra no se encuentre definida, lo cual no entra en contradicción con la primera propiedad de la media aritmética, que dice que ella es una constante para un **conjunto definido** de valores.

Esta conclusión puede ser generalizada del siguiente modo: cualquier cálculo muestral (es decir cualquier otra medida de posición e incluso de dispersión que se obtenga a partir de los datos muestrales) y no sólo la media aritmética debe considerarse como variable, por el mismo principio vigente para la media. Atendiendo al hecho de que la media muestral resulta una variable y que el cuadro precedente constituye la población de medias muestrales, se pueden obtener tanto la media poblacional como la variancia poblacional de la variable media muestral, es decir que se puede calcular $\mu_{\overline{x}}$:

$$\mu_{\overline{x}} = \frac{\sum \overline{x}}{M} = \frac{150}{25} = 6$$

lo cual permite formular la

4ta conclusión: la media de la población de medias muestrales es igual a la media de la variable x_i, es decir

$$\mu_{\overline{x}} = \mu_x \tag{B.1}$$

Para esa misma población de medias muestrales se calcula la variancia poblacional, es decir,

$$\sigma^2_{\overline{x}} = \frac{1}{M}\sum (x_i - \mu_{\overline{x}})^2 = \frac{135}{25} = 5{,}4$$

lo cual permite enunciar la

5ta conclusión: la variancia de la población de medias muestrales es igual a la variancia de la variable x_i dividida por n, es decir,

$$\sigma^2_{\overline{x}} = \frac{\sigma^2_x}{n} \quad \Rightarrow \quad \sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}} \tag{B.2}$$

Si se construye el gráfico de la Distribución muestral de las medias se obtiene la siguiente figura:

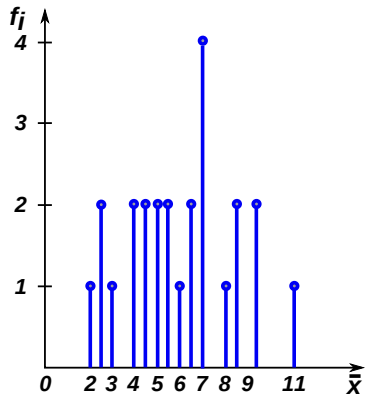


Figura B.2: Distribución de la variable $\overline{x}$.

en la cual se observa que:

a) los valores extremos de la variable media muestral son coincidentes con los de la variable x_i (en este caso son 2 y 11).

b) a medida que el tamaño n de la muestra crece, como la cantidad de muestras posibles aumentará considerablemente, aparecerán para la media muestral nuevos valores que siempre oscilarán entre los valores extremos ya determinados. Es decir que, en ese caso, la gráfica de bastones que se observa más arriba presentará nuevos valores y una mayor cantidad de bastones. En el límite, cuando n crezca indefinidamente, la variable media muestral se convertirá en continua y la gráfica de bastones se transformará en un área, lo que permite obtener la:

6^{ta} conclusión: en el muestreo con reposición, cuando $n \to \infty$, la variable media muestral se distribuye normalmente con parámetros $\mu_{\overline{x}} = \mu_x$ y $\sigma^2_{\overline{x}} = \frac{\sigma^2_x}{n}$.

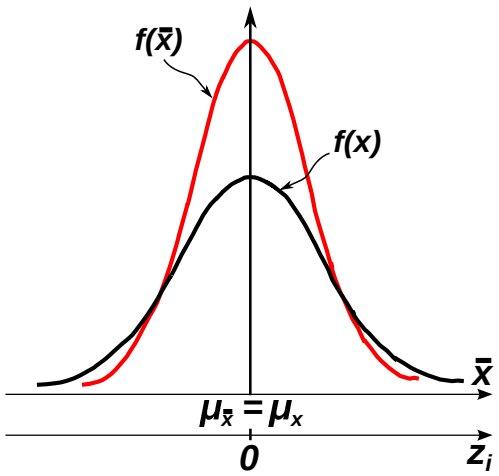


Figura B.3: Comparación de la distribución de la variable $\overline{x}$ *vs* la distribución de la variable x_i.

Es decir que $\overline{x} \sim N(\mu_x, \frac{\sigma^2_x}{n})$ si $n \to \infty$. Obsérvese que en esta última conclusión no se menciona para nada cuál es la distribución de la variable x_i, por lo cual se puede afirmar

que esta conclusión se cumple cualquiera sea la forma que toma la distribución de x_i. Si esa forma fuera normal, el gráfico de la figura (B.3) permite comparar cómo se verían tanto la distribución de las dos variables involucradas en este análisis. Esta última conclusión suele encontrarse en los libros de texto bajo la denominación de **Teorema Central del Límite**.

Caso sin reposición o para poblaciones finitas

Partiendo del ejemplo inicial en el cual la población tiene un tamaño $N = 5$, ahora se selecciona muestras de tamaño $n = 2$ sin reponer. El siguiente cuadro contiene todas las muestras posibles elegidas de acuerdo con ese procedimiento

2-3	2-6	2-8	2-11
	3-6	3-8	3-11
		6-8	6-11
			8-11

A partir del cuadro precedente se determina la

1^{ra} conclusión: El total de muestras sin reposición de tamaño n que pueden extraerse de una población de tamaño N es $\binom{N}{n}$.

Es decir que esta primera conclusión tiene una diferencia comparada con la indicada para el caso con reposición. El siguiente cuadro contiene las medias muestrales de cada una de las muestras que aparecen en el cuadro anterior:

2.5	4	5	6.5
	4.5	5.5	7
		7	8.5
			9.5

Las conclusiones 2^{da} y 3^{ra} no tienen modificación alguna en su texto, por lo que son válidas para los casos sin reposición. Ahora verificaremos cuál es el valor de la media poblacional de la variable media muestral.

$$\mu_{\overline{x}} = \frac{\sum \overline{x}}{M} = \frac{60}{10} = 6$$

con lo cual comprobamos que tampoco se modifica la cuarta conclusión, y que también en el caso sin reposición $\mu_{\overline{x}} = \mu_x$.

A continuación calcularemos la variancia de la variable media muestral,

$$\sigma^2_{\overline{x}} = \frac{1}{M}\sum (x_i - \mu_{\overline{x}})^2 = \frac{40{,}5}{10} = 4{,}05$$

lo cual permite decir que en el caso sin reposición la variancia de la variable aleatoria media muestral es diferente al caso con reposición. Por consiguiente, la 5^{ta} conclusión es:

$$\sigma^2_{\overline{x}} = \frac{\sigma^2_x}{n}\frac{N-n}{N-1} \quad \Rightarrow \quad \sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{B.3}$$

en la cual el coeficiente $\frac{N-n}{N-1}$ se denomina "factor de corrección para casos sin reposición", el cual se encuentra, según se ha visto en el capítulo 4, en la fórmula de la variancia de la Distribución Hipergeométrica, que como se ha visto es aplicable para casos sin reposición.

Finalmente, diremos que la 6^{ta} conclusión es similar a la señalada para el caso con reposición, es decir que en el muestreo sin reposición, cuando $n \to \infty$, la variable media muestral se distribuye normalmente con media poblacional $\mu_{\overline{x}} = \mu_x$ y variancia poblacional $\sigma^2_{\overline{x}} = \frac{\sigma^2_x}{n}\frac{N-n}{N-1}$. Luego,

$$\overline{x} \sim N\left(\mu_{\overline{x}}, \frac{\sigma^2_x}{n}\frac{N-n}{N-1}\right) \quad \text{si } n \to \infty \tag{B.4}$$