Universidad de los Andes
Facultad de Ciencias Económicas y Sociales
Escuela de Estadística
Departamento de Estadística
Prof. Gudberto León

Apuntes de Métodos Estadísticos I

Estadística Descriptiva

Junio 2013

Tema I: Introducción y Estadística Descriptiva[1]

Estadística: conceptos básicos y utilidad

Los conceptos y métodos que proporciona la estadística son de invaluable utilidad en la toma de decisiones ante *situaciones de incertidumbre*.
La estadística provee potentes herramientas analíticas que se emplean en una gran variedad de situaciones: en el gobierno, en la empresa privada, en los negocios, en la industria, en investigaciones: médicas, económicas, sociológicas, biológicas, agrícolas, genéticas, físicas, etc.

Definición de Estadística

La estadística es un conjunto de conocimientos y métodos que se utilizan en la recolección, organización, presentación y análisis de la información relativa a un fenómeno o hecho determinado y que le *permite al investigador tomar decisiones en situaciones donde está presente la incertidumbre*.

Como procedimiento para la toma de decisiones, la estadística se emplea hoy en día en toda clase de estudios científicos, siendo efectiva no solamente en los experimentos de laboratorio sino también lo es en estudios fuera de él.
La estadística es de mucha utilidad para dar respuesta, con justificación científica, a interrogantes como las siguientes:

- Cómo puede probar un gran laboratorio la eficiencia de un nuevo fármaco.
- Cómo el gobierno puede pronosticar la población para el año 2020 con fines de planificación en cuanto a seguridad social de los trabajadores.
- ¿Los cambios (disminución o crecimiento) en el índice de desempleo se deben a las políticas gubernamentales o a fluctuaciones estacionales?
- Para controlar la calidad de cierto artículo producido por una empresa, ¿cuántos de estos deben examinarse?
- Cómo es posible predecir el resultado de unas elecciones si solamente se entrevistan unos pocos votantes.
- ¿Existe relación entre el fumar y el cáncer del pulmón?
- Cómo medir y determinar los cambios (aumentos o disminuciones) en: los precios de los alquileres de viviendas, la inflación, el nivel de desempleo, el consumo de cierto producto, las muertes registradas los fines de semana, etc.

Nota:

1. La noción de "estadística" se derivó originalmente del vocablo "estado", porque ha sido función tradicional de los gobiernos centrales llevar registros de población, nacimientos, defunciones, vocaciones, cosechas, impuestos y muchas otras clases de cosas y actividades.

[1] Estos apuntes están basados en el libro: Armas, J. (1998). *Estadística Sencilla: Descriptiva*. Mérida: FACES-ULA.

2. Es importante en este momento hacer la siguiente aclaratoria, debido a la confusión de muchas personas en cuanto al significado de las palabras *estadístico* y *estadista*. Según el diccionario Larousse:
   Estadista: Político, persona que ejerce un alto cargo en la administración del estado.
   Estadístico: Persona que se ocupa de investigaciones estadísticas.

Universo Estadístico

Generalmente, existe un conjunto de elementos claramente definido en el que el investigador está interesado. Este conjunto se llama *universo*. Es un conjunto, finito o infinito de seres vivos, elementos o cosas, sobre las cuales están definidas características o variables que interesa analizar.
Los elementos individuales que conforman el universo se llaman **Unidades Elementales** (también se conocen como unidades individuales o unidades de observación) Las unidades elementales poseen las características de interés, las cuales pueden ser de naturaleza cuantitativa o cualitativa.

Ejemplo:

1. El Instituto de Investigaciones Ambientales lleva a cabo un estudio para determinar el grado de contaminación de los ríos en la ciudad de Mérida. Los elementos que poseen las características a estudiar son los ríos de Mérida y por tanto estos conforman el Universo Estadístico de esta investigación.
2. Un estudio sobre los ingresos mensuales de los hogares de la región andina es llevado a cabo por el instituto de investigaciones económicas de la ULA. El conjunto de elementos que poseen las variables a medir en el estudio, es decir, el universo, está conformado por todos los hogares de la región andina.
3. El Ministerio de Salud desea conocer si como consecuencia por el uso del teléfono celular, existen problemas de salud en los venezolanos. En este caso, el Universo estadístico está compuesto por las personas venezolanas que usan teléfono celular.
4. Se lleva a cabo una investigación para determinar la eficiencia en el consumo de combustible de los automóviles con caja dual de marcas asiáticas. Universo: _____
5. La oficina de registros estudiantiles de la ULA quiere llevar a cabo una investigación sobre el rendimiento estudiantil en el primer semestre de las carreras de la universidad. Universo:_____.

Población Estadística

La *población*, es un conjunto *de valores* asociados con los elementos del universo. Es la colección de todas las posibles mediciones que pueden hacerse de la característica en estudio.

Obsérvese que una *población estadística* es una colección de valores no una colección de personas.

Entonces, la población va a estar constituida por datos o valores y puede ser finita o infinita. Una *población finita* es aquella en la cual el número de elementos puede ser contado y es limitado. Una *población es infinita* si la cantidad de elementos que la componen es ilimitada o su composición es tal, que sus elementos no pueden ser contados. En la práctica, este concepto de infinito también expresa la idea de indeterminado o indefinido e incluso poblaciones finitas excesivamente grandes se les considera como infinitas.

Al número de elementos en la población se le denomina *tamaño de la población* y, en el caso finito, este tamaño se denota con la letra *N*.

Ejemplo:
En relación con los ejemplos de la página anterior se tiene que:

1. La población es el conjunto de valores que miden el grado de contaminación de los ríos de la ciudad de Mérida.
2. Los ingresos mensuales (en bolívares) de los hogares de la región andina conforman una población en esta situación
3. Población: _____
4. Población: _____
5. Población: _____

Nota:
De un mismo universo puede derivarse más de una población. Por ejemplo, del universo de estudiantes de la ULA podemos estar interesados en estudiar características tales como: Edad, sexo, ingreso y tipo de sangre. De cada una de estas cuatro características se origina una población, es decir que del universo de estudiantes de la ULA obtenemos la población de edades, la población de sexos, la población de ingresos y la población de tipos de sangre de los estudiantes de la ULA.
Nótese que en este ejemplo el universo está compuesto por personas (estudiantes) que son los que poseen las características de interés y las mediciones que se obtienen para cada una de estas características (edad, sexo, ingreso, tipo de sangre) constituyen las poblaciones. Así, una de estas poblaciones va a estar constituida por *N* números que representan cada una de las *N* edades de los estudiantes de la ULA. En la siguiente figura se puede observar de manera ilustrativa las relaciones explicadas arriba.

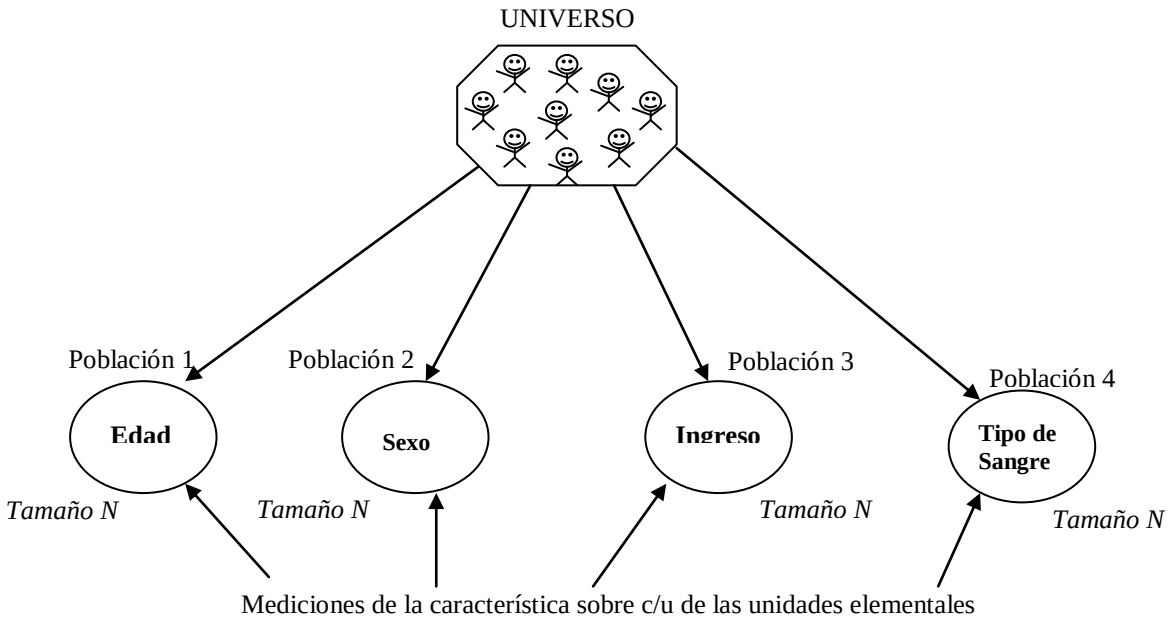


Figura 1. Ilustración didáctica de un universo con varias poblaciones

Muestra

Frecuentemente es imposible obtener o medir todos los valores en una población. Un subconjunto de valores de la población se conoce como una *muestra*. Es decir, una muestra es una parte de una población. De esta manera, como la población es un conjunto de mediciones de la característica bajo estudio, y la muestra es un subconjunto de la población, ésta va a estar constituida también por mediciones de la característica.

Así, una muestra está compuesta por ***n*** mediciones sobre las unidades elementales. En otras palabras, ***n*** representa el tamaño de la muestra y por lo tanto $n \leq N$. Fácilmente se puede deducir que de una misma población pueden seleccionarse diferentes muestras:

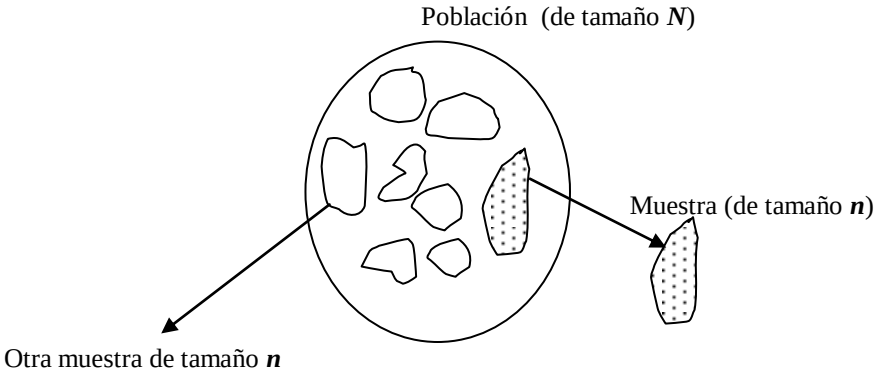


Figura 2. Ilustración de las muestras como subconjunto de una población

Ejemplo:
En el siguiente ejemplo determinar:

a. Unidades elementales
b. Universo
c. Población
d. Tipo de población (finita o infinita)
e. Muestra

Mediante un estudio se quiere conocer la opinión de los estudiantes de la ULA sobre el servicio de comedor que presta esta universidad. Con este fin se piensa entrevistar 500 estudiantes seleccionados aleatoriamente para conocer su opinión al respecto:

Característica en estudio: Opinión sobre el servicio del comedor.
Unidad elemental: Estudiante de la ULA
Universo: Todos los estudiantes de la ULA que asisten regularmente al comedor.
Población: La opinión de cada uno de los estudiantes de la universidad que asisten regularmente al comedor sobre el servicio de comedor.
Tipo de población: finita
Muestra: Las opiniones de los 500 estudiantes seleccionados al azar.

Censo

Se dice que se ha realizado un *censo* y se habla de enumeración completa, cuando una investigación es exhaustiva en el sentido de analizar toda la población estadística.

Muestreo

Cuando el estudio se hace sobre la base de una muestra de la población estadística, se habla de una *investigación por muestreo* o enumeración parcial.

Razones del uso del muestreo
Las razones que determinan la conveniencia de tomar muestras son entre otras las siguientes:

1. Menor costo que un censo
2. Mayor control en la recolección de la información y en consecuencia mejor calidad de la misma.

   En una muestra se puede dedicar más atención a la calidad de los datos, al entrenar al personal y realizar un seguimiento de quienes no contestan la encuesta. Es mucho mejor

tener buenas mediciones en una muestra representativa que mediciones poco confiables sobre toda la población.

3. Mayor rapidez en los resultados.
   Una estimación de la tasa de desempleo del año 2012 no es muy útil si para entrevistar a cada familia se tarda hasta el 2014.
4. El que la población sea excesivamente grande o infinita lo cual imposibilita cubrirla totalmente. Por ejemplo:
   a. Una evaluación de los recursos camaroneros del litoral venezolano
   b. O la evaluación de los recursos forestales de la región sur del estado Bolívar.
5. El que la población sea suficientemente homogénea.
   Este hecho permite que una muestra muy pequeña sea suficiente para inferir en la población con un margen de seguridad muy alto.
6. Que el proceso de medición sea auto destructivo en el sentido de ocasionar daño o pérdida de la unidad sobre la cual se mide.
   Por ejemplo:
   a. Cuando una galleta debe pulverizarse para determinar el contenido de grasa.
   b. Al probar los cinturones de seguridad para conocer su punto de ruptura, evidentemente se destruye el producto. Si todos se probaran de esa manera, no quedaría ninguno para vender.

Razones del uso del censo

1. La población es muy pequeña
   Por ejemplo, si se quiere conocer el historial de empleo de los graduados en Estadística de la Universidad de los Andes en el año 2010, se podría establecer contacto con ellos.
2. Si el tamaño de la muestra es relativamente grande con respecto al tamaño de la población, el esfuerzo adicional requerido para hacer un censo puede ser pequeño
3. Si se requiere una exactitud completa, un censo es la única forma de alcanzarla.
   Por ejemplo, un gerente bancario no tomaría una muestra al azar del dinero en las cajas para saber de cuánto efectivo dispone el banco, sino que contaría todo el dinero depositado en ellas.

Clasificación de la Estadística

Estadística Descriptiva

Cuando algunas personas escuchan la palabra "estadística", inmediatamente se imaginan cosas como: promedios de bateo, índices de accidentes, tasas de mortalidad, promedio de goles como visitante (en fútbol) etc. Esta rama de la estadística que utiliza números para describir hechos, recibe el nombre de *estadística descriptiva*, la cual consiste en organizar, resumir, simplificar,

presentar los datos en cuadros y gráficos y del cálculo de medidas numéricas que permitan destacar los aspectos más importantes de los datos.
Los métodos estadísticos descriptivos permiten obtener una visión completa de un fenómeno en el sentido de describir lo que está ocurriendo en determinado momento. En lenguaje figurado, *la estadística descriptiva* proporciona una fotografía o inventario de una situación y pone de relieve los aspectos de mayor interés.
El promedio industrial Dow-Jones, el índice de desempleo, el costo de la vida, la precipitación pluvial, el rendimiento medio de un auto en kilómetros por litro y los promedios de calificación, quedan todos en esta categoría.

Nota:
Un análisis descriptivo puede realizarse en una muestra o en toda una población.

Inferencia Estadística
Consiste en el análisis e interpretación de una muestra de datos. Más formalmente la *inferencia estadística* se encarga de estudiar las características y las leyes propias de la población mediante una muestra seleccionada de ella.
El *muestreo* es un ejemplo vivo en la siguiente situación familiar para nosotros: no hay que comerse todo el queso para saber si esta salado. Por tanto, la idea básica en el muestreo es medir una porción pequeña pero *típica*, de alguna población, y posteriormente utilizar dicha información para inferir (conjeturar inteligentemente) qué características tiene la población total. Otros ejemplos comunes son:

- Meter la punta del pie en el agua para calcular su temperatura en la piscina
- Sacar un auto nuevo para probarlo
- Realizarse un examen de sangre, etc.

Además hay muchas formas de aplicar este concepto a la industria y los negocios. Considérese los siguientes ejemplos:

- ♦ Un estudio cinematográfico somete a diversas pruebas a algunos actores y actrices antes de decidir quién interpretará cada papel.
- ♦ Las fábricas suelen producir un pequeño número de piezas (producción piloto) antes de pasar a la producción en gran escala.
- ♦ Muchas compañías almacenan cientos de artículos en inventario y, mediante técnicas de muestreo, pueden estimar su valor en unidades monetarias sin tener que contar por completo todos los artículos.
- ♦ Algunas veces se llevan a cabo estudios de mercado en ciudades claves, para establecer el grado de aceptación por el consumidor.

Veamos de una manera ilustrativa la definición de inferencia Estadística:

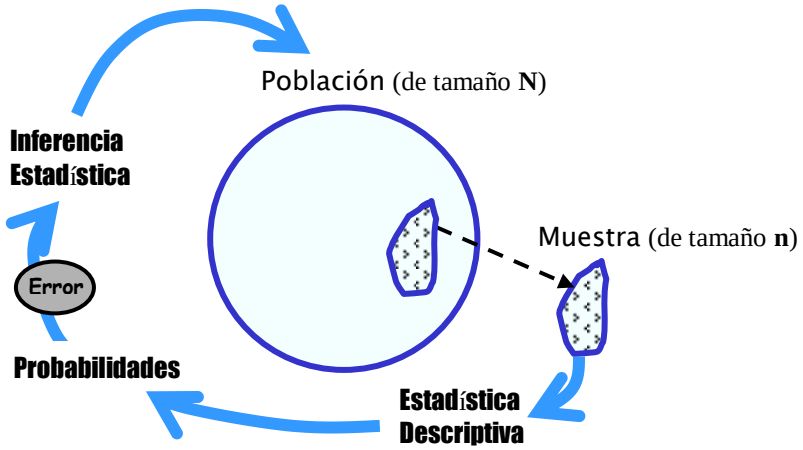


Figura 3. Ilustración didáctica de estadística descriptiva e inferencia estadística

Nota:

1. Los métodos estadísticos inferenciales tienen su base de apoyo en la *Teoría de Probabilidades* y en la *Teoría del Muestreo*.
2. Como los datos provienen de un conjunto menor que la población, se cometen errores al hacer una inferencia. Estos errores pueden ser cuantificados, así como la ***probabilidad*** de cometerlos, la cual, además de tratar con situaciones influenciadas por factores no controlados por el analista, proporciona un modelo racional para trabajar con la variabilidad inherente a la naturaleza del fenómeno bajo estudio y también con las situaciones relacionadas con el azar. El conocimiento de las probabilidades relacionadas con una situación, suministra la base para el desarrollo de las técnicas para la toma de decisiones, explica el funcionamiento de esas técnicas, e indica la manera en que las conclusiones pueden ser presentadas e interpretadas correctamente.
3. La recolección de información constituye un aspecto importante en una investigación y en ese sentido *la estadística* proporciona al investigador el apoyo necesario para su ejecución en todo lo referente a los instrumentos de recolección a utilizar y al tipo y cantidad de información a recoger de tal manera que se obtengan resultados confiables, especialmente en aquellos casos en que se piensa inferir los resultados a una población. La parte de la *estadística* encargada de estos aspectos se conoce como *diseño de experimentos* y *teoría del muestreo*.
4. En este curso se debe asumir que los datos ya se han recogidos mediante técnicas estadísticas y se trabajará en estas notas a partir de este supuesto.

En resumen se tiene la siguiente clasificación de la estadística:

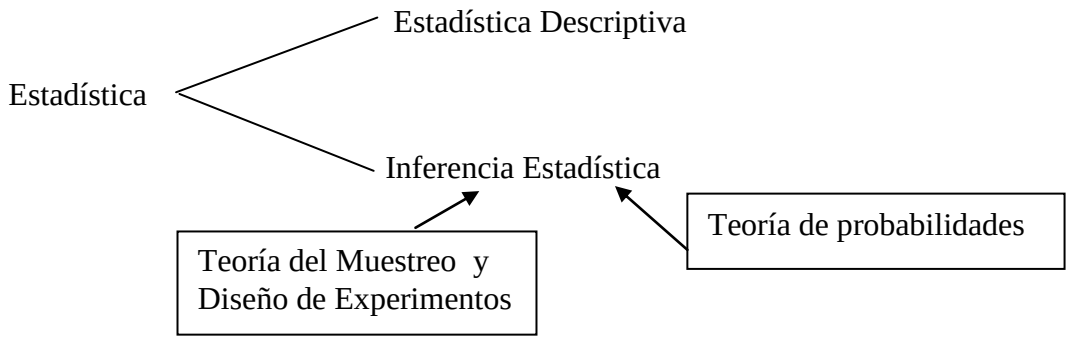


Figura 4. Clasificación de la Estadística

La Estadística y el método científico

Los métodos estadísticos utilizan *el método científico*, que en general, consiste en cinco pasos básicos:

1. Definir cuidadosamente el problema.
   Asegurarse de que esté claro el objeto de un estudio o un análisis.
2. Formular un plan para recopilar los datos necesarios.
3. Reunir los datos.
4. Analizar e interpretar los mismos.
5. Anotar las conclusiones y otros descubrimientos, de manera que sean fácilmente comprendidos por los que utilizarán los resultados al tomar decisiones.

Datos Estadísticos

Los datos estadísticos se obtienen mediante un proceso que comprende la observación o medición de conceptos como:

- Ingresos anuales en una comunidad.
- Calificaciones de exámenes.
- Cantidad de café por taza despachada por una máquina vendedora.
- Resistencia a la rotura de fibras de plástico.
- Porcentaje de azúcar en cereales, etc.

Tales conceptos también reciben el nombre de **variables**, ya que producen valores que tienden a mostrar cierto grado de variabilidad, al efectuarse mediciones sucesivas.

Tipos de Datos

Variable continua
Son aquellas variables que pueden asumir cualquier valor en determinado intervalo de valores. Características tales como altura, peso, longitud, espesor, velocidad, temperatura, etc., quedan dentro de esta categoría

Datos continuos
Son aquellos datos que se obtienen de variables continuas.
Ejemplo:
La cantidad de café que se vende por día, la gasolina que se expende por hora, la velocidad del aire, etc.

Nota:
En términos prácticos, los instrumentos de medición presentan ciertas limitaciones de tipo físico que restringen el grado de precisión, a pesar de esto los datos siguen siendo continuos. Este es el caso de datos que representan la estatura de una persona. Usando una cinta métrica tradicional, se habla por ejemplo, de que una persona mide 71,3 metros. Pero si tuviésemos a disposición

algún instrumento electrónico sofisticado podría obtenerse que esta persona mide 71, 287253046301 metros.

Variable Discreta
Es la que puede asumir sólo ciertos valores, por lo regular, números enteros.

Datos Discretos
Los datos discretos surgen al contar el número de conceptos que posee cierta característica. Ejemplos de datos discretos son: el número de clientes por día, la cantidad de alumnos en un salón de clase, los defectos de un auto, número de goles en un partido de fútbol.

Datos Cuantitativos
Tanto los datos discretos como los continuos se conocen como datos cuantitativos ya que son inherentemente numéricos. Es decir, ciertos valores numéricos se relacionan de manera natural con las variables que se miden. Las variables de donde se obtienen este tipo de datos se denominan **variables cuantitativas**.

Variables Nominales
Se caracterizan porque la única relación que está definida entre los valores que puede tomar la variable es la igualdad o diferencia.
Ejemplo:

- Sexo (masculino, femenino)
- Color de los ojos (azul, marrón, negro, verde)
- Campo de estudio (medicina, administración, ingeniería, economía), etc.

Nota:
Ninguna de las características anteriores es numérica por naturaleza. En caso de utilizar números, estos simplemente constituyen un indicador de distinción cualitativa y en ningún caso el orden y la distancia entre ellos tiene otra interpretación. Es decir, si se usan números más bien se deben considerar como un código y no como el valor numérico que representa. Por ejemplo, la variable sexo puede tomar los valores 0 y 1, donde un 0 representa a *masculino* y un 1 representa a *femenino*. En este caso sumar 0 + 1 no tiene sentido porque es como querer sumar masculino + femenino. Nótese que en este ejemplo el 0 y el 1no son inherentemente numéricos, son códigos.

Los datos asociados a este tipo de variable se conocen como **datos nominales**.

Variables Ordinales
Se caracterizan porque entre dos valores de la variable, además de la relación de igualdad o diferencia se pueden dar las relaciones "mayor que" o "menor que". Es decir, dados dos valores de la variable se puede decir si son iguales o diferentes y además saber cuál valor está antes que el otro de acuerdo a un orden, es decir se jerarquizan los valores.

Ejemplo:
- Concurso de belleza o cocina
- Jerarquías del ejército, etc.

Los **datos ordinales** son los valores que toman las variables ordinales.

Variables Cualitativas
Son aquellas comprendidas por variables nominales u ordinales. Éstas no son inherentemente numéricas, es decir, presenta modalidades no cuantitativas y en caso de utilizar números para representar esas modalidades, estos no tienen significado en sí mismos. Las variables cualitativas también son conocidas como *atributos.* Los datos que se obtienen de este tipo de variables se llaman **datos cualitativos**.

En la siguiente ilustración se puede observar la clasificación de las variables:

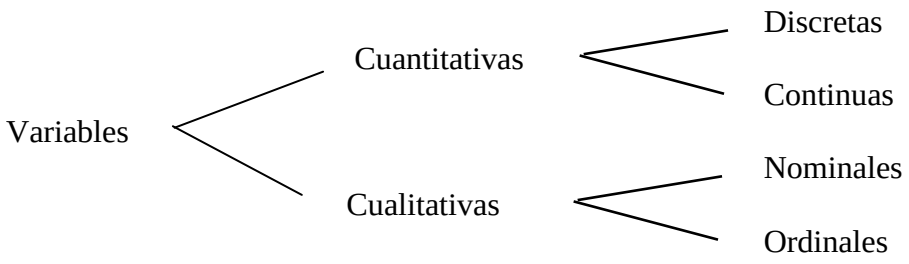


Figura 5. Clasificación de las variables

Notación:
Se acostumbra denotar a las variables por letras latinas mayúsculas, en general las últimas del alfabeto: X, W, Y, Z, etc. A los valores que toma la variable se habitúa denotar con la misma letra en minúscula enumerada con un subíndice. Por ejemplo, si $\boldsymbol{Y}$ representa a la variable Edad, entonces $\boldsymbol{y_3}$ indica la edad que toma el tercer individuo.

Series Cronológicas o Series de Tiempo

Una *serie cronológica* o *serie de tiempo* es una sucesión de observaciones tomadas secuencialmente en el tiempo. Así, una *serie de tiempo* refleja las variaciones de una variable en el tiempo.

Ejemplos
a. Producción anual de petróleo (en número de barriles) en Venezuela
b. La cotización diaria del dólar
c. El índice mensual de precios al consumidor
d. Las pruebas de electrocardiograma en un hospital

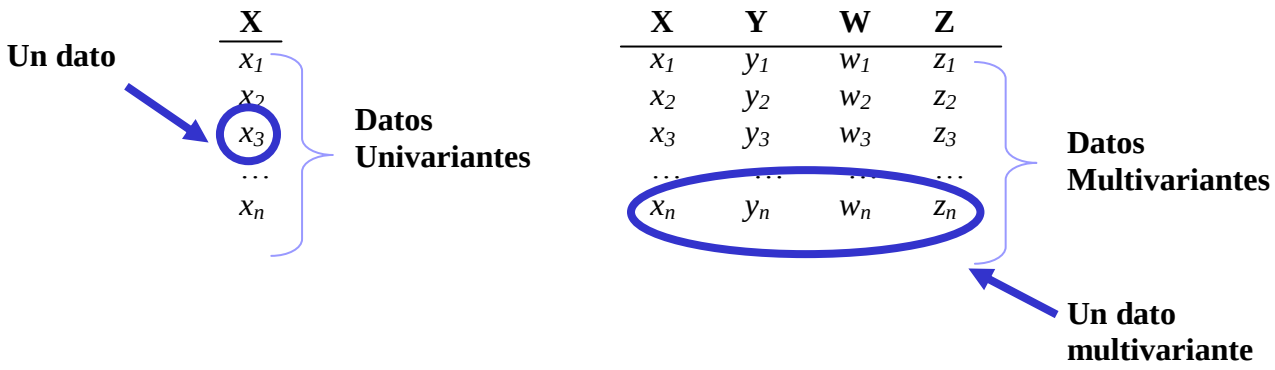
Variables Univariantes y Multivariantes (también unidimensionales y multidimensionales)

Existe otra clasificación de acuerdo al número de variables que se analizan conjuntamente. Cuando las variables se presentan y analizan individualmente, se habla de *variable univariante*. Alternativamente, cuando se analizan simultáneamente dos, tres o más variables se habla de *variable bivariante, trivariante o multivariante*.

Por ejemplo, de una encuesta se obtienen los datos sobre tipo de sangre, peso, ingreso y sexo de los estudiantes de Métodos Estadísticos I; y se analiza cada una de estas variables separadamente. En este caso se tienen cuatro variables univariantes.

Por otro lado, si es de interés analizar conjuntamente las variables tipo de sangre y peso se está ante la presencia de una variable bivariante. Pero, si se analizan simultáneamente las cuatro variables entonces se habla de una variable multivariante.

Sea,
X: Tipo de sangre, Y: Peso, W: Ingreso, Z: Sexo.



Estadístico y Parámetro

Es conveniente describir una población en términos de unas pocas medidas que resumen características de interés. Una medida calculada de los valores poblacionales es llamada **Parámetro**. Muchos parámetros distintos pueden ser definidos para medir diferentes aspectos de una población.

Un **Estadístico** es una medida que es calculada sobre la base de los datos de la muestra. Más formalmente, un *estadístico* es una función matemática de una muestra.

Por tanto, un *parámetro* es un valor único mientras que un *estadístico* puede tomar distintos valores dependiendo de la muestra seleccionada.

Por ejemplo, de una muestra de 100 estudiantes de la ULA, se puede obtener que el porcentaje de estudiantes que trabajan es 9%; este 9% ¿es el valor de un estadístico o de un parámetro?

Se acostumbra a denotar los parámetros con letras griegas y los estadísticos con letras latinas. Por ejemplo, generalmente se denota la media de una población como μ y la media de una muestra como $\bar{X}$.

Ejercicios:
Para cada una de las siguientes investigaciones describa:

a) Característica en estudio
b) Tipo de variable
c) El universo estadístico
d) La unidad elemental
e) La población estadística
f) La muestra
g) El parámetro de interés del estudio
h) El estadístico

1. Se realiza un estudio para determinar el peso promedio de las vacas en la zona del Sur del Lago de Maracaibo. Se seleccionan 500 vacas al azar de varias granjas de esa región. Luego se registra el peso de cada vaca. Así se obtiene que el peso promedio de las vacas seleccionadas es de 425 Kg.
2. Para estimar cuántos libros de la biblioteca deben ser encuadernados de nuevo, un bibliotecario elige aleatoriamente 100 libros de los estantes de la biblioteca y registra si el libro debe encuadernarse o no.
3. Para una encuesta sobre los ingresos familiares en un pequeño pueblo con 1000 familias al sur del estado Mérida, se seleccionó aleatoriamente 50 familias de ese pueblo y se les preguntó sobre sus ingresos mensuales, luego se obtuvo de estos datos el ingreso promedio por familia.
4. El Ministerio del Trabajo realiza una investigación sobre industrias manufactureras venezolanas, para estimar el porcentaje de obreros que sufren accidentes laborales mensualmente. De esta manera selecciona 100 industrias al azar y encuentra que aproximadamente el 3% de sus obreros sufren accidentes de trabajo en un mes.
5. Una empresa de televisión por cable desea instalarse en la ciudad de Mérida y dentro del estudio de factibilidad requiere conocer el nivel de ingreso promedio mensual de las familias de la ciudad. Se lleva a cabo una investigación en donde se seleccionan 1500 familias de forma aleatoria.
6. Una fábrica de vigas (de cierta aleación de hierro y otros minerales) para la construcción, desea conocer la resistencia de sus productos. Para llevar a cabo un estudio al respecto se seleccionan 850 vigas producidas por esta fábrica y se mide su resistencia.

Notación de suma con sigma (sumatoria)

Antes de comenzar la siguiente sección, se introducirá un tipo de notación matemática que sirve para expresar muchas de las fórmulas que se utilizan en los procedimientos estadísticos que se estudiaran más adelante. En muchas ocasiones será necesario obtener la *suma* de un conjunto de números.

Supóngase que alguna variable **X** toma los siguientes valores:

9 4 3 1 6

Nótese que puede considerarse 9 como el primer valor de **X**, 4 como el segundo valor de **X**, 3 como el tercer valor de **X**, 1 como el cuarto valor de **X,** y 6 como el quinto valor de **X**. Una manera sencilla de expresar esto consiste en utilizar subíndices que representen la posición del valor en la lista. De este modo, el 9 que es el primer valor de **X** será representado por x_1; de manera similar, debido a que 4 es el segundo valor de **X**, estará representado por x_2**.** Es decir:

$x_1 = 9$ $x_2 = 4$ $x_3 = 3$ $x_4 = 1$ $x_5 = 6$

Cuando se desee referir a un valor de X de forma general sin hacer especificaciones, se utilizará el subíndice *i* y al valor se le llamará x_i (léase "equis sub i")

La letra griega **Σ** (sigma mayúscula) se utiliza para denotar una suma. Entonces $\sum_{i=1}^{5} x_i = 23$. El símbolo Σ en la expresión anterior indica que se deben sumar los valores de **X**. Además, la expresión "i = 1" que se encuentra debajo de sigma comienza con el valor de **X** que tiene el subíndice $i = 1$ (x_1). De esta manera, se suman sucesivamente los valores de **X**, uno cada vez, y la operación es finalizada cuando se alcanza el valor de **X** cuyo subíndice es igual al número entero que se encuentra encima de sigma, 5 (x_5). Por consiguiente en la suma anterior se tiene paso por paso:

$$\begin{aligned}\sum_{i=1}^{5} x_i &= x_1 + x_2 + x_3 + x_4 + x_5 \\ &= 9+4+3+1+6 \\ &= 23\end{aligned}$$

Si sólo se desea sumar algunos valores, se utilizan los subíndices anotados por debajo y por encima de Σ. Por ejemplo:

$$\begin{aligned}\sum_{i=2}^{4} x_i &= x_2 + x_3 + x_4 \\ &= 4+3+1 \\ &= 8\end{aligned}$$

Al invertir este proceso, se puede utilizar este método para abreviar la expresión de los datos que se quiere sumar, por ejemplo:

$x_3 + x_4 + x_5$ se convierte en $\sum_{i=3}^{5} x_i$

$\sum_{i=1}^{n} x_i$ significa que *n* observaciones (todas) han de ser sumadas, y a menudo esto se abrevia con los símbolos $\sum x_i$ o $\sum x$.

La notación sigma puede también utilizarse con expresiones más complicadas, como se demuestra en los siguientes ejemplos:

a. $\sum_{i=1}^{3} i = 1+2+3 = 6$

b. $\sum_{i=1}^{5} x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 = 9^2 + 4^2 + 3^2 + 1^2 + 6^2 = 143$

c. $\sum_{i=0}^{n} \frac{1}{2^i} = \frac{1}{2^0} + \frac{1}{2^1} + \frac{1}{2^2} + ... + \frac{1}{2^n}$

Propiedades de Σ

Teorema 1
Si *a* es una constante y cada uno de los *n* valores diferentes de *i* es igual a *a*, entonces

$$\sum_{i=1}^{n} a = na$$

Prueba
Como cada una de las *x* es igual a una cantidad constante *a*:

$$\sum_{i=1}^{n} a = a + a + \cdots + a = na$$

Teorema 2
Sea *a* una constante cualesquiera de todos los valores individuales que intervienen en la suma,

$$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i$$

Prueba

$$\begin{aligned}\sum_{i=1}^{n} ax_i &= ax_1 + ax_2 + \cdots + ax_n \\ &= a(x_1 + x_2 + \cdots + x_n) \\ &= a\sum_{i=1}^{n} x_i\end{aligned}$$

Teorema 3
La notación sigma se puede distribuir respecto de la suma (o de la diferencia):

$$\sum_{i=1}^{n}(x_i + y_i - z_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} z_i$$

Prueba
Lo anterior se cumple porque:

$$\begin{aligned}\sum_{i=1}^{n}(x_i + y_i - z_i) &= (x_1 + y_1 - z_1) + (x_2 + y_2 - z_2) + \cdots + (x_n + y_n - z_n) \\ &= (x_1 + \cdots + x_n) + (y_1 + \cdots + y_n) - (z_1 + \cdots + z_n) \\ &= \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} z_i\end{aligned}$$

Ejercicio:
Calcule cada una de las siguientes cantidades sirviéndose de los datos proporcionados (Nota: *n* es el número de observaciones)

Y
21
6
54
60
34
9
68
73
8

a. $\sum_{i=1}^{6} 8$

b. $\sum_{i=1}^{n} y_i$

c. $\sum_{i=1}^{n} \frac{y_i}{n}$

d. $\sum y^2$

e. $\left(\sum y\right)^2$

f. $\sum_{i=1}^{n}(y - 37)$

g. $\sum_{i=1}^{n}(y - 37)^2$

h. $\frac{\sum_{i=1}^{n}(y - 37)^2}{n}$

i. $\frac{\left\{\sum_{i=1}^{n} y_i^2 - \left[\frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]\right\}}{n}$

j. $\sum_{i=1}^{n}(y - 10)^2$

Ejercicio:
Calcule las siguientes cantidades según los datos que se indican:

X	**Y**
3	10
5	11
9	15
10	19
2	21
1	26

a. $\sum x$

b. $\sum_{i=1}^{n} y_i$

c. $\sum_{i=1}^{n} x_i y_i$

d. $\sum_{i=1}^{n} x_i^2 y_i$

e. $\sum_{i=1}^{n}(x_i - y_i)$

f. $\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$

Estudio descriptivo de una colección de datos

Cuando se ha recolectado la información correspondiente al fenómeno que se está investigando, se cuenta con una colección de datos individuales, la cual constituye la materia prima para el investigador. Comúnmente, este conjunto de datos es bastante grande y por ende es muy difícil obtener algunas conclusiones que sean de utilidad para el estudio. Por tal razón se hace necesario utilizar los métodos estadísticos descriptivos tanto para resumir y presentar convenientemente los datos, como también para conseguir algunos indicadores numéricos que sean de utilidad para la interpretación de los aspectos más importantes y de interés de los datos.

Organización de datos cualitativos

La manera de condensar o agrupar los datos cualitativos es muy intuitiva. Sólo es necesario un conteo de las distintas modalidades que presenta la variable en cuestión, lo que se conoce como frecuencia:

VARIABLE			
Modalidad 1	**. . .**	**Modalidad k**	**Total**
f_1	. . .	f_k	n

Tabla de doble entrada o tabla de contingencia

También se pueden organizar dos variables en una tabla. Este tipo de organización de datos se conoce como *tabla de doble entrada o tabla de contingencia*:

		VARIABLE A a_1	a_2 …	a_i	**TOTALES**
VARIABLE B	b_1				
	b_2				
	⋮				
	b_j				
TOTALES					

Nota:
También pueden organizarse en una misma tabla tres o más variables.

Organización de datos cuantitativos

Cuando se agrupan datos cuantitativos generalmente el tipo de organización visto antes no es adecuado. Esto se debe a que las variables cuantitativas por lo regular presentan muchos valores distintos, con lo cual la finalidad de condensar la información no se cumple.
La idea ahora consiste en establecer intervalos que cubran todos los datos que se tienen a disposición sobre la variable en estudio. De esta manera, se construye una tabla en la que se cuenta el número de observaciones contenidas en cada intervalo previamente especificado. Estos intervalos se llaman clases o intervalos de clases y el número de datos en cada intervalo se denomina frecuencia. Esta forma de agrupar los datos tendrá esta apariencia:

Intervalos de clase	**frecuencia**
LI_1 - LS_1	f_1
LI_2 - LS_2	f_2
...	...
LI_i - LS_i	f_i
...	...
LI_k - LS_k	f_k
	Total de observaciones

De este modo, se puede definir una **distribución de frecuencias** como una ordenación tabular de los datos en intervalos de clase con sus respectivas frecuencias.

Nota:
Cuando los datos se presentan en distribuciones de frecuencias, se habla de *datos agrupados*, mientras que cuando se presentan individualmente, se habla de *datos no agrupados*.

Pasos para la construcción de una distribución de frecuencias

1. Determinar el valor máximo y el valor mínimo de los datos.
2. Calcular el rango (o recorrido) de la variable el cual viene dado por la diferencia entre el valor máximo y el valor mínimo. El rango se denota por ***R***.
3. Determinar el *número de clases* (***K***) y las *amplitudes de clase* (***C_i***):
   A la anchura de un intervalo de clase se le conoce como *amplitud de clase*, es decir, la amplitud de clase de un intervalo viene dada por la diferencia entre el límite superior y el límite inferior de dicho intervalo. Podemos determinar la amplitud o el número de clases tomando en cuenta lo siguiente:

   a. Si se conoce el número de clases: $C_i = \frac{R}{K}$

   b. Si se conoce la amplitud de las clases: $K = \frac{R}{C_i}$

c. Regla de Sturges:
*K = 1 + 3,3 * Log n*

Nota:

i. La fórmula de Sturges sólo proporciona una orientación sobre cuál debe ser el número de clases. También se puede usar la regla de la raíz cuadrada: $K = \sqrt{n}$.

ii. Pueden existir clases abiertas, es decir, clases que sólo tienen un límite superior o solamente un límite inferior. Si ese es el caso, a esta clase abierta no se le podrá determinar la amplitud.

iii. En la práctica no se conoce de antemano el número de clases y la amplitud de estas. Sin embargo existen dos recomendaciones importantes al construir una distribución de frecuencias:

- Que el número de clases no sea inferior a 5 ni mayor que 15.
- De ser posible es deseable que todas las clases tengan la misma amplitud.

4. Proceder a construir los intervalos de clase.
En este punto ya se debe conocer el número de intervalos de clase a construir y las amplitudes de clase de cada uno de ellos, las cuales pueden ser iguales o no. Para la construcción de las clases se deben seguir los siguientes pasos:
   a. Establecer el límite inferior del primer intervalo de clase. Esto se puede realizar arbitrariamente de acuerdo a las siguientes alternativas:
      - Utilizando el valor mínimo de los datos
      - Utilizando otro valor menor al mínimo, pero no muy alejado.
   b. Fijado el primer límite inferior se le suma a este la amplitud de la primera clase, C_1, y se obtiene el límite superior de esta primera clase, el cual se constituye a la vez como el límite inferior de la segunda clase, a este se le suma la amplitud C_2 y se obtiene el límite superior de la segunda clase. Y de la misma manera se construyen los *K* intervalos. Naturalmente el último intervalo de clase debe incluir el valor máximo de los datos.
   c. Para calcular la frecuencia de cada intervalo, se debe asumir lo siguiente: En términos matemáticos los intervalos de clase van a ser intervalos cerrados por su límite inferior y abiertos por su límite superior. Es decir, el intervalo de la i-ésima clase será $[LI_i - LS_i)$, con $i = 1, \dots, K$.

5. Determinar el número de datos contenidos en cada clase. Es decir, determinar las *frecuencias absolutas* de clase (f_i). Evidentemente se debe cumplir que $\sum_{i=1}^{K} f_i = n$, siendo *n* el número total de datos

6. Determinar el resto de las frecuencias.
   a. Frecuencia relativa de una clase:
   Se va a denotar por fr_i y se obtiene de la siguiente manera:

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

Siempre se cumple que $\sum_{i=1}^{K} fr_i = 1$. La frecuencia relativa de una clase representa la *proporción de datos* contenidos en ese intervalo de clase.

b. Frecuencia acumulada de una clase:
Se denota por F_i. Se obtiene sumando las frecuencias absolutas de todas las clases anteriores a ella más la frecuencia absoluta de la *i*-ésima clase considerada. Por tanto la frecuencia acumulada de la última clase es $F_k = n$.
La frecuencia acumulada, F_i, representa el *número de observaciones que son menores* que el límite superior de la *i*-ésima clase.

c. Frecuencia relativa acumulada de una clase:
Se denota por Fr_i y se obtiene de la siguiente manera:

$$Fr_i = \frac{F_i}{n}$$

También, $Fr_i = fr_1 + fr_2 + fr_3 + \ldots + fr_i$
La frecuencia relativa acumulada, Fr_i, representa la *proporción de todas las observaciones que son menores* que el límite superior de la i-ésima clase.

d. Marca de clase o punto medio de clase:
La *marca de clase* o *punto medio de clase*, denotado por m_i se define como el punto central de la clase particular:
$m_i = \frac{LI_i + LS_i}{2}$, en donde LI_i es el límite inferior de la *i*-ésima clase y LS_i es el límite superior de esa clase.

Ejercicio:
Con base en los datos recogidos en clase, construir una distribución de frecuencias para la variable peso.

Nota 3:
Existen algunas situaciones en que uno o más intervalos de clase en una distribución de frecuencias no tienen límite inferior o superior. Estos se conocen como ***Clases Abiertas***.
Por ejemplo, la siguiente distribución de frecuencias tiene dos clases abiertas:

Clases	fi
Menos de 5	73
5 - 10	58
10 - 15	35
…	…
50 y más	39

Observe que a las clases abiertas no se les puede determinar la amplitud y tampoco la marca de clase.

Ejercicio:
Completar la siguiente distribución de frecuencias:

Clases	m_i	f_i	fr_i	F_i	Fr_i
- - -	15	--	---	--	0,16
[20 -30)	--	--	0,08	--	---
[30 - 40)	--	6	---	12	---
[40 - 50)	45	8	---	20	---
- - -	65	--	---	--	---
Totales		--	---		

Distribución de frecuencias cuyas clases son valores individuales de la variable en estudio

En muchas ocasiones se presentan colecciones de datos en las cuales el número de valores diferentes que toma la variable de interés es pequeño y por consiguiente, no es apropiado agrupar estos datos en una distribución de frecuencias cuyas clases sean intervalos. Generalmente, en estos casos, los datos son de tipo discreto.
En tal situación, se toman como clases los diferentes valores de la variable y las frecuencias se calculan de la forma habitual.

Ejemplo:
Con base en los datos recogidos en clase, construir una distribución de frecuencias para la variable *Número de veces al mes que va al cine*.

Nota:
a. Nótese que en este tipo de distribución de frecuencias no existen límites de clase, amplitudes y las marcas de clase m_i coinciden con las clases.
b. Obsérvese también que en las distribuciones de frecuencias cuyas clases son valores individuales, se puede reconstruir fácilmente la colección de datos originales. Recuerde que esto no es posible cuando las clases son intervalos.

Ventajas y desventajas de agrupar los datos en distribuciones de frecuencias

- Facilita la presentación y resumen de los datos, lo que permite analizar sus aspectos más resaltantes.
- La desventaja principal es que se pierde la individualidad de los datos. Se sabe que en determinado intervalo está contenido cierta cantidad de datos pero no se conoce exactamente qué valores toman.

En conclusión, al agrupar datos se gana en simplicidad y accesibilidad, pero se pierde el nivel de detalle en el caso de distribuciones de frecuencias con intervalos.

Presentación de los datos

Presentación Escrita
Este método consiste en presentar un informe que reseña los rasgos de mayor importancia de los datos. Debido a que es necesario leer el informe íntegramente para conocer los aspectos de interés de los datos, este método no es muy efectivo y por consiguiente es poco empleado. Sin embargo, posee la virtud de poder resaltar las cifras y las comparaciones que se consideren esenciales.

Ejemplo:
En el boletín mensual sobre el *índice nacional de precios al consumidor* (INPC) presentado por el Banco Central de Venezuela, en fecha 06/06/2013, se puede leer:

INDICE NACIONAL DE PRECIOS AL CONSUMIDOR
PARA EL MES DE JUNIO DE 2013

"El índice nacional de precios al consumidor (INPC), elaborado por el Banco Central de Venezuela (BCV) y el Instituto Nacional de Estadística (INE), registró en el mes de mayo de 2013 una variación intermensual de 6,1%, mayor a la del mes previo (4,3%) y a la del mismo mes del año 2012 (1,6%).
Con este resultado el indicador de precios al consumidor acumuló un incremento relativo de 19,4% en los primeros cinco meses del año 2013, por encima del 6,0% obtenido en igual período del año anterior. La variación anualizada correspondiente a mayo de 2013 se situó en 35,2%, superior a la observada en mayo de 2012 (22,6%).

Al desagregar la variación intermensual por agrupaciones se aprecia una amplia dispersión que abarca un rango entre 0,1% y 10,0%, con sólo un grupo por encima del 6,1% global. En efecto, las tasas intermensuales de variación fueron, en orden de magnitud: Servicios de la vivienda (0,1%), Alquiler de vivienda (0,8%), Comunicaciones (1,5%), Servicios de educación (1,7%), Esparcimiento y cultura (2,1%), Bienes y servicios diversos (2,4%), Salud (2,5%), Vestido y calzado (3,4%), Transporte (3,9%), Equipamiento del hogar (5,0%), Bebidas alcohólicas y tabaco (5,1%), Restaurantes y hoteles (6,0%) y Alimentos y bebidas no alcohólicas (10,0%). Estas cifras se dieron en un escenario afectado por el efecto residual del ajuste cambiario del mes de febrero y por el aumento del salario mínimo a partir del 1° de mayo.

En el caso particular del registro de la agrupación Alimentos y bebidas no alcohólicas resultaron determinantes, por una parte, el aumento de 20,4% en los productos agrícolas y, por otra, el ajuste del precio oficial de algunos rubros sujetos a control.

En la variación del grupo Transporte se recogen los efectos del aumento de las tarifas del traslado terrestre de pasajeros, vigentes a partir del mes de abril.

En el ámbito geográfico se observó bastante uniformidad en las tasas intermensuales del índice de precios al consumidor, con un mínimo de 4,6% y un máximo de 6,6%: Caracas, 6,2%; Maracay, 6,1%; Ciudad Guayana, 6,1%; Barcelona-Puerto La Cruz, 5,6%; Valencia, 6,1%; Barquisimeto, 4,6%; Maracaibo, 6,6%; Mérida, 5,1%; Maturín, 5,0%; San Cristóbal, 5,7% y Resto nacional, 6,3%.

La variación consolidada correspondiente a los Servicios que pertenecen a la canasta del INPC fue de 3,8% en el mes de mayo, cerca de la mitad del 7,5% que se obtuvo para los Bienes , con ambas categorías acelerando respecto al mes anterior, desde 2,9% y 5,1%, respectivamente.

El índice del núcleo inflacionario mostró una variación de 5,0% en mayo, significativamente por debajo de la tasa global (6,1%) pero mayor al 4,0% registrado en abril, aceleración que responde a los mayores crecimientos ocurridos en las cuatro categorías que conforman el indicador: Alimentos elaborados, de 5,5% a 5,7%; Textiles y prendas de vestir, de 2,9% a 3,2%; Bienes industriales, distintos de alimentos y textiles, de 4,0% a 4,5%, y Servicios no administrados, de 3,5% a 5,2%.

El indicador de escasez descendió a 20,5%, mientras que el índice de diversidad retrocedió a 113,3.

(Disponible en: http://www.bcv.org.ve/c4/notasprensa.asp?Codigo=10632&Operacion=2&Sec=False [consulta: 2013, Junio 6])

Cuadros Estadísticos

Los cuadros estadísticos son tablas en las cuales se exhibe de manera ordenada a los datos. Un cuadro estadístico debe ser capaz de explicarse por sí solo. Para cumplir esto debe poseer principalmente título, encabezados, cuerpo y fuente.
El *título* debe ser breve y suficientemente explicativo de la situación estudiada, la época y el sitio. *Los encabezados* son los nombres de las filas y columnas de la tabla. *El cuerpo* son los datos ya condensados y organizados. *La fuente* indica el origen de la información, por esta razón nunca debe faltar en todo cuadro estadístico.
Adicionalmente a los elementos anteriores, estos cuadros pueden llevar *notas preliminares* en el título del cuadro, *notas explicativas* debajo del cuadro y *numeración del cuadro* cuando existen varios de ellos.

Ejemplo:

Numeración del cuadro

CUADRO No. 1.3

Título

ÍNDICE NACIONAL DE PRECIOS AL CONSUMIDOR

Nota preliminar

Serie Mayo 2012 – mayo 2013 (Variaciones Porcentuales)
(Base Diciembre 2007=100)

Encabezados

Cuerpo

Meses	INPC
Mayo	1,6
Junio	1,4
Julio	1,0
Agosto	1,1
Septiembre	1,6
Octubre	1,7
Noviembre	2,3
Diciembre	3,5
Enero	3,3
Febrero	1,6
Marzo	2,8
Abril	4,3
Mayo	6,1

Fuente

Fuente: INE

Ejercicio:
Presentar en un cuadro estadístico la distribución de frecuencias de la variable peso.

Construcción de gráficos

Los gráficos facilitan la visualización de las cifras y son ampliamente utilizados en la representación de los datos estadísticos. Cuando se elabora cualquier clase de gráfico se pierde información, pues ya no existen las observaciones originales. Sin embargo, frecuentemente esa pérdida de información es pequeña comparada con la síntesis y facilidad de la interpretación.
Al igual que los cuadros estadísticos, los gráficos deben llevar un *título* que explique de lo que trata la información allí presentada y la *fuente*. También pueden llevar *notas explicativas, numeración correlativa*.

Algunos tipos de gráficos

- Diagrama de puntos
- Diagrama de dispersión
- Curvas
- Gráfico de barras
- Gráfico circular
- Pictogramas
- Gráficos especiales para distribuciones de frecuencias:
  - Histograma
  - Diagrama de líneas
  - Polígono de frecuencias
  - Ojiva
  - Diagrama de frecuencias acumuladas
- Diagrama de tallo y hojas
- Diagramas de caja

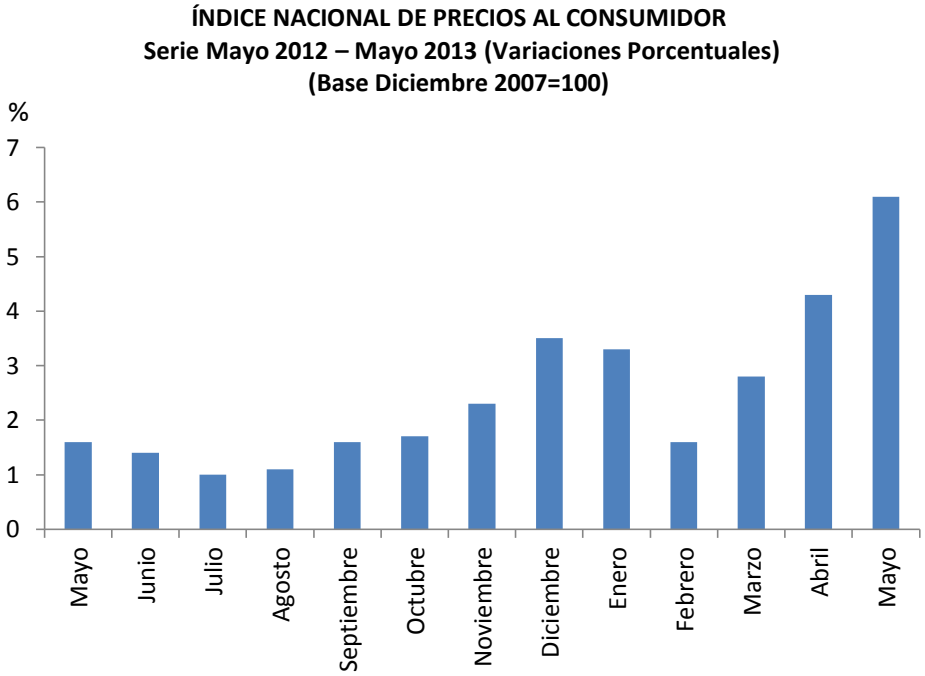
Gráfico de barras

Los gráficos de barras constituyen una herramienta muy adecuada para representar *series cronológicas*, para *datos cualitativos ordinales* y en general para datos donde exista algún orden. En algunas ocasiones también se utiliza en *datos nominales.*

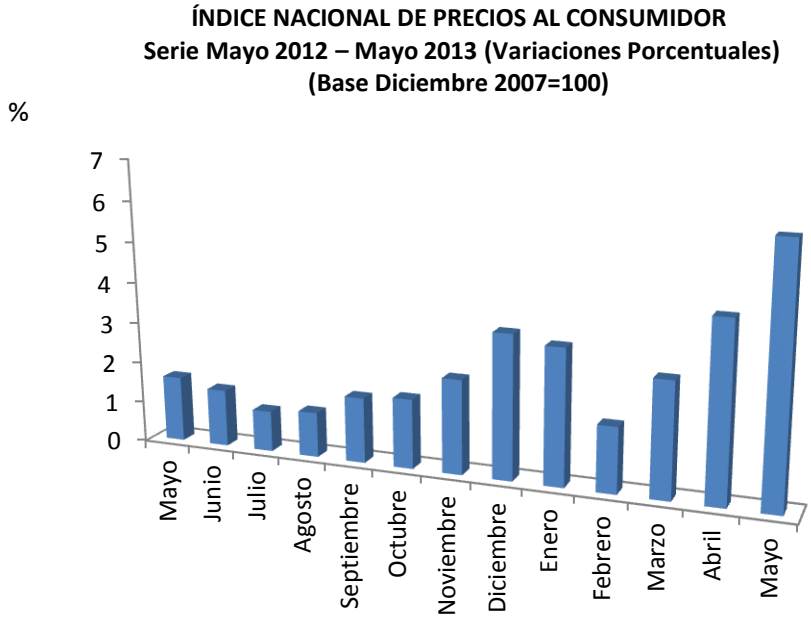
Construcción del gráfico

- Paso 1: Establezca un orden (arbitrario cuando la variable es cualitativa nominal) para la colocación, en el eje horizontal de:
  - Las distintas modalidades en el caso de variables cualitativas.
  - El tiempo de ser una serie de tiempo.
- Paso 2: Teniendo en cuenta el valor máximo de la frecuencia (o porcentaje) de los datos, escoja una escala vertical para representar los valores correspondientes.
- Paso 3: En el eje horizontal, para la primera modalidad (o tiempo), dibuje un rectángulo de base cualquiera y altura proporcional al valor de la modalidad.
- Paso 4: Repita el proceso del paso 3 para las demás modalidades.

Ejemplo:



Fuente: INE

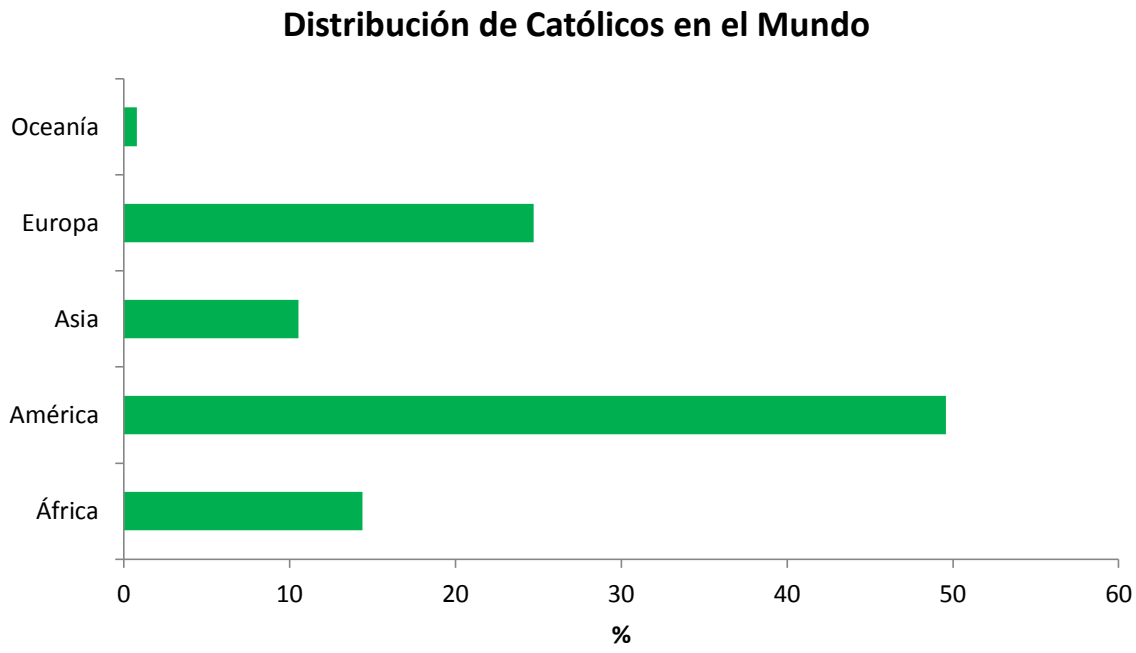


Fuente: INE

Como se puede apreciar en el gráfico anterior, la presentación puede realizarse en tres dimensiones lo cual puede mejorar la estética del gráfico. Nótese que la información representada en estos gráficos de barras es la misma recogida en la tabla de la página 25. Puede advertirse que resulta más fácil y rápido hacerse una idea del comportamiento del INPC observando el gráfico de barras que mirando los números de la tabla.

El gráfico de barras también puede construirse de tal manera que las barras aparezcan de forma horizontal. Se acostumbra utilizar esta variante cuando se comparan datos cualitativos o datos que se refieren a zonas geográficas.

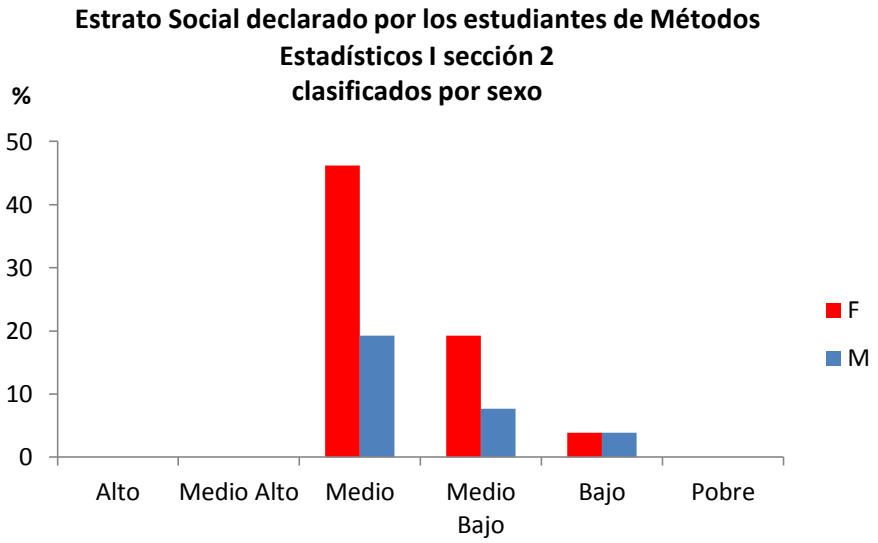
Ejemplo:



Fuente: Las estadísticas de la Iglesia Católica (Disponible en: http://www.vicariadepastoral.org.mx/domund_11/imagenes/Estadisticas.pdf *[consulta: 2012, Mayo 7])*

Con un gráfico de barras también existe la posibilidad de presentar dos o más variables en un mismo gráfico, de tal manera de que se pueda apreciar el comportamiento individual y además poder hacer comparaciones entre ellas. Veamos esto con un ejemplo:

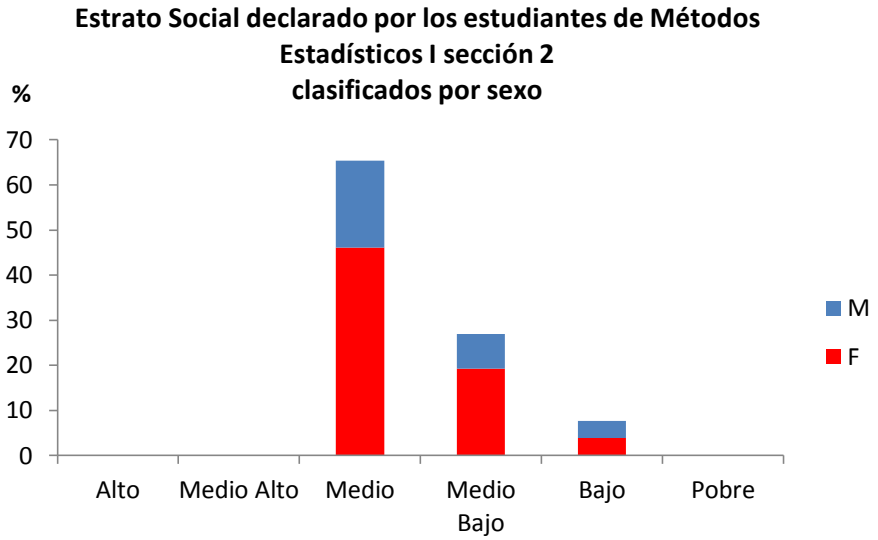
Ejemplo:



Fuente: Encuesta realizada por la cátedra de Estadísticas Básicas a los estudiantes de Métodos Estadísticos I sección 2 de FACES-ULA. Abril 2012

Otra variante de esta clase de gráficos es el de *columnas yuxtapuestas o adyacentes*, el cual representa la relación que hay entre los valores o categorías individuales y el total.

Ejemplo:



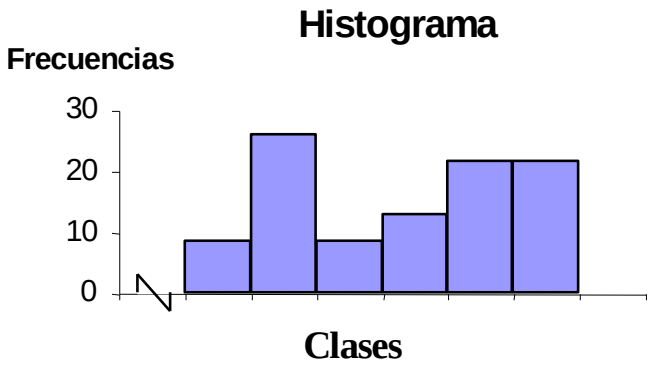
Fuente: Encuesta realizada por la cátedra de Estadísticas Básicas a los estudiantes de Métodos Estadísticos I sección 2 de FACES-ULA. Abril 2012

La desventaja de este tipo de gráfico es que puede ser dificultoso apreciar el comportamiento de la variable de arriba (Masculino en el ejemplo).

Histograma
El histograma es el gráfico adecuado para ilustrar el comportamiento de los valores agrupados en intervalos de clase, siendo un gráfico de barras compuesto por varios rectángulos adyacentes, que representan a la tabla de distribución de frecuencias de cierta variable cuantitativa. En el eje horizontal se marcan los intervalos, y cada intervalo es la base de cada rectángulo; en el eje vertical se marcan las alturas de los rectángulos la cual viene dada por las frecuencias respectivas (absolutas simple o relativas)

Construcción
- Paso 1: En el eje horizontal, marque sucesivamente los límites de cada clase.
- Paso 2: En el eje vertical, marque, en la escala, los valores correspondientes a las frecuencias absolutas o frecuencias relativas de las clases.
- Paso 3: Para la primera clase, construya un rectángulo cuya base es el intervalo de clase y la altura es la frecuencia absoluta simple (o relativa) de esa clase;
- Paso 4: Para la clase siguiente, construya un rectángulo adyacente al primero cuya base es el intervalo de la clase y la altura es la frecuencia absoluta o relativa de esa clase.
- Paso 5: Repita el procedimiento para las demás clases.

En la siguiente figura se representa la apariencia que tendrá un histograma



Nota
Cuando se construyen histogramas, el eje vertical debe mostrar el cero verdadero para no distorsionar o representar equivocadamente el tipo de datos. Sin embargo, no es necesario que el eje horizontal especifique el punto cero del fenómeno de interés. Por razones de estética, el rango de la variable debe constituir la principal porción de la gráfica y, cuando no se incluye el cero, resulta apropiado incluir "fracturas" ( ) en el eje.

Ejercicio:
Construir un histograma para la distribución de frecuencias de la variable peso.

Cuando todas las clases de una distribución de frecuencias tienen la misma amplitud, y el histograma se construye utilizando como altura las frecuencias relativas de clase, se permite la comparación de histogramas correspondientes a distribuciones de frecuencias de datos de la misma naturaleza que difieren en cuanto al número de datos. Cuando se utilizan las frecuencias

absolutas como alturas, no es posible comparar histogramas diferentes excepto en casos especiales.

Histograma para distribuciones de frecuencias con clases de diferente amplitud

Cuando en la distribución de frecuencias que se ha construido no todas sus clases tienen la misma amplitud, se debe tener en cuenta lo siguiente para la construcción del histograma. La altura de cada rectángulo del histograma debe ser igual a la frecuencia relativa de la clase dividida entre la respectiva amplitud, es decir:

$$h_i = \frac{fr_i}{c_i}$$

De esta manera, las áreas de los rectángulos del histograma serán iguales a la proporción de datos contenidos en la respectiva clase, es decir, a las frecuencias relativas correspondientes. En otras palabras, el tamaño de los rectángulos construidos de este modo, refleja la proporción de datos contenida en esa clase.

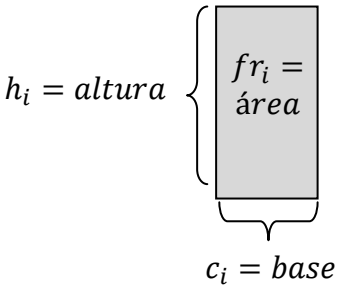
Para entender por qué esto es así, hay que recordar que el área de un rectángulo es:

$$área = base \times altura$$

y si se quiere hallar la altura, despejando queda que:

$$altura = \frac{área}{base}$$

Esta última expresión, en la notación utilizada en distribuciones de frecuencia y para el i-ésimo rectángulo es equivalente a:



De esta forma, el área total del histograma es igual a 1. Esto es así, ya que al sumar todas las áreas de los rectángulos, lo que realmente se está haciendo es $\sum fr_i$ que es igual 1.

El histograma construido de este modo es válido tanto para distribuciones de frecuencias cuyas clases tienen todas la misma amplitud como para las que las amplitudes no son todas iguales.

El histograma construido de esta manera permite su comparación con otros histogramas correspondientes a distribuciones de frecuencias de datos de la misma naturaleza que difieran en cuanto al número de datos y a la manera como están establecidas las clases.

Ejercicio:

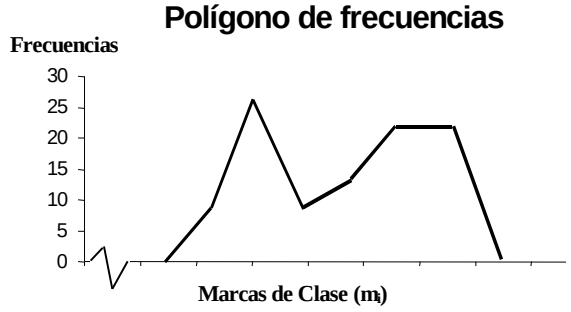
1. Para la distribución de frecuencias de la variable ingreso:
   a. Construir un histograma con altura de los rectángulos igual a fr_i (frecuencias relativas)
   b. Construir un histograma con altura de los rectángulos igual a h_i.
   c. ¿Cuál de los dos histogramas anteriores representa adecuadamente a los datos agrupados en la distribución de frecuencias de la variable ingreso? Explique las razones de su elección.
2. Construya dos histogramas con altura de los rectángulos igual a h_i, uno para los pesos de los varones y otro para los pesos de las hembras. Compare ambos gráficos.

Nota

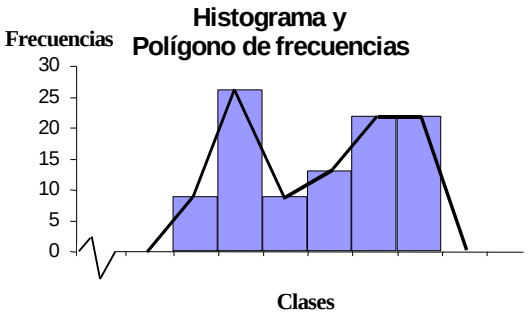
En un histograma, al eliminar los espacios entre las barras se logra que la gráfica lleve consigo una "continuidad" que refleja que los datos son continuos. En los *gráficos de barras* (para datos cualitativos), deben aparecer los espacios entre las barras para evitar que el lector interprete una "continuidad" de los datos, ya que las categorías de las variables cualitativas son ordenadas en general de forma arbitraria y por definición no pueden tomar todos los valores en el eje horizontal.

Polígono de frecuencias

Una alternativa para un histograma, es el *polígono de frecuencias.* En el eje horizontal se escriben las *marcas de clase* de cada intervalo y para cada una de estas m_i se colocan las alturas en el eje vertical, las cuales vienen dadas por las frecuencias respectivas (absolutas simple o relativas). Luego, se marcan los puntos (m_i , fr_i *ó* m_i , f_i) y se une con rectas en el plano cartesiano. Para cerrar la curva resultante con el eje de las abscisas, se crean dos puntos medios ficticios, uno anterior al de la primera clase y otro posterior al de la última clase cada uno con frecuencia igual a cero. De esta manera se obtiene el polígono de frecuencias:



Nótese como el polígono puede obtenerse directamente del histograma:

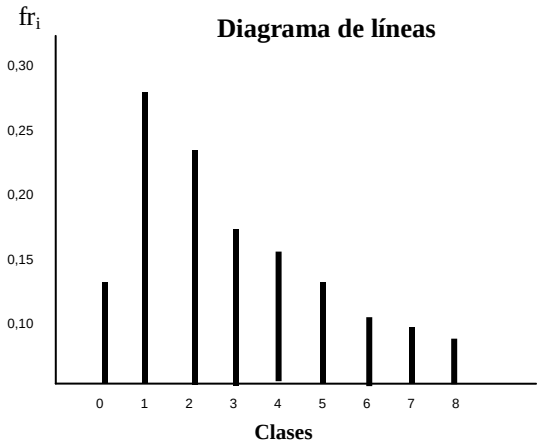


Ejercicio:
Construir un polígono de frecuencias para la distribución de la variable peso.

Diagrama de líneas de frecuencias
Es el equivalente al histograma en una distribución de frecuencias cuyas clases son valores individuales de la variable.

Construcción
- Paso 1: En el eje horizontal, marque sucesivamente las clases.
- Paso 2: En el eje vertical, marque, en la escala, los valores relativos a las frecuencias absolutas o frecuencias relativas de las clases.
- Paso 3: Para la primera clase, trace una línea vertical cuya altura es la frecuencia absoluta simple (o relativa) de esa clase;
- Paso 4: Repita el procedimiento para las demás clases.

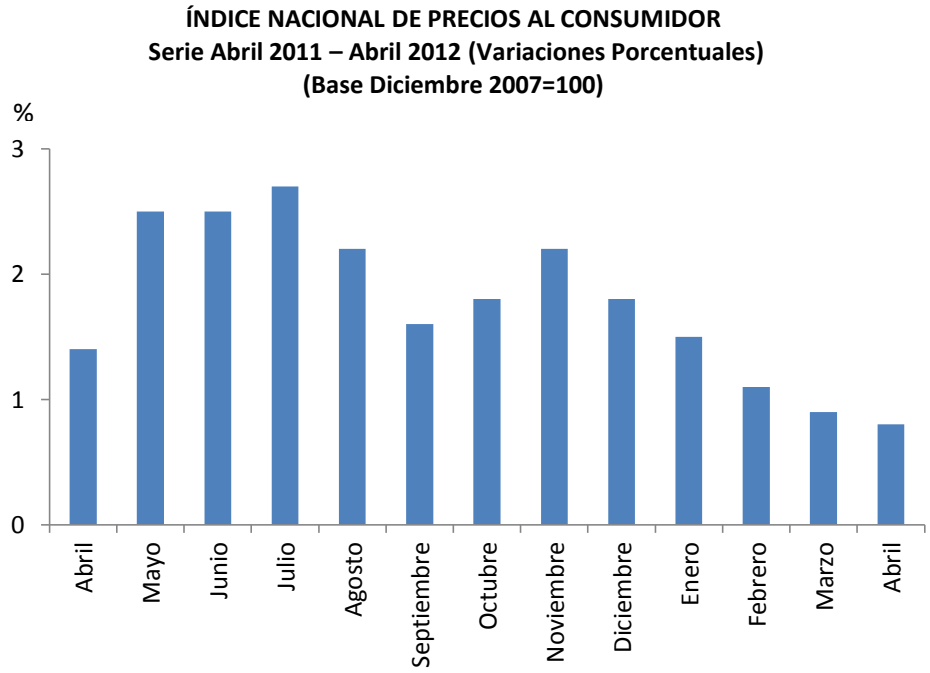
En la siguiente figura se representa un diagrama de líneas.

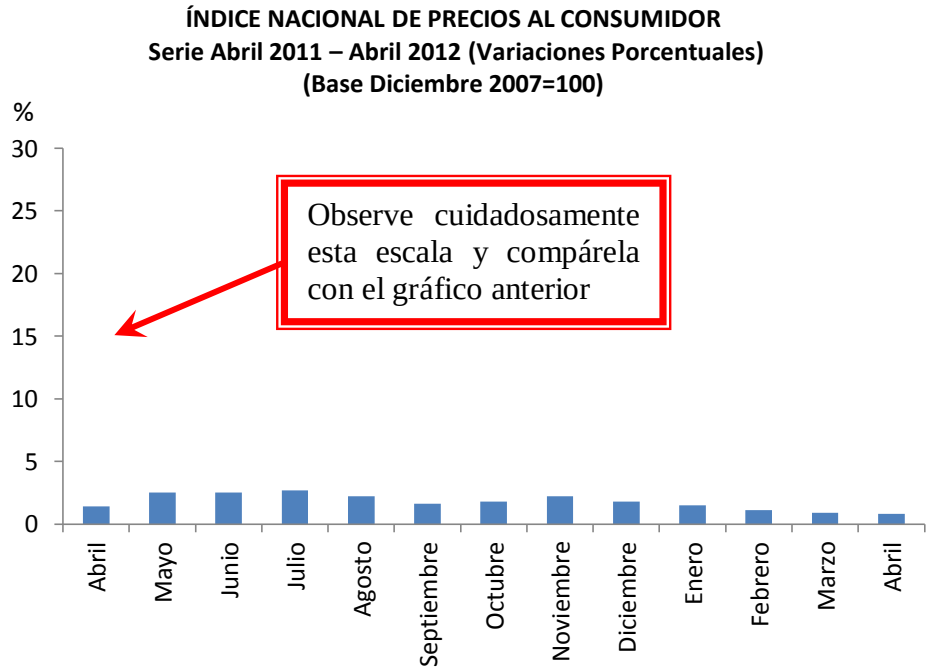


Ejercicio:
Construir un diagrama de líneas para la distribución de frecuencias de la variable número de hermanos.

Gráficos engañosos

¡Cuidado! Cuando se observa un gráfico, particularmente como parte de un anuncio, sea cauteloso. Fíjese en las escalas utilizadas en los ejes vertical y horizontal. Se puede distorsionar la verdad con las técnicas estadísticas, tal como se muestra a continuación:





Descripción de la forma en que se distribuyen los datos

Los gráficos para distribuciones de frecuencias vistos anteriormente sirven para proporcionar una idea a primera vista acerca de la forma en que se distribuyen los datos. En este sentido se tienen las siguientes definiciones:

Distribución simétrica

Una distribución de frecuencias es *simétrica* (o con sesgo cero) con respecto al valor central de la distribución, digamos x_0, cuando el gráfico a la izquierda de x_0 es el "espejo" de la derecha. En otras palabras, si a la izquierda y a la derecha de x_0 existe la misma cantidad de datos la distribución será simétrica.

Distribución asimétrica

Si una distribución no es simétrica, se dice que es asimétrica (o sesgada). Existen dos casos de asimetría:

Asimetría Positiva o por la derecha

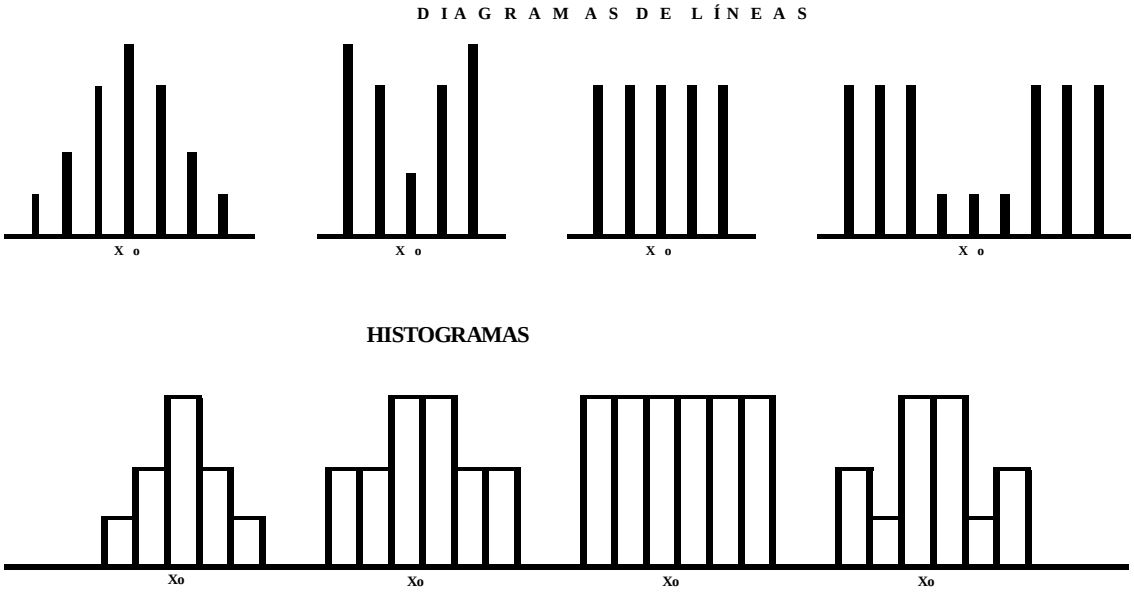
Este tipo de asimetría se presenta cuando existe una mayor concentración de datos en las primeras clases en comparación con las últimas. Se puede visualizar fácilmente cuando el extremo o "cola" de la derecha del gráfico se prolonga más que el de la izquierda.

Asimetría Negativa o por la izquierda

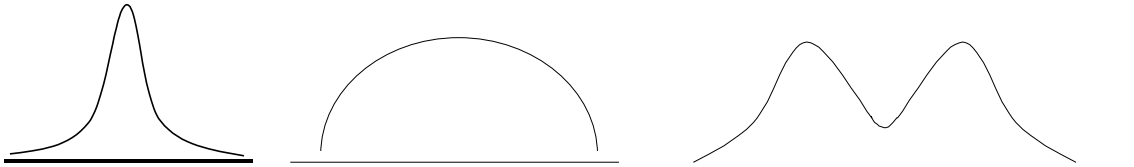
Cuando hay mayor concentración de datos en las últimas clases en comparación con las primeras, es decir, cuando la "cola" izquierda de la curva se prolonga más que la derecha se dice que, la distribución de frecuencias es *asimétrica negativa o por la izquierda*.

Ejemplo:

A continuación se presentan algunos casos de distribuciones simétricas:



POLÍGONOS



Ejemplo**:**

A continuación se presentan dos casos de distribuciones asimétricas**:**

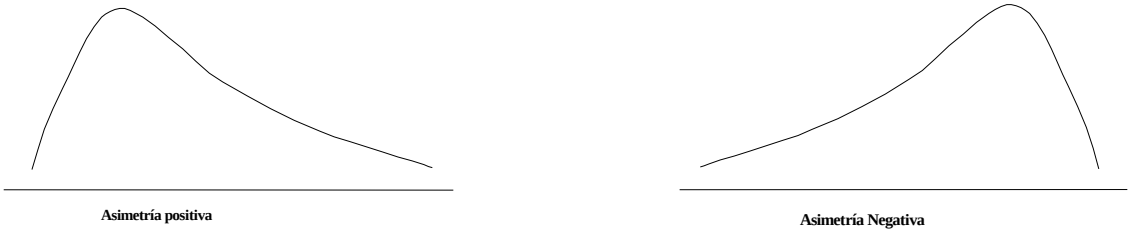


Diagrama de Tallo y Hojas[2]

Un diagrama de *tallo y hojas* es una representación visual de los datos que es a la vez una tabla y un gráfico. Es como un histograma horizontal con el cual se puede visualizar rápidamente la distribución de los datos. El diagrama de *tallo y hojas* provee más detalles que un histograma, ya que cada punto del gráfico representa un valor individual de los datos.

Construcción

Para ejemplificar la elaboración de un diagrama de tallo y hojas, considérese los datos de la Tabla 1, que representan la duración de 40 baterías de carro similares. Las baterías estaban garantizadas para durar 3 años.

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Tabla 1. Duraciones de las baterías de un automóvil.

Primero, se divide cada observación en dos partes que consisten en un tallo y una hoja de tal forma que el primero represente el dígito que es el entero y la hoja corresponda a la parte decimal del número. En otras palabras, para el número 3.7 el dígito 3 se designa como el tallo y el dígito 7 como la hoja. Los cuatro tallos, 1, 2, 3 y 4 quedan listados consecutivamente en el lado izquierdo de la línea vertical de la Tabla 2; las hojas se escriben en el lado derecho de la línea en contraposición al valor de tallo apropiado.

Entonces, la hoja 6 del número 1.6 se escribe a la altura del tallo 1; la hoja 5 del número 2,5 se escribe a la altura del tallo 2; y así sucesivamente. La cantidad de hojas registradas para cada tallo se resume en la columna de frecuencia.

El diagrama de tallo y hojas de la Tabla 2 contiene sólo 4 tallos y, en consecuencia, no proporciona una imagen adecuada de la distribución. Para evitar este problema, se necesita incrementar el número de tallos en la tabla. Una forma sencilla de llevar a cabo esto es escribir

El *tallo* de 1.9 → 1 69 ← La *hoja* de 1.9

Tallos	Hojas	Frecuencia
1	69	2
2	25696	5
3	43185147236282971300971 45	25
4	71354172	8

Tabla 2. Diagrama de tallo y hojas de las duraciones de las baterías.

[2] Basado en Walpole, R. y Myers, R. (1993) *Probabilidad y Estadística*. Págs. 59-61.

Tallos	Hojas	Frecuencia
1 S	69	2
2 I	2	1
2 S	5696	4
3 I	431142322130014	15
3 S	8576897975	10
4 I	13412	5
4 S	757	3

Tabla 3. Diagrama de doble tallo y hojas para las baterías.

cada valor de tallo dos veces en el lado izquierdo de la línea vertical y entonces registrar las hojas 0, 1, 2, 3 y 4 en el nivel del tallo que les correspondió originalmente; y las hojas 5, 6, 7, 8 y 9 a la altura del tallo registrado la segunda vez. Este diagrama modificado de doble tallo y hojas se muestra en la Tabla 3, donde los tallos que corresponden a las hojas 0, 1, 2, 3 y 4 han sido identificados con el símbolo I (inferior), y los tallos correspondientes a las hojas 5 a 9 por el símbolo S (superior).

Puede lograrse un incremento adicional en el número de tallos al escribir cada valor del tallo cinco veces en el lado izquierdo de la línea vertical donde se podría, ahora, identificar con la letra *a* al tallo para las hojas 0 y 1, con la *b* para las hojas 2 y 3, la *c* para las hojas 4 y 5, la *d* para las 6 y 7 y la *e* para las hojas 8 y 9. Para los datos de la Tabla 1 se usarían entonces, los tallos 1*d*, 1*e*, 2*a*, 2*b*, 2*c*, 2*d* y 2*e* para dibujar una tabla de cinco tallos y hojas.

En cualquier problema dado, debe decidirse el valor apropiado de tallos. Esta decisión se toma de manera un poco arbitraria, a pesar de que el tamaño de la muestra sirve de guía. Puede usarse la regla de Sturges como guía de cuantos tallos utilizar, de manera similar a como se hace en la determinación del número de intervalos de clases de una distribución de frecuencias. Usualmente se seleccionan entre 5 y 15 tallos. Entre más pequeña es la cantidad de datos disponibles, menor es el número de tallos seleccionados. Por ejemplo, si los datos consisten en números del 1 al 21 que representan la cantidad de personas en una cafetería en 40 días de trabajo seleccionados aleatoriamente y se decide utilizar un diagrama de doble tallo y hojas, los tallos serían 0I, 0S, 1I, 1S y 2I de tal forma que a la observación más pequeña 1 le corresponde el tallo 0I y la hoja 1, al número 18 le corresponde el tallo 1S y la hoja 8, y a la observación más grande 21 le corresponde el tallo 2I y la hoja 1.

Por otro lado, si los datos consisten en cantidades, desde \$8.800 a \$9.600, las cuales representan los mejores tratos posibles de 100 automóviles nuevos de un cierto distribuidor y se construye un diagrama de tallo y hojas, los tallos serían 88, 89, 90, . . , 96 y ahora cada hoja tendría dos dígitos. Un vehículo que se vende a \$9,385 tendría un valor de tallo de 93 y la hoja de dos dígitos 85.

Las hojas de múltiples dígitos que pertenecen a un mismo tallo generalmente se separan con comas en el diagrama de tallo y hojas. Los puntos decimales en los datos casi siempre se ignoran cuando todos los dígitos a la derecha del punto decimal representan la hoja. Tal es el caso de la Tabla 2 y la Tabla 3. Sin embargo, si los datos consisten en los números en el rango de 21,8 a

74,9, se podrían seleccionar como tallos los dígitos 2, 3, 4, 5, 6, 7, de tal manera que un número, por ejemplo el 48,3, tendría un valor de tallo de 4 y de hoja 8,3.
También se acostumbra a redondear el valor de la variable a la hoja más cercana para presentarla sólo con un dígito. Por ejemplo, si una variable toma el valor 6,25 tendrá un tallo de 6 y una hoja de 3.

Nota
Ante la presencia de valores atípicos el diagrama de tallo y hojas puede requerir una "fractura" ( ) en los tallos.

Ejemplo[3]
A continuación se ilustra el diagrama de tallo y hoja para 25 observaciones del rendimiento por lote de un proceso químico. En el diagrama (a) se han utilizado los números 6, 7, 8, y 9 como tallos. Esto produce muy pocos tallos, con lo que el diagrama no proporciona mucha información sobre los datos. En el diagrama (b) se ha dividido cada tallo en dos partes, con lo que se tiene una presentación más adecuada de los datos. El diagrama (c) ilustra un gráfico en el que cada tallo se ha dividido en cinco partes. En esta gráfica existen muchos tallos, lo que da como resultado una presentación que no dice mucho con respecto a la distribución de los datos.

(a)

Tallo	Hoja
6	134556
7	011357889
8	1344788
9	235

(b)

Tallo	Hoja
6I	134
6S	556
7I	0113
7S	57889
8I	1344
8S	788
9I	23
9S	5

Las *hojas* están ordenadas

(c)

Tallo	Hoja
6a	1
6b	3
6c	455
6d	6
6e	
7a	011
7b	3
7c	5
7d	7
7e	889
8a	1
8b	3
8c	44
8d	7
8e	88
9a	
9b	23
9c	5
9d	
9e	

[3] Tomado de Newbold, P. (1998). *Estadística para los Negocios y la Economía.* Pág. 7.

Ventajas y Desventajas del Diagrama de Tallo y Hoja

Los diagramas de tallo y hojas muestran el centro, la dispersión y forma de la distribución de la misma manera en que lo hace un histograma. Sus ventajas incluyen que cada valor de la variable es representado de forma exacta, es muy fácil determinar la mediana y los percentiles; y también es muy fácil de construir con sólo papel y lápiz. Entre las desventajas se encuentra que es difícil comparar distribuciones cuando el número de observaciones en los conjuntos de datos son muy diferentes; y cuando el conjunto de datos es muy grande el diagrama de tallo y hojas se vuelve impráctico.

Nota

Observe que en el ejemplo anterior las hojas correspondientes a cada tallo están listadas en orden creciente. El orden de las hojas facilita el uso del diagrama de tallo y hojas en la determinación de la mediana y los percentiles. También se debe notar que todos los posibles valores de los tallos están listados, tengan o no observaciones (hojas) como se ve en el diagrama (c).

Construcción de una distribución de frecuencias a partir del diagrama de tallo y hojas

Una distribución de frecuencias en la cual los datos se agrupan en diferentes clases o intervalos, puede construirse con facilidad contando simplemente las hojas que pertenecen a cada tallo y notando que cada uno de ellos define un intervalo. En la Tabla 2 el tallo 1 con dos hojas define el intervalo [1.0-2.0) y contiene dos observaciones; el tallo 2 con 5 hojas define el intervalo [2.0-3.0) y contiene 5 observaciones; el tallo 3 con 25 hojas define el intervalo [3-4) y contiene 25 observaciones y el tallo 4 con 8 hojas define el intervalo [4-5) y contiene 8 observaciones. Para el diagrama de doble tallo y hojas presentado en la Tabla 3, los tallos definen a los 7 intervalos de clase [1.5-2), [2-2.5), [2.5-3), [3-3.5), [3.5-4), [4-4.5) y [4.5-5), con frecuencias 2, 1, 4, 15, 10, 5 y 3, respectivamente.

Clases	m_i	f_i	fr_i
[1.5 – 2.0)	1.75	2	0.050
[2.0 – 2.5)	2.25	1	0.025
[2.5 – 3.0)	2.75	4	0.100
[3.0 – 3.5)	3.25	15	0.375
[3.5 – 4.0)	3.75	10	0.250
[4.0 – 4.5)	4.25	5	0.125
[4.5 – 5.0)	4.75	3	0.075

Tabla 4. Distribución de frecuencias relativas de las duraciones de las baterías.

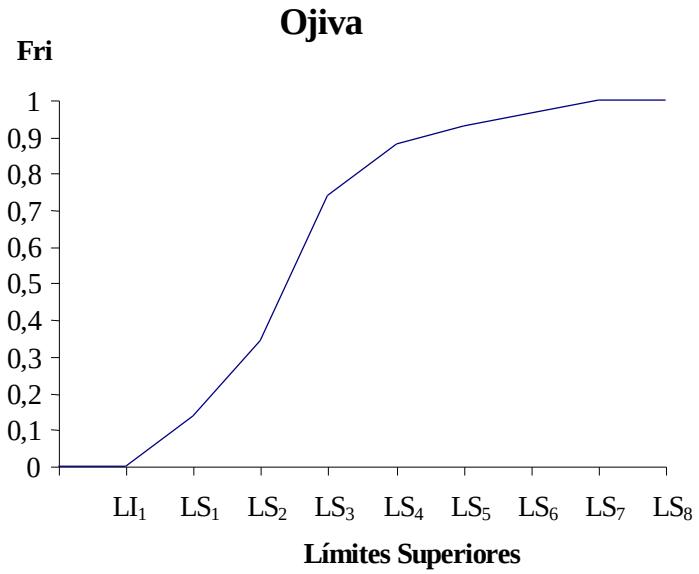
Ojiva (Polígono de frecuencias acumuladas)

Este gráfico se emplea en distribuciones de frecuencias cuyas clases son intervalos. Es un tipo especial de gráfico de curvas en el cual se representan las frecuencias acumuladas.

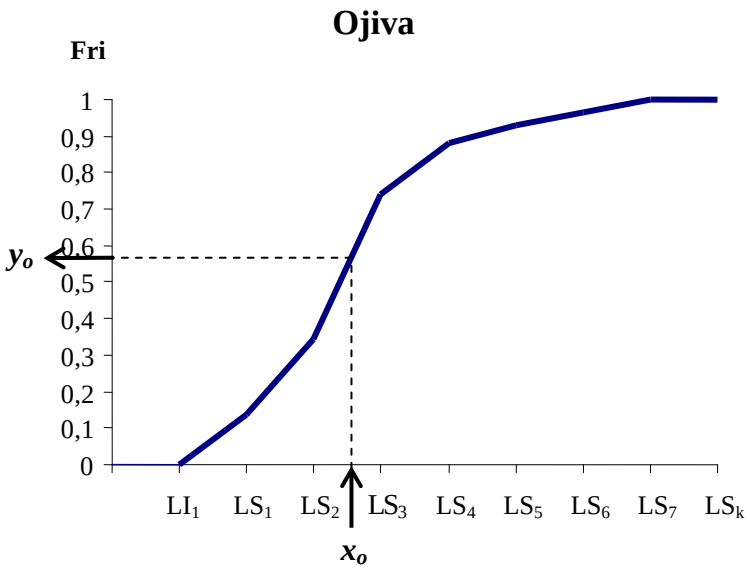
Construcción

- Paso 1: En el eje horizontal, marque sucesivamente los límites superiores de cada clase.
- Paso 2: En el eje vertical, marque los valores correspondientes a las frecuencias acumuladas o frecuencias relativas acumuladas.
- Paso 3: Para cada límite superior de clase se marca con un punto su correspondiente frecuencia acumulada.
- Paso 4: El límite inferior de la primera clase también se señala con un punto en el eje horizontal, asignándole una frecuencia acumulada igual a 0.
- Paso 5: Se unen todos los puntos con segmentos de recta.

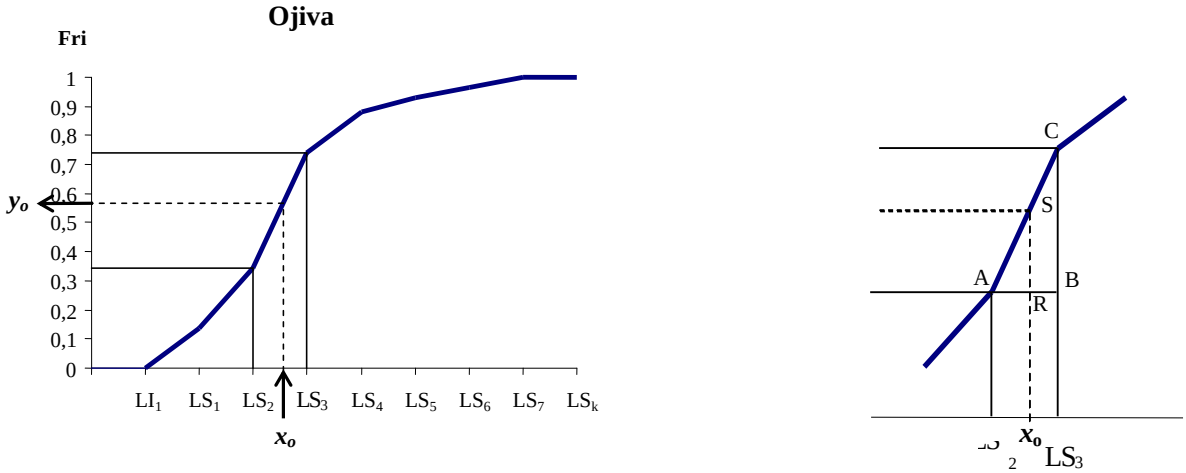
Así se obtiene la *Ojiva*. Nótese que este gráfico es no decreciente.



Las *ojivas* son principalmente usadas para determinar gráficamente y de forma aproximada el número o proporción de datos *que son menores, o que son iguales o mayores* a una valor de interés. Si se usa papel milimetrado para graficar la *ojiva*, se fija el valor x_0 de interés de la variable en estudio el cual es ubicado en el eje horizontal y se levanta desde este valor x_0 una línea perpendicular al eje que llegue hasta la curva. Luego, a partir del punto de intersección se traza una línea paralela al eje de las abscisas; y el punto de corte con el eje vertical, y_0, representa el número o proporción de datos (dependiendo si la *ojiva* se construyó con las F_i ó con las Fr_i) que son inferiores al valor x_0 especificado.



También se puede encontrar la proporción de datos y_0 que son menores que el valor x_0, mediante un proceso de interpolación (si no se usa papel milimetrado) usando la propiedad de triángulos semejantes:



Nótese en el gráfico anterior que el triángulo ABC es equivalente con el triángulo ARS, con lo cual se cumple la propiedad:

$$\frac{\overline{AR}}{\overline{AB}} = \frac{\overline{RS}}{\overline{BC}}$$

Que al aplicarse en nuestro caso, tenemos que:

$\overline{\text{AR}} = x_0 - LS_2$

$\overline{\text{AB}} = LS_3 - LS_2$

$\overline{\text{RS}} = y_0 - Fr_2$

$\overline{\text{BC}} = Fr_3 - Fr_2$

Entonces, sustituyendo en la propiedad de los triángulos semejantes, queda:

$$\frac{x_0 - LS_2}{LS_3 - LS_2} = \frac{y_0 - Fr_2}{Fr_3 - Fr_2}$$

despejando y_0 de la igualdad anterior, se encuentra que:

$$y_0 = \left(\frac{x_0 - LS_2}{LS_3 - LS_2} \right) * \left(Fr_3 - Fr_2 \right) + Fr_2$$

Así, a través de ese método de interpolación se puede encontrar una aproximación a la proporción de datos, igual a y_0, que es menor que el valor x_o.

Ejercicios:

Usando la distribución de frecuencias construida en clase para la variable peso, encuentre mediante el método gráfico de interpolación:

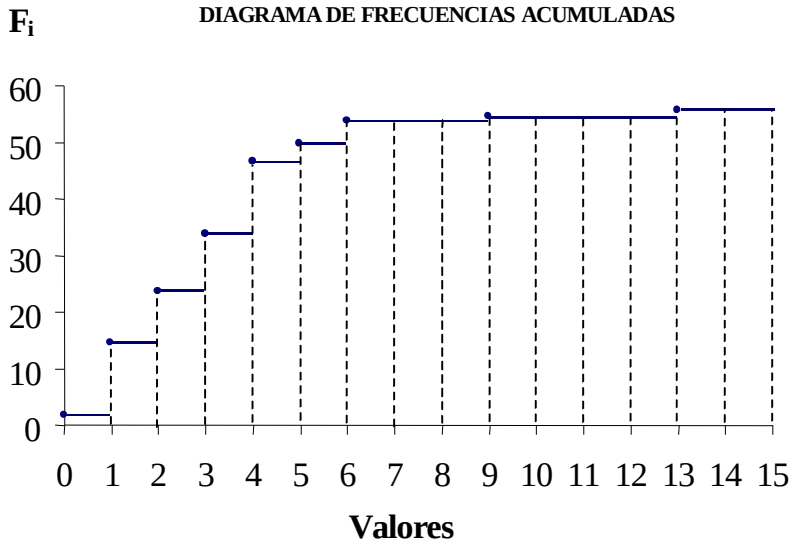
1. La proporción de estudiantes que pesan menos de 78 Kg.
2. La proporción de estudiantes cuyos pesos son menores a 56 Kg.
3. La proporción de estudiantes que tienen un peso mayor o igual a 56 Kg.
4. El porcentaje de estudiantes que pesan menos de 56 Kg.
5. El valor del peso por encima del cual se encuentra el 50% de los pesos de los estudiantes.
6. El peso tal que el 10% de los estudiantes está por debajo de él.
7. El peso tal que la mitad de los datos está por debajo de su valor.

Diagrama de Frecuencias Acumuladas (Gráfico de escalera)

El gráfico equivalente a la *ojiva* en el caso de *distribuciones de frecuencias cuyas clases son valores individuales de la variable en estudio* se denomina *diagrama de frecuencias acumuladas*.

Construcción

- Paso 1: En el eje horizontal, marque sucesivamente los valores de la variable que representan las clases.
- Paso 2: En el eje vertical, marque los valores correspondientes a las frecuencias acumuladas o frecuencias relativas acumuladas (o porcentaje).
- Paso 3: A cada valor de la variable se le representa su frecuencia acumulada mediante una línea horizontal que se prolonga hasta donde está señalado el próximo valor de la variable.
- Paso 4: Al trazar las líneas anteriores, se les coloca un punto al comienzo. Esto indica que al correspondiente valor en el eje horizontal le corresponde esa frecuencia acumulada.



Nótese que el gráfico suministra visualmente el número de datos *menores o iguales* que un valor particular de la variable en estudio.

Comentarios finales

En las aplicaciones prácticas, el objetivo de construir una distribución de frecuencias y su respectivo histograma es conseguir información relevante sobre los datos. En este sentido, la decisión más difícil, pero inevitable, es cuanto ha de detallarse. Si se realiza una presentación muy poco detallada, es decir con muy pocas clases, se pueden ocultar características importantes, mientras que si se cae en el otro extremo, podríamos perdernos en un exceso de detalle. La mejor guía a seguir es el sentido común, aunque pueden enumerarse unas cuantas reglas generales:

1. Como se había comentado antes, para lograr una interpretación más fácil es preferible establecer intervalos de igual amplitud. Sin embargo, en algunas ocasiones habrá que descartar este principio. Si un conjunto de datos tiene muchas observaciones contenidas en muy pocos intervalos de clase, mientras que las otras están muy dispersas en el resto de las clases, será preferible dividir en intervalos de longitud pequeña la zona donde las observaciones están más concentradas, y en intervalos más amplios las observaciones fuera de esta zona. Si se hace esto, es importante tener muy en cuenta que son las *áreas* y no las alturas de los rectángulos del histograma, las que han de ser proporcionales a las frecuencias.

2. Es importante asegurarse que los puntos medios de los intervalos sean representativos de los miembros de esa clase. Por ejemplo, muchos artículos en las tiendas tienen precios de Bs. 9.999, Bs. 10.999, etc. Si se clasifican los precios en intervalos [Bs. 9.000 - Bs. 10.000), [Bs. 10.000 - Bs. 11.000), etc., es muy probable que en cada intervalo de clase predominen los precios próximos al límite superior. Una mejor solución consistiría en establecer clases como: [Bs. 9.500 - Bs. 10.500), [Bs. 10.500 - Bs. 11.500) y así sucesivamente.
   Una razón para elegir puntos medios de clase que sean representativos de los valores de los miembros de esa clase, es que el histograma tendrá un aspecto visual más fidedigno. Además como se verá más adelante, en muchos casos se calculan medidas de centralización y dispersión para datos agrupados. Estos cálculos dependen del supuesto de que *los puntos medios de cada intervalo de clase son representativos de la clase*.

3. Muchas veces, la decisión más difícil de tomar es decidir el número de clases a incluir en una distribución de frecuencias. Si el número de clases es demasiado pequeño, la clasificación resultante puede esconder aspectos importantes de los datos. Si hay demasiadas clases, puede resultar un gráfico quebrado y desigual, difícil de interpretar. En general, como se había recomendado antes, debe usarse un número de clases mayor que cinco y menor que 15. Para conjuntos de datos muy grandes, con muchas observaciones, será razonable establecer más clases. Subdividir, por ejemplo, un conjunto de 20 observaciones en 15 clases pequeñas conllevaría a tener muchas clases vacías o casi vacías. Esto puede ser un problema menos grave si se tienen 200 observaciones.

   Aún teniendo en cuenta los factores enunciados, no siempre estará clara la elección del número de intervalos. En muchos casos, una buena idea es probar varias posibilidades y ver cuál de los histogramas resultantes presenta un aspecto más claro.

Medidas Descriptivas Numéricas

Frecuentemente una colección de datos se puede reducir a una o unas cuantas medidas numéricas sencillas que resumen al conjunto total. Tales medidas son más fáciles de comprender que el conjunto de datos originales o ya agrupados. Tres características importantes de los datos que las medidas numéricas ponen de manifiesto son:

1. El valor central o típico de los datos
2. La dispersión de los datos
3. La forma de la distribución de los datos

Medidas de Posición o Localización (tendencia central)

Las medidas de posición se utilizan para indicar un valor que tiende a tipificar o a ser el más representativo de un conjunto de datos. Las tres medidas que más comúnmente se emplean son la media, la mediana y la moda.

1. **Media**

a. ***Media Aritmética***

La media aritmética es lo que viene a la mente de la mayoría de las personas cuando se menciona la palabra "promedio". Como este término tiene ciertas propiedades matemáticas deseables, es la más importante de las medidas de tendencia central. La media aritmética se calcula al sumar los datos y al dividir este resultado entre el número de valores.

Ejemplo:

Si un granjero quiere conocer el peso promedio de sus ocho cerdos cuyos pesos en kilogramos son: 172, 177, 178, 173, 177, 174, 176, 173; realizará el siguiente cálculo:

$$\frac{172+177+178+173+177+174+176+173}{8}=\frac{1400}{8}=175$$

Es decir, el peso promedio de esos cerdos es 175 Kg.

Dada una colección de datos representada por $x_1, x_2, \ldots, x_n$, la media aritmética de una muestra se denotará por el símbolo $\bar{x}$ (que se lee "equis barra"), y su calculo se puede expresar matemáticamente como:

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n}$$

El procedimiento para calcular la media aritmética es el mismo, independientemente si un conjunto de datos se refiere a las observaciones de la muestra o a todos lo valores de la población. Sin embargo, se utiliza el símbolo μ para la media de una población y N para el número de elementos en la misma:

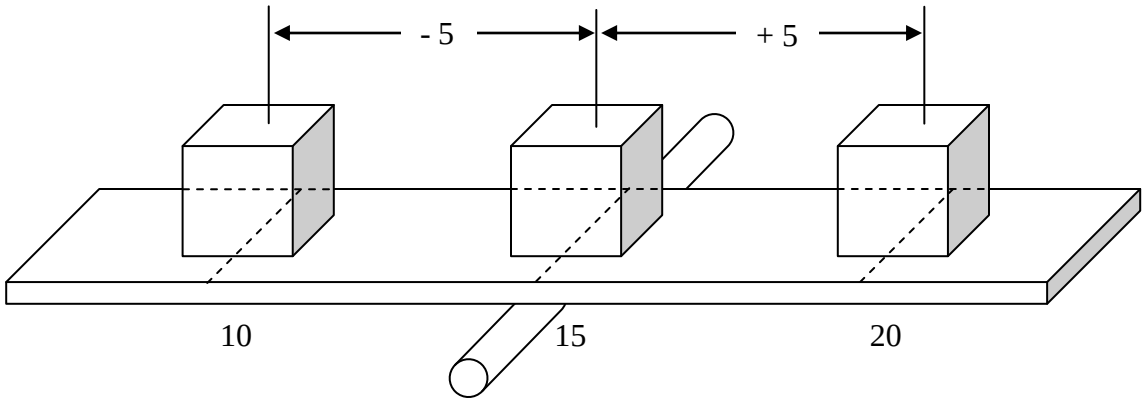
$$\mu=\frac{\sum_{i=1}^{N}x_i}{N}$$

Nota

i. La media aritmética viene expresada en las mismas unidades que los datos originales.

ii. La media aritmética no tiene que coincidir con alguno de los datos de la colección. Como se observa en el último ejemplo el valor $\bar{x}$ =175 Kg. no aparece en los pesos del grupo de cerdos.

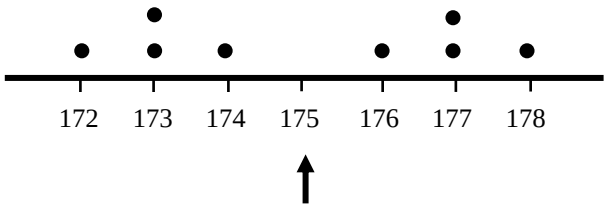
iii. Quizás la manera más adecuada de interpretar la media aritmética sea la que se hace desde el punto de vista de la física, en el sentido de que la media de una serie de datos representa el *centro de gravedad* o *punto de equilibrio* de esos datos.
Una representación física de la media es imaginar una barra con un punto de apoyo central que sostiene pesos iguales en sitios correspondientes a los valores de un conjunto. La media de los números 10, 15 y 20 se puede ilustrar como se observa en la siguiente figura:



Nótese como la media es el *punto de equilibrio* de la tabla; las diferencias positivas y negativas se contrabalancean entre sí.

En el último ejemplo también podemos observar visualmente el punto de equilibrio o centro de gravedad de esos datos:

$\bar{x}$ =175 Kg. constituye el punto en donde se logra el equilibrio.



Nota

No debe interpretarse la media como punto medio de los datos. La media representa el punto de equilibrio de las observaciones, el cual no tiene que ser igual al punto medio. En el gráfico anterior el punto de equilibrio coincide con el punto medio debido a que esos datos se distribuyen *simétricamente*.

Ejercicio:
Para los datos no agrupados, de estudio en clase, calcule la media aritmética para las variables peso, número de hermanos, visitas a la discoteca, visitas al cine, estatura e ingreso mensual del hogar.

b. ***Media Ponderada***
La fórmula de la *media aritmética* supone que cada observación es de igual importancia. Habitualmente, suele suceder así, sin embargo, existen algunas excepciones. Por ejemplo, un profesor informa a su clase que efectuará cuatro parciales. Estos, con respecto a la calificación final del curso equivalen a:
Parcial 1: 20%, Parcial 2: 30%, Parcial 3: 20% y Parcial 4: 30%
El cálculo de la media deberá considerar las diferentes *ponderaciones* de los exámenes. Se conoce como *peso o ponderación* a los factores cuantitativos que modifican a cada uno de los datos.
La media ponderada de una colección de datos $x_1, x_2, \ldots, x_n$, cuyas respectivas ponderaciones son $w_1, w_2, \ldots, w_n$ se define como:

$$\overline{x}_p = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Así un alumno que logre las siguientes calificaciones:

Evaluación	Calificación	Ponderación
1	15	0,30
2	12	0,20
3	19	0,20
4	12	0,30

$$\overline{x} = \frac{0{,}30(15) + 0{,}20(12) + 0{,}20(19) + 0{,}30(12)}{0{,}30 + 0{,}20 + 0{,}20 + 0{,}30} = 14{,}3$$

Obtendrá un promedio de 14,3 puntos. Si todas las evaluaciones poseen la misma importancia, entonces el promedio sería 14,5 puntos. ¿Por qué?

Ejemplo:
Supóngase que el semestre anterior un estudiante cursó Matemática I, Inglés, Métodos Estadísticos I y Sociología, obteniendo las siguientes calificaciones:

Materia	Unidades Crédito	Calificación
Matemáticas I	6	10
Sociología	3	16
Métodos Estadísticos I	5	13
Inglés	3	20

Así el promedio ponderado del estudiante fue de:

$$\bar{x}_p = \frac{6(10)+3(16)+5(13)+3(20)}{6+3+5+3} = 13{,}71 \text{ puntos}$$

y su promedio aritmético simple:

$$\bar{x} = \frac{10+16+13+20}{4} = 14{,}75 \text{ puntos}$$

¿A qué se debe que los dos promedios anteriores sean distintos?

c. ***Media aritmética para datos agrupados en distribuciones de frecuencias***

Es posible utilizar una variante de la fórmula para calcular la media ponderada, a fin de obtener la media de una distribución de frecuencias. Las ponderaciones son sustituidas por las frecuencias absolutas simples y la fórmula se convierte en:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i m_i}{n}$$

Ejercicio:

Calcular la media aritmética para las distribuciones de frecuencias de las variables peso, visitas a la discoteca, estatura, número de hermanos, visitas al cine e ingreso mensual del hogar.

Nota:

En el caso de una *distribución de frecuencias para valores individuales de la variable*, mediante la fórmula se obtendrá la misma respuesta como si se trabajara con datos originales. Si las clases de la distribución de frecuencias son intervalos, el agrupamiento hace que se pierda información y por tanto la media resultante es una aproximación. El uso de los puntos medios de clase (marcas de clase) los considera como promedios de clase, que representan a la clase respectiva, lo cual no siempre se cumple. Sin embargo, si no se dispone de datos originales, no existe otra alternativa razonable. Además la aproximación de esta fórmula a la verdadera media es generalmente buena.

Propiedades de la media aritmética

La media aritmética presenta ciertas propiedades útiles e interesantes, que explican por qué es la medida de tendencia central que se utiliza más ampliamente.

Sea $x_1, x_2, \ldots, x_n$, una colección de datos cuya media aritmética es $\bar{x}$, entonces se cumple que:

i. La suma de las desviaciones o diferencias de cada uno de los datos con respecto a su media, es cero:

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

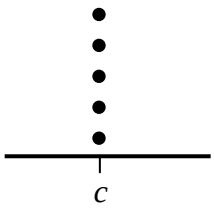
Ejemplo:
En el ejemplo de los pesos de los cerdos se obtuvo que la media aritmética es 175 Kg. Ahora calculando las desviaciones con respecto a $\bar{x}$ =175 se tiene que:

$$\begin{aligned} 172 - 175 &= -3 \\ 177 - 175 &= +2 \\ 178 - 175 &= +3 \\ 173 - 175 &= -2 \\ 177 - 175 &= +2 \\ 174 - 175 &= -1 \\ 176 - 175 &= +1 \\ 173 - 175 &= -2 \\ \hline & \quad 0 \end{aligned}$$

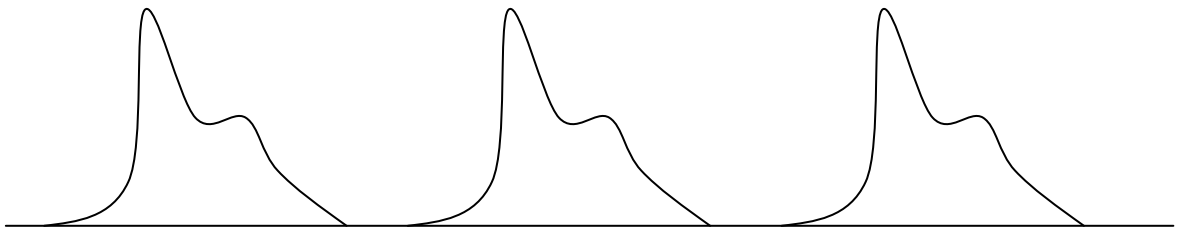
ii. $\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$ es un valor mínimo.

Si se calcula la expresión anterior sustituyendo $\bar{x}$ por cualquier otro valor arbitrario que se nos ocurra, se obtiene un valor mayor al que se consigue utilizando $\bar{x}$.

iii. Si todos los datos son iguales a un mismo valor fijo o constante *c*, entonces la media de esos datos también es igual a *c*:



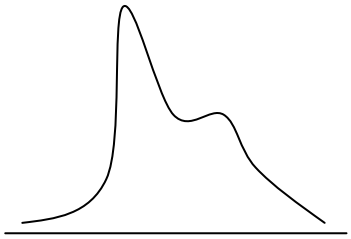
iv. Si a cada uno de los datos originales se le suma un mismo número real *c*, entonces se tiene una nueva colección de datos $x_1 + c, x_2 + c, \ldots, x_n + c$, cuya media viene dada por $\bar{x} + c$.
Esta situación se puede visualizar gráficamente de la siguiente manera:



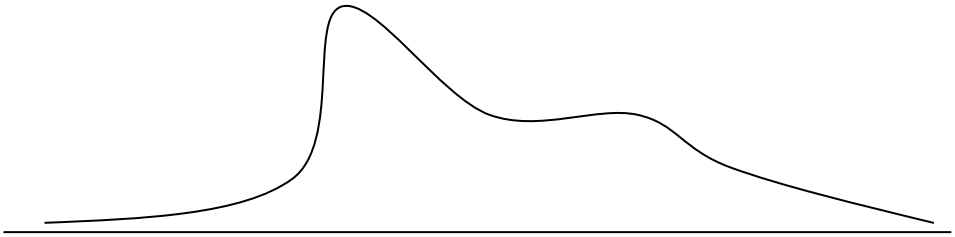
Al sumar la misma constante a cada uno de los datos, realmente lo que estamos haciendo es desplazar sobre el eje horizontal los datos hacia la derecha si la constante

es positiva o hacia la izquierda si la constante es negativa. Entonces la media aritmética se "corre" con los datos.

v. Si cada uno de los datos originales se multiplica por un mismo número real *c*, entonces se genera una nueva colección de datos $x_1 * c, x_2 * c, \ldots, x_n * c$, cuya media viene dada por $\bar{x} * c$.
En la siguiente ilustración se puede observar como se ensancha la distribución de los datos originales cuando estos han sido modificados al multiplicar cada uno por una constante, con lo cual la media se ve afectada.



$x_1, x_2, \ldots, x_n$



$x_1 * c, x_2 * c, \ldots, x_n * c$

vi. Si se tienen *m* diferentes grupos de datos de distintos tamaños $n_1, n_2, \ldots, n_m$ respectivamente, entonces la media de todos esos datos juntos viene dada por:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{m} n_i \bar{x}_i}{\sum_{i=1}^{m} n_i}$$

***Nota*:**

Obsérvese que: $\bar{\bar{x}} \neq \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_m}{m}$

Ejemplo:
Si en un semestre un estudiante aprobó sus cuatro materias con 15 puntos ¿Cuál fue su calificación promedio?
De acuerdo a la propiedad iii. la media aritmética de sus calificaciones fue de 15 puntos.
Ejemplo:

Haciendo referencia al ejemplo de los pesos de los cerdos, suponga que al granjero le han recomendado un nuevo alimento para cerdos que según parece los engorda 20 Kg. en quince días. ¿Cuál será el peso promedio de los cerdos dentro de quince días, luego de utilizar el nuevo alimento?
Nótese que todos los cerdos aumentan 20 Kg., así que a cada uno de los pesos originales se le debe sumar la constante c = 20. En consecuencia, de acuerdo a la propiedad iv. dentro de quince días el peso promedio de los cerdos debe ser 175+20 = 195 Kg.

Ejemplo:
Suponga ahora que todos los cerdos del granjero se enferman a causa de un virus y se detecta cinco días después que todos estos animales han disminuido exactamente 10 Kg. ¿cuál es ahora el peso promedio de los cerdos?

Ejemplo:
Las secciones 03 y 05 de la asignatura Estadística I, tienen 66 y 73 alumnos respectivamente. Se realiza la primera evaluación y se obtienen las siguientes notas promedio por sección: $\bar{x}_1$ = 15 y $\bar{x}_2$ =12. Entonces la nota promedio del primer parcial para las dos secciones juntas es:

$$\bar{\bar{x}} = \frac{66*15+73*12}{139} = 13,42 \text{ puntos}$$

¡OJO es falso que:! $\bar{\bar{x}} = \frac{12+15}{2} = 13,5$

Ejemplo:
Si en el ejemplo de los cerdos, se incluye otro de esos animales cuyo peso es de 490 Kg., Calcule la media aritmética.

$$\bar{x} = \frac{172+177+178+173+177+174+176+173+490}{9} = \frac{1890}{9} = 210$$

Este valor 210 Kg. transmite una idea equivocada de la realidad en cuanto al peso de la mayoría en ese grupo de cerdos. ¿Qué es lo que provoca que $\bar{x}$ no sea representativa de los pesos de los cerdos?
En estos casos no debe utilizarse la media para calcular el peso promedio, sino que se recomiendan otras medidas de tendencia central.

Desventajas de la media aritmética

- No puede calcularse cuando los datos están agrupados en distribuciones de frecuencias que tienen un intervalo de clase abierto.
- La principal desventaja es que se ve afectada por la presencia de *valores extremos o atípicos* en los datos.

Ventajas de la media aritmética

- Es un promedio que toma en cuenta todos los valores de una colección de datos.
- Es fácil de calcular y se presta a operaciones algebraicas, lo que la convierte en la medida de tendencia central más utilizada tanto en estudios descriptivos como para realizar inferencias.
- En general, para una serie dada de datos existe una buena aproximación entre el valor de la media para los datos no agrupados y la media de los datos agrupados.

2. **Mediana**

La mediana de una colección de datos, que previamente han sido *ordenados*, es aquél valor más central o que está más en medio en el conjunto de datos. En otras palabras, la mediana es *mayor* que aproximadamente la mitad de los datos y *menor* que (aproximadamente) la otra mitad. Así se tiene que aproximadamente 50% de las observaciones se encuentran por arriba y 50% (aproximadamente) por debajo de ella. La mediana se denota *Md* (también algunos autores la denotan como $\tilde{X}$.

Ejemplo:
Los tiempos de los miembros de un equipo de atletismo en una carrera de 1,6 Km están dados en la siguiente tabla, calcule la mediana.

Miembro	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Tiempo (en minutos)	4.2	9.0	4.7	5.0	4.3	5.1	4.8

En primer lugar se deben ordenar los datos: 4.2 4.3 4.7 4.8 5.0 5.1 9.0.

Mediana

Md = 4.8 minutos, es el valor que está en el centro de los datos.

Ejemplo:
Calcule la mediana para el número de pacientes tratados en la sala de emergencias de un hospital durante ocho días consecutivos:

Día	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
No. de pacientes	49	52	86	30	35	31	43	11

Los datos ordenados son: 86 52 49 43 35 31 30 11

Centro de los datos

La mediana en este caso puede ser 43 ó 35, o también cualquier valor entre 43 y 35. Para evitar esta imprecisión, se acepta tomar como mediana la suma de los dos valores centrales y se dividen entre dos: $Md = \frac{43+35}{2} = 39$.

Nota:
Si se tienen *n* observaciones ordenadas, la mediana es la observación que ocupa la posición $\frac{n+1}{2}$ cuando *n* es impar y la media de las observaciones que ocupan las posiciones $\frac{n}{2}$ y $\frac{n+2}{2}$ cuando *n* es par.

Ejemplo:
Regresando al ejemplo de los tiempos del equipo de atletismo, se pide calcular la media y comparar este resultado con el de la mediana ya obtenida.

Entonces, se obtiene que $\bar{x}$ = 5.3 minutos y antes se obtuvo que *Md = 4.8* minutos. Nótese que en esos datos existe un *valor atípico*: 9.0 minutos. Por tanto, la media aritmética $\bar{x}$ se distorsiona. La mediana, en cambio, *no se ve distorsionada* por la presencia del valor 9.0. Este valor pudo haber sido 15.0 o incluso 45.0 y la mediana ¡seguirá siendo la misma!

Cálculo de la mediana para datos agrupados en distribuciones de frecuencias

i. Cuando las clases son intervalos

- Se ubica la *clase medianal*, la cual viene dada por aquella clase que contiene a la frecuencia acumulada $n/2$ o equivalentemente a la frecuencia relativa acumulada 0,5.
- Luego de ubicada la clase medianal, el cálculo de la mediana se hace mediante un proceso de interpolación el cual conduce a la siguiente fórmula:

$$Md = LI_m + \left(\frac{\frac{n}{2} - F_{am}}{f_m} \right) * C_m$$

en donde,
LI_m: Límite inferior de la clase medianal
n: No. total de observaciones o datos
F_{am}: Frecuencia acumulada anterior a la clase medianal
f_m: Frecuencia absoluta de la clase medianal
c_m: Amplitud de la clase medianal

Ejemplo: Calcular la mediana para la distribución de frecuencias de la variable peso.

En primer lugar se debe ubicar la clase medianal, para esto se debe calcular:

$$\frac{n}{2} = \frac{43}{2} = 21{,}5$$

Ahora se ubica la frecuencia acumulada que contiene a 21,5:

Clases	mi	fi	fri	Fi	Fri
[40-49)	44,5	4	0,0930	4	0,0930
[49-58)	53,5	10	0,2326	14	0,3256
[58-67)	62,5	15	0,3488	29	0,6744
[67-76)	71,5	7	0,1628	36	0,8372
[76-85)	80,5	5	0,1163	41	0,9535
[85-94)	89,5	1	0,0233	42	0,9767
[94-103)	98,5	1	0,0233	43	1
Totales		**43**	**1**		

Clase medianal →

Frecuencia acumulada que contiene a 21,5

También se puede ubicar la clase medianal encontrando la frecuencia relativa acumulada que contiene a 0,5000.

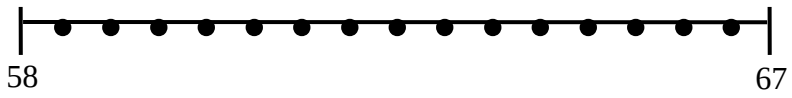
Entonces, se tiene que:

$$Md = 58 + \left(\frac{21{,}5 - 14}{15}\right) * 9$$

$$Md = 62{,}5$$

De esta manera, *Md* = 62,5 Kg. representa el valor central de los pesos. Es decir, aproximadamente la mitad de los estudiantes de Métodos Estadísticos I tienen un peso inferior a 62,5 Kg. y aproximadamente la otra mitad pesa más de 62,5 Kg.

Nota:
En la fórmula de la mediana se está suponiendo que los valores en el intervalo de clase que contiene la mediana están *uniformemente espaciados (o equidistantes)*. Entonces, en el ejemplo anterior se está suponiendo que los 15 valores que contiene la clase medianal están uniformemente espaciados en [58 -67):



Ejercicio:
Calcule la mediana para las distribuciones de frecuencias correspondientes a las variables estatura, índice académico e ingreso mensual del hogar.

ii. Cuando las clases son valores individuales

- Se calcula $n/2$ (o se considera el valor 50% de las observaciones)
- Si el valor $n/2$ **NO APARECE** en la columna de la F_i, entonces se ubica aquella frecuencia acumulada que lo contiene y la mediana será el valor de la variable correspondiente a esa frecuencia acumulada.

- Una forma equivalente de hacer lo anterior es la siguiente, si el valor 50% no aparece en la columna de las Fr_i * 100 entonces se ubica aquella frecuencia que lo contenga y la mediana será el valor de la variable correspondiente a esa clase.
- Si el valor $n/2$ **APARECE** en la columna de las F_i , es decir que coincide con la frecuencia acumulada de alguna clase, entonces la mediana viene dada por la media aritmética de ese valor de la variable y el siguiente valor.
- También, si el valor 50% coincide con alguna de las Fr_i * 100, entonces la mediana viene dada por el promedio de los valores de la variable correspondiente a esa clase y a la siguiente.

Ejemplo:
La siguiente distribución de frecuencias corresponde al número de materias que cursan 112 estudiantes de la carrera de Contaduría Pública. Calcule la mediana.

Inicialmente se debe calcular $\frac{n}{2} = 56$. Entonces 56 no aparece en la columna de las F_i.

Por tanto, *Md* = 4 materias.

Número de materias	***fi***	***fri***	***Fi***	***Fri***
1	1	0,0089	1	0,0089
2	1	0,0089	2	0,0179
3	12	0,1071	14	0,1250
4	56	0,5000	70	0,6250
5	40	0,3571	110	0,9821
6	2	0,0179	112	1,0000
Totales	**112**	**1,0000**		

Mediana → 4

Frecuencia acumulada que contiene a $\frac{n}{2} = 56$ → 70

Ejemplo:
Calcule la mediana para la siguiente distribución de frecuencias, en donde $\frac{n}{2} = 30$. Es decir, $n/2$ aparece en la columna de las frecuencias acumuladas:

Clase	**f_i**	**F_i**	**Fr_i**
5	8	8	0,1333
6	9	17	0,2833
7	13	30	0,5000
8	10	40	0,6667
9	6	46	0,7667
10	14	60	1
	n = 60		

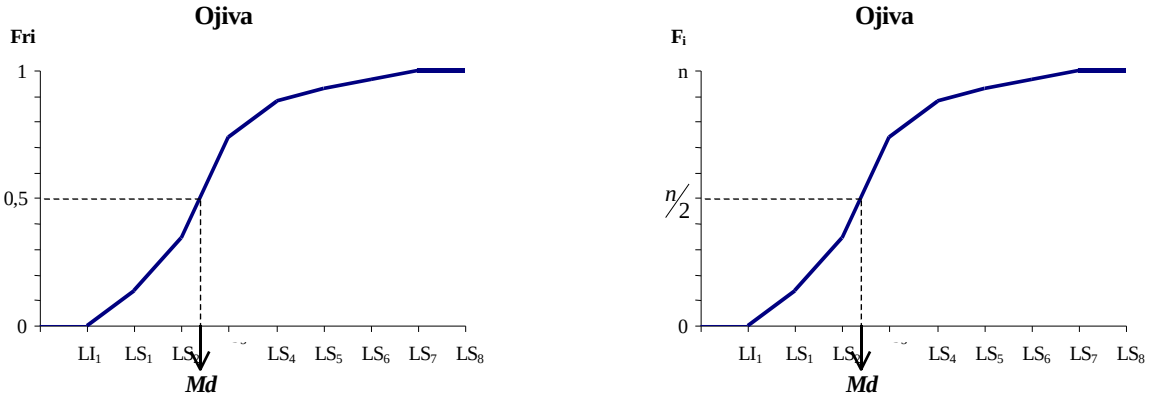
Mediana → 7

Frecuencia acumulada que coincide con $\frac{n}{2} = 30$ → 30

Entonces la mediana viene dada por: $Md = \frac{7+8}{2} = 7,5$

La mediana gráficamente

Mediante la ojiva y a través del método de interpolación visto en esa sección se puede obtener de manera gráfica el valor de la mediana de una colección de datos agrupados en una distribución de frecuencias cuyas clases son intervalos. Si se usa la ojiva construida con la frecuencia acumulada Fi la mediana será aquél valor en el eje horizontal cuya ordenada sea $n/2$. En el caso de usar la ojiva construida con la frecuencia relativa acumulada Fri (o Fri*100), la mediana vendrá dada por el valor en el eje de las abscisas que corresponda a la ordenada 0,5 (o 50%).



Así aplicando el método de interpolación visto antes se obtiene la fórmula del cálculo de la mediana:

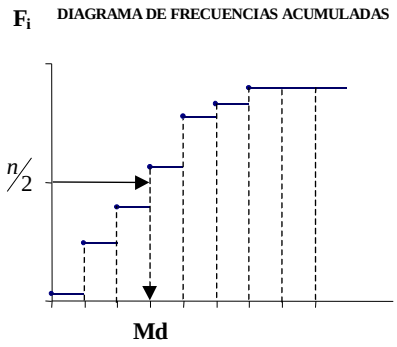
$$Md = LI_m + \left(\frac{\frac{n}{2} - F_{am}}{f_m} \right) * C_m$$

Ejercicio:

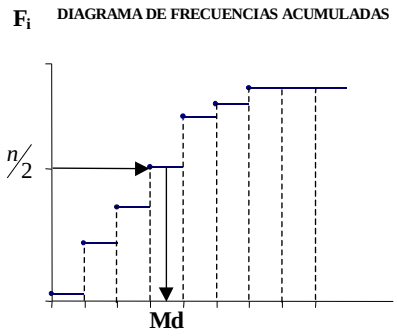
Obtenga gráficamente la fórmula anterior para el cálculo de la mediana.

Para el caso de distribuciones de frecuencias cuyas clases son valores individuales de la variable, se puede hallar gráficamente la mediana por medio del diagrama de frecuencias acumuladas. El procedimiento es similar que cuando se usa la ojiva. Se ubica en el eje vertical $n/2$ (o 50% si se usó Fri*100) y se traza una línea paralela al eje horizontal, así se presentan las dos situaciones siguientes:

- Si la línea intercepta el gráfico, entonces la mediana viene dada por el valor en el eje de las abscisas que corresponde a la ordenada $n/2$ (o 50%).



- Si la línea coincide con uno de los escalones del gráfico, la mediana vendrá dada por el punto medio de ese escalón.



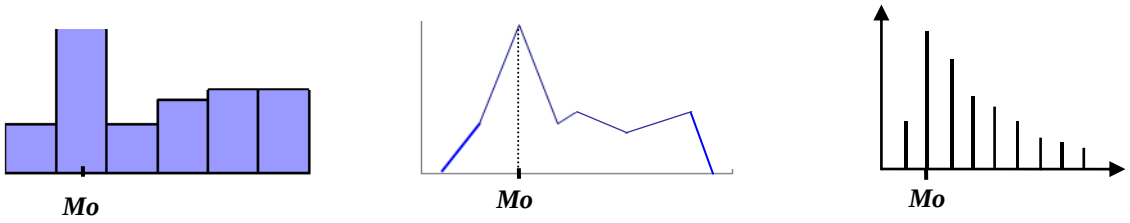
Propiedades de la mediana

i. La mediana es una medida de tendencia central de fácil comprensión pero que solamente toma en cuenta la posición que ocupan las observaciones y no el valor en sí de las mismas. Esto hace que la mediana no sea susceptible de operaciones algebraicas y en consecuencia limita su utilidad, por ejemplo para fines de inferencia estadística.

ii. Puede calcularse en el caso de distribuciones de frecuencias con clases abiertas siempre y cuando se disponga de la información correspondiente a la clase medianal.

iii. No se ve afectada ante la presencia de unos pocos *valores atípicos* y es por ello que se recomienda su uso en el caso de distribuciones marcadamente asimétricas.

Distribuciones unimodales:



3. **Moda**

La moda es el valor que más se repite, es decir el que aparece con mayor frecuencia. En otras palabras la moda es el valor más común de los datos, se denota por *Mo* y viene expresada en las mismas unidades que los datos.

Ejemplo:
Calcule la moda de los siguientes datos: 5, 3, 6, 5, 4, 5, 2, 4.
En este caso el valor que más se repite es el 5, por tanto $Mo = 5$.

Ejemplo:
Calcule la moda de los siguientes datos: 5, 3, 6, 5, 4, 5, 2, 4, 4.
En este conjunto de datos existen dos valores que se repiten con la misma frecuencia: 4 y 5. Así, se tienen dos modas: $Mo_1 = 4$ y $Mo_2 = 5$.

Ejemplo:
Calcule la moda de los siguientes datos: 5, 3, 3, 5, 6, 2, 6, 4, 2, 4.
En este caso no existe la moda dado que no hay datos que se repitan más que otros.

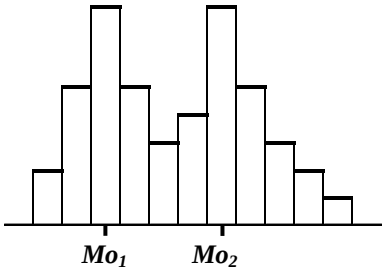
En conclusión, una colección de datos puede que no tenga moda o puede ser que posea una o más modas.

Nota:
Cuando hay una sola moda la distribución de datos se llama *unimodal*, con dos modas *bimodal*, con tres modas *trimodal* y con 4 o más modas se llama *polimodal* o *multimodal*. Si todos los valores se presentan la misma cantidad de veces, la distribución se llama *amodal*.

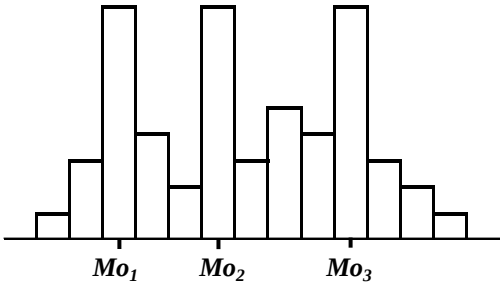
Cuando los datos están agrupados en distribuciones de frecuencias cuyas clases presenten igual amplitud, se toma el *punto medio* de la clase con mayor frecuencia absoluta como la moda.

Representación gráfica de la moda:

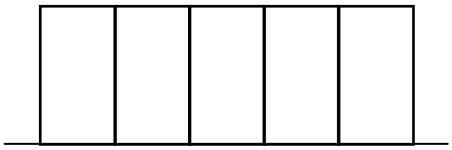
Distribución bimodal:



Distribución trimodal:



Distribución amodal:



Ejercicio:
Calcular la moda para las distribuciones de frecuencias correspondientes a las variables peso, número de hermanos, estatura, ingreso mensual del hogar, número de veces que visita la discoteca e índice académico.

*Propiedades de la Moda***:**

i. La moda en realidad no es una medida de tendencia central, sino más bien indica punto(s) de concentración de datos.
ii. No es susceptible de operaciones algebraicas y de allí que su uso es limitado.
iii. Es la única de las medidas descriptivas que puede utilizarse para datos cualitativos de cualquier tipo.
iv. Es posible su cálculo en algunos casos de distribuciones de frecuencias con intervalos de clase abiertos.
v. Es una medida muy imprecisa e inestable. En una distribución de frecuencias depende de la forma en como se construyen las clases.

Ejemplo:
Considere la siguiente distribución de frecuencias:

Clases	$\mathbf{f_i}$
[0 - 5)	3
[5 - 10)	5
[10 - 15)	6
[15 - 20)	6
[20 - 25)	4
[25 - 30)	7
[30 - 35)	2
Total	**33**

La clase modal es [25 - 30) y la moda es *Mo* = 27,5

Si se introduce una pequeña modificación en las clases, por ejemplo agrupando las dos primeras, se tiene:

Clases	$\mathbf{f_i}$
[0 - 10)	8
[10 - 15)	6
[15 - 20)	6
[20 - 25)	4
[25 - 30)	7
[30 - 35)	2
Total	**33**

La clase modal pasa a ser [0 - 10) y Mo = 5. Obsérvese el cambio tan grande que se produce en la moda ya que pasa de 27,5 a 5.

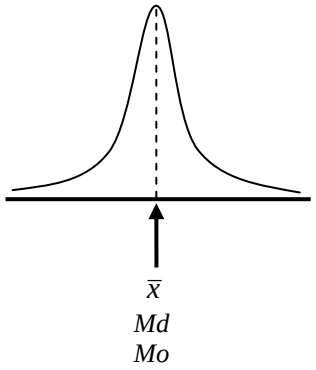
vi. La moda es de utilidad en aquellos casos donde la naturaleza de los datos así lo indique.

Ejemplo:
Para una fábrica de zapatos, el interés está en conocer la o las tallas más frecuentes en la población.

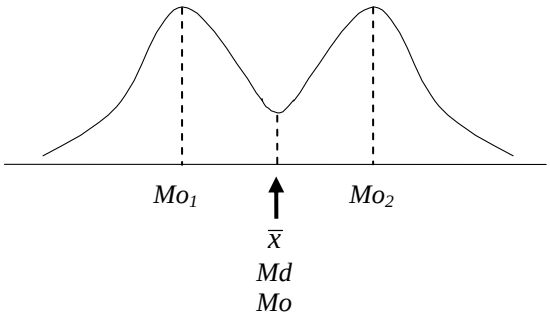
Relación entre la Media Aritmética, la Mediana y la Moda

En función de la simetría de una distribución se presentan las siguientes relaciones entre esas tres medidas:

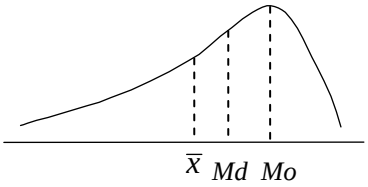
1. En distribuciones simétricas unimodales la media, la mediana y moda coinciden:



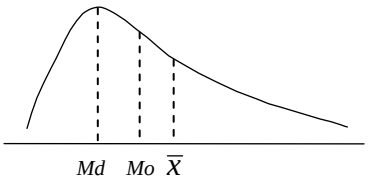
2. En distribuciones simétricas bimodales, la media y la mediana son iguales pero no coinciden con las modas.



3. En distribuciones asimétricas negativas o por la izquierda, se cumple que $\bar{x} < Md < Mo$



4. En distribuciones asimétricas positivas o por la derecha, se cumple que $\bar{x} > Md > Mo$



Selección de la Medida de Posición adecuada

Los siguientes factores deben tomarse en cuenta en el momento de la selección de la medida numérica apropiada para describir la posición o tendencia central de los datos:

1. De acuerdo al tipo de dato se puede utilizar una u otra medida de tendencia central. Las medidas que pueden aplicarse con cada tipo de dato son las siguientes:
   i. Datos Nominales: Moda
   ii. Datos Ordinales: Moda y Mediana
   iii. Datos Discretos: Todas
   iv. Datos Continuos: Todas

2. Teniendo en cuenta lo anterior se recomienda tener presente los siguientes aspectos:

   a. ***La naturaleza de la distribución de los datos***. Gráficamente se puede observar la forma general en que se distribuyen los datos. Esto es determinante en la selección del promedio adecuado.
      - Si se trata de una *distribución simétrica* o aproximadamente simétrica, se sabe que la media, la mediana y la moda coinciden y en consecuencia se puede utilizar cualquiera de ellas.
      - Si la *distribución es asimétrica*, la media aritmética no va a ser adecuada y es preferible inclinarse por la moda o la mediana.

   b. ***El concepto de tendencia central o de posición que interese reflejar en una situación dada.***
      - Si interesa conocer el valor más común de una serie de datos como por ejemplo la estatura típica de las personas que ingresan al ejército, es necesario usar *la moda*.
      - Si se desea ubicar a una persona en cuanto a su salario anual diciendo que gana por encima o por debajo de lo que gana la mitad de los trabajadores del país, entonces habrá que usar *la mediana***.**
      - Cuando interesa el total de datos o reflejar el punto de equilibrio de los mismos se utiliza *la media aritmética*.

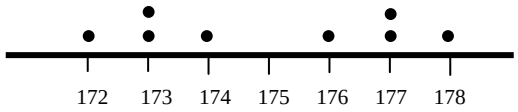
   c. ***Riesgos que se corren ante la presencia de valores atípicos.***
      Si existen valores atípicos, hay que verificar si se incurrió en algún error en la recolección de la información o puede ser el alerta de alguna situación no esperada por el investigador. En todo caso hay que tener presente que la media aritmética se ve seriamente afectada ante la presencia de valores atípicos y será necesario recurrir a alguna de las otras medidas conocidas.

   d. ***Posibilidad de realizar inferencia estadística***
      Cuando el análisis estadístico se realiza sobre una muestra de la población con la intención de generalizar a la totalidad, lo que se conoce como inferencia estadística, prácticamente la única medida de tendencia central utilizada hasta ahora satisfactoriamente es la media aritmética y esto se debe a que existe un fundamento teórico bien fundamentado que la respalda.

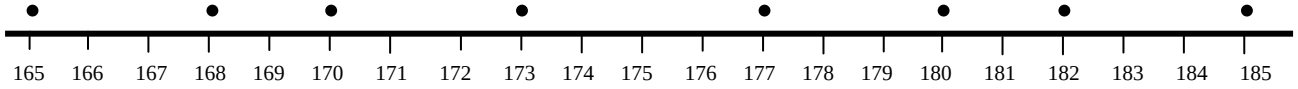
Medidas de Dispersión

Además de obtener la información que reúnen las medidas de tendencia central es muy conveniente tener conocimiento sobre el grado de dispersión o variabilidad que presentan los datos. Las *medidas de dispersión* indican si los valores están relativamente cercanos uno del otro o si se encuentran dispersos. Esta idea se ilustra en las siguientes figuras.

Recuérdese que en el ejemplo de los pesos de los cerdos tenemos los siguientes datos: 172, 177, 178, 173, 177, 174, 176, 173. El diagrama de puntos para esos valores es:



Si los cerdos de otro granjero tienen los siguientes pesos: 165, 182, 185, 168, 170, 173, 180, 177. Entonces el diagrama de puntos está dado por:



Obsérvese que ambos grupos de datos poseen la misma media aritmética y la misma mediana, $Md = \bar{x} = 175$ Kg. Además, también se puede advertir como las observaciones en el primer gráfico tienen valores relativamente más cercanos entre sí en comparación con los pesos del segundo grupo de cerdos.

Por consiguiente, además de las medidas de tendencia central, siempre es importante contar con indicadores que midan la dispersión de los datos. Una medida de tendencia central, casi nunca es suficiente por sí sola, para resumir adecuadamente las características de un conjunto de datos. Por lo general, es necesario, adicionalmente, una medida de la *dispersión* de los datos.

En general, se pueden clasificar las medidas de dispersión en *absolutas* y *relativas*. Las *medidas de dispersión absolutas* son aquellas que vienen expresadas en las mismas unidades que los datos. Las *medidas de dispersión relativas* no vienen expresadas en las unidades de los datos sino en porcentaje.

A pesar de que existen diferentes medidas de dispersión, sólo se van a considerar las más usadas:

Medidas de dispersión absolutas:

- Rango o recorrido
- Varianza
- Desviación Estándar
- Basadas en Percentiles

Medida de dispersión relativa:

- Coeficiente de Variación

Todas estas medidas, excepto el rango, toman la media como punto de referencia. En cada caso un valor cero indica que no hay dispersión, mientras que la dispersión aumenta a medida que se incrementa el valor del indicador (varianza, coeficiente de variación, etc.)

1. **Rango o recorrido**

Esta es la medida más sencilla de calcular y comprender. Se concentra en el valor máximo y mínimo de la colección de datos y viene dada por:

$$R = Valor\ máximo - Valor\ mínimo$$

En el caso de distribuciones de frecuencias, el rango se obtiene restándole al límite superior de la última clase el límite inferior de la primera clase:

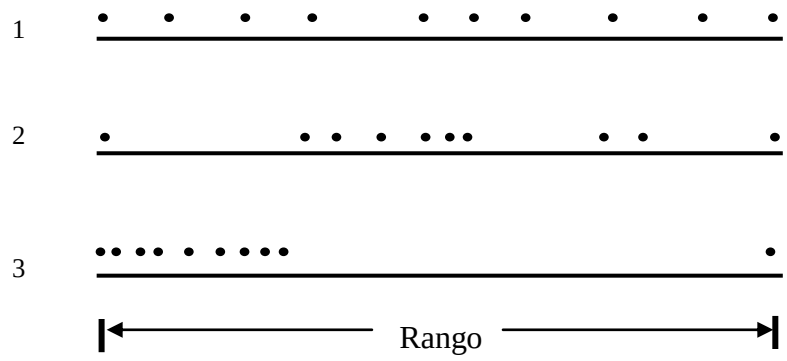
$$R = LS_K - LI_1$$

En los ejemplos anteriores para los dos grupos de cerdos se tiene que el recorrido para el grupo 1 es R = 178 - 172 = 6 Kg., y para el grupo 2 es R = 185 - 165 = 20Kg.

La ventaja de utilizar el rango como medida de dispersión, es la sencillez de su cálculo, aun cuando se trate de un conjunto bastante grande de datos. Además, el significado de esta medida es fácil de comprender.

La principal limitación del rango es que considera solamente los valores extremos de los datos, y no proporciona información respecto a los demás valores.

En el siguiente ejemplo se presentan tres conjuntos de datos bastante diferentes, que poseen el mismo rango.



Nótese como en el primer grupo de datos, los valores se distribuyen en forma uniforme, y esta medida cumple con su objetivo. En el segundo conjunto, los datos se encuentran más agrupados y acá el rango mide de una "forma cruda" la dispersión. Sin embargo, la tercera colección demuestra cómo se puede influir fácilmente en el rango mediante unos cuantos *valores extremos*

(o valores atípicos), y presentar información bastante engañosa respecto a la dispersión de una colección de datos. Debido a estos problemas, el rango tiene una limitada utilidad ya que no resulta una medida de dispersión confiable.

2. **Varianza**

Supóngase que $x_1, x_2, \ldots, x_n$ son las observaciones una muestra aleatoria, cuya media es $\bar{x}$. Dado que se está interesado en analizar la dispersión de estos valores, será natural fijarse en sus distancias con respecto a la media, esto es, en las diferencias:

$$x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x}$$

Puesto que algunos valores de la muestra son mayores que la media y otros son menores, algunas de estas diferencias serán positivas y otras negativas. Es más, las diferencias están "equilibradas", en el sentido de que su suma es 0 (por propiedad i. de la media aritmética, ver Pág.49)

Sin embargo, para analizar la dispersión de los datos, no interesa el signo de las diferencias, Así se tratará una diferencia negativa exactamente igual que una diferencia positiva de la misma cantidad. Por ejemplo, un salario que esté 100.000 bolívares por debajo de la media deberá ser tratado exactamente igual que uno que esté 100.000 bolívares por encima de la media. Una forma de conseguir este objetivo consiste en fijarse, no en las diferencias, sino en sus cuadrados:

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \cdots, (x_n - \bar{x})^2$$

El promedio de los cuadrados de las diferencias proporciona una medida de la dispersión que se conoce con el nombre de *varianza*.

Este es un indicador que mide la dispersión de los datos con respecto a su media aritmética y se denota por S_*^2.

Dada una colección de datos $x_1, x_2, \ldots, x_n$, cuya media aritmética es $\bar{x}$, se define la *varianza* de esos datos como el promedio de las diferencias elevadas al cuadrado de cada uno de esos valores con respecto a su media. Es decir:

$$S_*^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n}$$

Nota

De la fórmula anterior se deduce que:

i. Mientras más alejados estén los valores de su media mayor será el valor de la varianza y mientras más concentrados se encuentren alrededor de su media, menor será el valor de la varianza.
ii. La varianza nunca es negativa, ya que se está sumando cantidades elevadas al cuadrado.
iii. El valor mínimo que puede tomar es cero, el cual se logra cuando todos los valores son iguales entre sí, es decir, que no existe variabilidad entre ellos.

iv. Si se desarrolla la fórmula anterior, se obtiene otra expresión equivalente de la varianza, más cómoda de calcular y además reduce los errores de redondeo:

$$S_*^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\overline{x}\right)^2$$

Ejemplo:
Los datos de los pesos de los cerdos son: 172, 177, 178, 173, 177, 174, 176, 173 y la media es 175 Kg. Calcular la varianza.

$$S_*^2 = \frac{(172-175)^2 + (177-175)^2 + (178-175)^2 + (173-175)^2 + (177-175)^2 + (174-175)^2 + (176-175)^2 + (173-175)^2}{8}$$

$$S_*^2 = 4{,}5 \text{ Kg}^2$$

Por la otra fórmula:

$$S_*^2 = \frac{172^2 + 177^2 + 178^2 + 173^2 + 177^2 + 174^2 + 176^2 + 173^2}{8} - 175^2$$

$$S_*^2 = 4{,}5 \text{ Kg}^2$$

Nótese que la varianza viene expresada en las unidades de los datos pero elevadas al cuadrado. Por esta razón, la varianza resulta difícil de interpretar. Para solucionar esta situación se define la *desviación estándar*.

3. **Desviación Estándar (Desviación Típica)**:
La *desviación estándar* o *desviación típica* de una colección de datos, denotada por S_*, se define como:

$$S_* = +\sqrt{S_*^2}$$

La cual viene dada en las mismas unidades de los datos.

Ejemplo:
Tomando el ejemplo anterior se tiene que:

$$S_* = \sqrt{4{,}5 \text{ Kg}^2}$$
$$= 2{,}12 \text{ Kg}$$

Para distribuciones de frecuencias
Para el caso de datos agrupados en distribuciones de frecuencias, las expresiones para la varianza y la desviación estándar son:

$$S_*^2 = \frac{\sum_{i=1}^{k} \left(m_i - \overline{x}\right)^2 f_i}{n} = \frac{\sum_{i=1}^{k} m_i^2 f_i}{n} - \left(\overline{x}\right)^2$$

$$\text{y} \quad S_* = \sqrt{S_*^2} \qquad \text{donde,} \quad \bar{x} = \frac{\sum_{i=1}^{n} f_i m_i}{n}$$

Ejercicio:
Calcular la varianza y la desviación estándar para las distribuciones de frecuencias de las variables peso, número de hermanos, estatura, ingreso mensual del hogar, número de visitas a la discoteca y número de visitas al cine.

Propiedades de la varianza y la desviación estándar

Sea $x_1, x_2, \ldots, x_n$ una colección de datos, cuya media, varianza y desviación estándar son $\bar{x}$, S_*^2 y S_* respectivamente.

i. S_*^2 y S_* son no negativas, es decir $S_*^2 \geq 0$ y $S_* \geq 0$ para cualquier conjunto de datos.

ii. Si cada uno de los datos $x_1, x_2, \ldots, x_n$ es igual a un mismo valor fijo o constante c, entonces la varianza S_*^2 y la desviación estándar S_* son iguales a cero.

En el diagrama de puntos que se muestra en la página 50 se puede observar esta situación.

iii. Si a cada uno de los datos originales se le suma un mismo número real c, positivo o negativo, entonces la nueva colección de datos que se origina $x_1+c, x_2+c, \ldots, x_n+c$ tiene la misma S_*^2 y S_* que los datos originales.

Obsérvese el gráfico correspondiente a la propiedad iv de la media aritmética en la página 50. Nótese como la dispersión se mantiene invariante al sumar o restar una constante.

iv. Si cada uno de los datos se multiplica por un mismo número real cualquiera c, la varianza y la desviación estándar de los "nuevos datos" $x_1*c, x_2*c, \ldots, x_n*c$ vienen dadas por $c^2 S_*^2$ y $|c| S_*$ respectivamente.

Esto se ilustra en los gráficos que corresponden a la propiedad v de la media aritmética en la página 51. Obsérvese como se produce una alteración en la dispersión de los nuevos datos, ya sea disminuyendo o aumentando dependiendo del valor de c.

Ejercicio:

En un estudio realizado en un hospital se determinó que se gastaba en medicinas un promedio de Bs. 80.000 semanalmente por paciente con una desviación estándar de Bs. 15.000.

a. Si se produce un aumento del 100% en el precio de las medicinas, cuanto será el gasto promedio por paciente y cuanto será la varianza.
b. Cuanto será el gasto promedio por paciente y cuanto será la desviación estándar si el aumento es del 20%.

Varianza muestral y varianza poblacional

Al definir la varianza S_*^2 se ha estado suponiendo que la colección de datos $x_1, x_2, \ldots, x_n$ constituye una muestra de tamaño *n* de una población y que $\bar{x}$ es la media de esa muestra. La *varianza poblacional*, denotada por σ^2, de una población de *N* elementos cuya media poblacional es μ, se define por:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$$

y la *desviación estándar poblacional* es:

$$\sigma = +\sqrt{\sigma^2}$$

Nota:

La varianza muestral también puede definirse como:

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

Se utiliza con la finalidad de, además de tener fines descriptivos, realizar inferencias sobre una población usando S^2 y no S_*^2 por cuanto se demuestra que S^2 es un mejor estimador de la varianza poblacional σ^2 que S_*^2 como se verá en el tema de *estimación*.

<u>Coeficiente de variación</u>

La *medida de dispersión relativa* más conocida es el *coeficiente de variación*. En algunas ocasiones es de interés comparar la dispersión de dos colecciones de datos. Si los datos están medidos en las mismas unidades y las respectivas medias aritméticas son iguales o muy parecidas es posible utilizar la desviación estándar. Si esto no se cumple, no se puede utilizar la desviación estándar para comparar las dispersiones de los dos grupos de datos.

Una medida de dispersión que permite la comparación de la dispersión en cualquier situación, que no viene expresada en ninguna unidad es el *coeficiente de variación*.
El *coeficiente de variación* se define como:

$$CV = \frac{S_*}{\bar{x}} * 100\%$$

El *coeficiente de variación* es la proporción o porcentaje de la media que representa la desviación estándar. Obsérvese como la fórmula anterior proviene de una regla de tres simple:

$$\bar{x} \rightarrow 100\%$$
$$S_* \rightarrow ?$$

Si por ejemplo el CV=20%, significa que la desviación estándar representa el 20% <u>del valor de la media aritmética</u>.

Ejercicio:
Supóngase que se desea comparar las dispersiones de los sueldos de los empleados de las empresas "Cervecería El Cóndor" y "Aguardiente Tropical". Los sueldos promedio para estas empresas son Bs. 7000 y Bs. 2500 respectivamente; las desviaciones estándar correspondientes son Bs. 3000 y Bs. 300.

Ejercicio:
En una encuesta sobre bienes raíces en la Urbanización Santa Cecilia de una ciudad, se obtiene entre otras cosas, información sobre el valor actual de la casa y el tamaño del lote de terreno. Se está interesado en determinar si el valor de avalúo tiene mayor variabilidad que el tamaño del lote. De la mencionada encuesta se consigue lo siguiente:

Valor de la casa	**Tamaño del terreno**
$\bar{x}$ = 1.550.000 Bs.	$\bar{x}$ = 650 mts^2
S_* = 500.000 Bs.	S_* = 350 mts^2

Ejercicio:
Compare la dispersión de la distribución de frecuencias del peso de los varones con la dispersión de la distribución del peso de las hembras.

Percentiles, Deciles y Cuartiles
Además de las medidas de tendencia central, dispersión y forma, también existen algunas medidas interesantes de posición que se utilizan al resumir y analizar las características o propiedades de grandes colecciones de datos.

1. **Percentiles**

Los *percentiles* son aquellos valores que dividen a los datos *ordenados* de forma creciente, en cien partes iguales. Existen noventa y nueve percentiles que se denotan por P_1, P_2, ... , P_{99}. Entre dos percentiles consecutivos se encuentra el 1% de los datos. Así, por ejemplo, entre los percentiles P_{10} y P_{20} se encuentran 10% de los datos.

Para denotar un percentil cualquiera usamos P_h, donde h = 1, 2, 3, ... , 99. Así, la definición formal de percentil es la siguiente:

El percentil P_h de una colección de datos que previamente han sido ordenados (de forma creciente), es un valor tal que como máximo el h% de los datos son menores que él, y también como máximo un (100-h)% de los datos son mayores que él.

Como en el caso de la mediana, si dos valores consecutivos del conjunto de datos cumplen con la definición anterior, se conviene en tomar como percentil al promedio de ellos dos.

Ejemplo:
Suponga que los pesos de ocho personas (en Kg) son: 52, 97, 108, 63, 90, 74, 86, 73. Hallar lo percentiles: P_{20}, P_{50} y P_{80}.

En primer lugar se deben ordenar de forma creciente los datos:

52 63 73 74 86 90 97 108

El P_{20} es el valor tal que el 20% de los datos, es decir el 20% de 8 = 1,6 datos como máximo son menores que él, y también como máximo el 80% de 8 = 6,4 datos son mayores que él.
Observe que el valor 63 cumple con estas condiciones. Por tanto, P_{20} = 63 Kg.

Ahora, en el cálculo de P_{50} se observa que existen dos valores 74 y 86, que cumplen con la definición. De esta manera, P_{50} = (74 + 86) / 2 = 80 Kg.

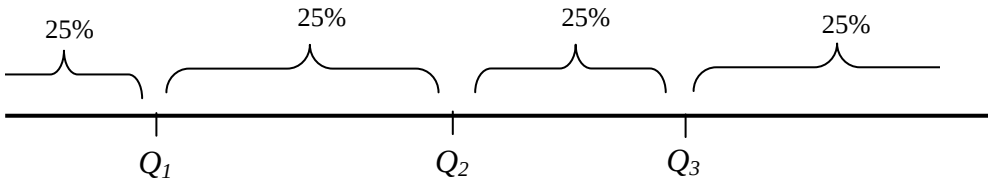
Para estos datos, P_{80} tiene como máximo 6,4 datos por debajo de él y a lo sumo 1,6 datos por encima. El valor 97 satisface esto, así P_{80} = 97 Kg. Nótese que ni el valor 90 ni 108 cumple con las condiciones. Por ejemplo, el valor 90 tiene cinco datos por debajo que cumple con lo que se exige pero por encima tiene a dos datos (el 25% de los datos), lo que no satisface los requerimientos para ser percentil 80.

2. **Deciles**

Los *Deciles* son los valores que dividen a los datos ordenados (de forma creciente) en diez partes iguales. Existen nueve deciles que se denotarán por D_1, D_2, ... , D_9. Entre dos deciles consecutivos se encuentra un 10% de los datos.

3. **Cuartiles**

Los *cuartiles* son los valores que dividen a una colección de datos que previamente han sido ordenados en forma creciente, en cuatro partes iguales. De esta manera, existen tres cuartiles que se denotan Q_1, Q_2 y Q_3. Nótese que entre dos cuartiles consecutivos se encuentra un 25% de los datos. Además, por debajo de Q_1, se encuentra un 25% de los datos y por encima un 75%; mientras por debajo del cuartil tres, se encuentra un 75% de los datos y por encima de él existe un 25% de los datos.



Nótese que el segundo cuartil, Q_2, es igual a la mediana. Además, puede dejarse ver las siguientes relaciones entre los cuartiles deciles y percentiles:

$$Q_1 = P_{25}$$
$$Q_2 = D_5 = P_{50} = Md$$
$$Q_3 = P_{75}$$
$$D_1 = P_{10}$$
$$D_2 = P_{20}$$
$$\vdots$$
$$D_9 = P_{90}$$

Nota:
A los cuartiles, deciles y percentiles en general se les denominan *cuantiles*

Ejercicio:
Para los datos no agrupados de estatura, calcular e interpretar: los cuartiles, el decil tres y el percentil diez.

Cálculo de Percentiles en distribuciones de frecuencias

1. En el caso de distribuciones de frecuencias cuyas clases son intervalos, los percentiles, de la misma manera como se hizo con la mediana, se pueden calcular mediante un método de interpolación tanto de forma algebraica como gráfica.

   Algebraicamente, para el cálculo del percentil *h*-ésimo, P_h, se sigue el siguiente procedimiento:

   a) Se ubica la clase del percentil *h*, que es aquella que contiene la frecuencia acumulada $n*\left(\frac{h}{100}\right)$.

   b) Una vez ubicada la clase del percentil *h*, mediante un proceso de interpolación se puede obtener la siguiente fórmula para el cálculo de los percentiles:

$$P_h = LI_p + \left[\frac{n*\left(\frac{h}{100}\right) - F_{ap}}{f_p}\right] * C_p$$

   en donde,

   LI_p: Límite inferior de la clase del percentil h.

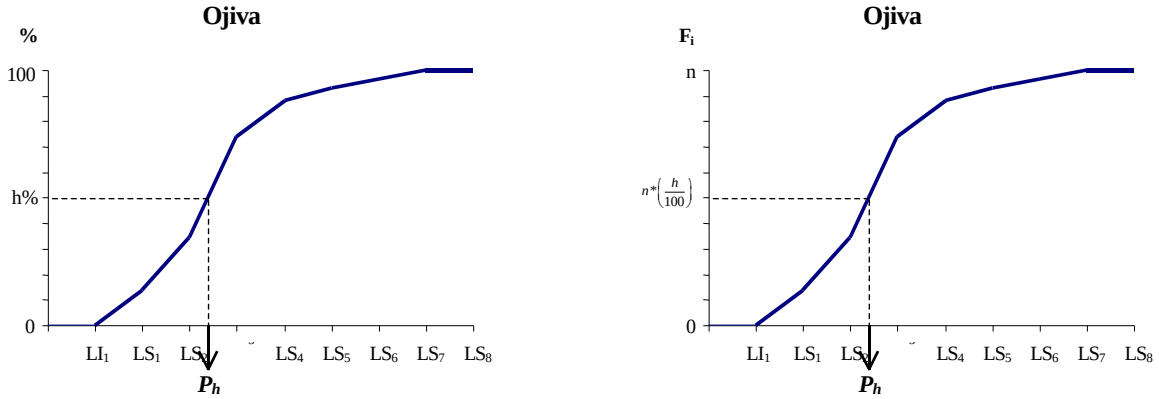
   n: No. total de observaciones o datos

   F_{ap}: Frecuencia acumulada anterior a la clase del percentil h.

   f_p: Frecuencia absoluta de la clase del percentil h.

   C_p: Amplitud de la clase del percentil h.

   Como se vio antes con la mediana, se pueden obtener gráficamente los percentiles utilizando la ojiva:



Nótese que por el mismo método de interpolación gráfico de la ojiva para distribuciones de frecuencias con intervalos; si se conoce algún valor de los datos, digamos P_h, entonces puede ser encontrada la proporción (o porcentaje) de datos, *h*, que son menores (o puede ser, mayores o iguales) que el valor P_h. Simplemente, despejando *h* de la fórmula para calcular percentiles en distribuciones de frecuencias. Así de ese despeje queda que:

$$h = \left[\left(P_h - LI_p \right) \frac{f_p}{C_p} + F_{ap} \right] \frac{100}{n}$$

Ejercicio:

i. Obtenga la fórmula anterior:
   a. Despejándola de la fórmula para el cálculo de los percentiles.
   b. Deduciéndola mediante el método gráfico de interpolación, con la ojiva.

ii. Del ejercicio para la ojiva de la página 44, obtenga las respuestas de los ítems 1 a 4, usando la última fórmula.

2. Si las clases de la distribución de frecuencias son *valores individuales* de la variable en estudio, se procede similarmente a como se hizo con la mediana. En este caso, no hace falta hacer alguna interpolación. Se puede encontrar cualquier percentil mediante la definición.

Ejercicio:
Calcular los cuartiles Q_1 y Q_3, el decíl D_9 y el percentil P_{90} en las distribuciones de frecuencias de las variables peso y estatura usando el método algebraico y el método gráfico.

Ejercicio:
Calcular los percentiles P_{15} y P_{80} para las distribuciones de frecuencias correspondientes a las variables número de hermanos y número de visitas al cine usando el método algebraico y el método gráfico.

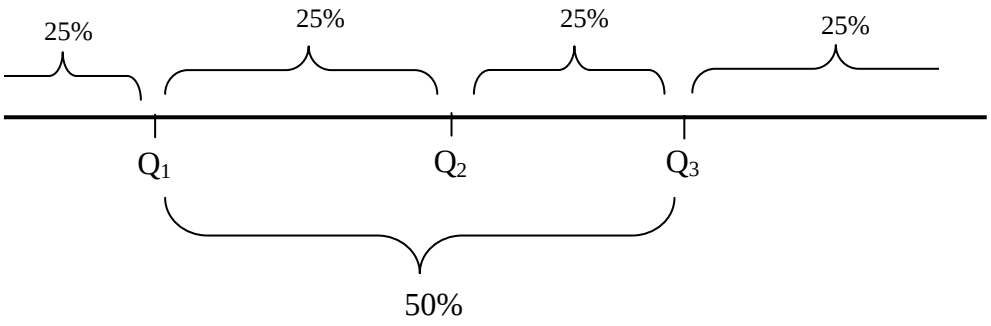
Los percentiles también son utilizados como indicadores de la dispersión de los datos. Con ellos se construyen algunas medidas de dispersión. Veamos algunas de ellas:

Recorrido Intercuartil

El *recorrido intercuartil*, viene dado por:

$$RQ = Q_3 - Q_1$$

Esta medida refleja la dispersión de la parte central de la distribución ya que toma en cuenta al 50% de los datos del centro de la distribución:



Desviación Cuartil ó Recorrido Semi-Intercuartil

La *desviación cuartil* se obtiene mediante la siguiente expresión:

$$Q = \frac{Q_3 - Q_1}{2}$$

Si se calcula:

$$Md \pm Q$$

se obtiene un intervalo que contiene aproximadamente el 50% de los datos.

Fácilmente puede notarse que las dos medidas anteriores no toman en cuenta a todos los datos, lo cual puede representar una seria desventaja ya que es posible que por debajo de Q_1 o por encima de Q_3, los datos se encuentren muy concentrados o muy dispersos y el efecto sobre RQ y Q será el mismo. Aunque por otro lado, y por la misma razón, el recorrido intercuartil y la desviación cuartil no son afectados por valores atípicos.

Recorrido Percentil

Es una medida basada en la misma idea que el RQ, la cual viene dada por:

$$RP = P_{90} - P_{10}$$

Este indicador refleja el 80% de los datos ubicados en la parte central de la distribución

Ejercicio:
Para las distribuciones de frecuencia correspondientes a las variables peso e ingreso hallar:

a) *RQ*
b) *RP*
c) El intervalo que contiene aproximadamente el 50% de los datos de la parte central de la distribución.

Medidas de Forma

En una sección anterior se examinó la forma en que se distribuyen los datos analizando el respectivo gráfico. Se observó la simetría (o asimetría) que presentan los datos y también se podía percibir el grado de apuntamiento (o achatamiento) del gráfico que representa la distribución de los datos.

Existen indicadores que cuantifican la asimetría y el apuntamiento de una distribución, los cuales son de utilidad cuando no se dispone del gráfico o para confirmar las conclusiones obtenidas gráficamente.

Tanto las medidas de asimetría como las de apuntamiento son indicadores relativos ya que no vienen expresados en alguna unidad de medida.

1. **Medidas de Asimetría**

Los resultados que se discutirán se refieren a distribuciones unimodales:

a. ***Coeficiente de Asimetría de Pearson***

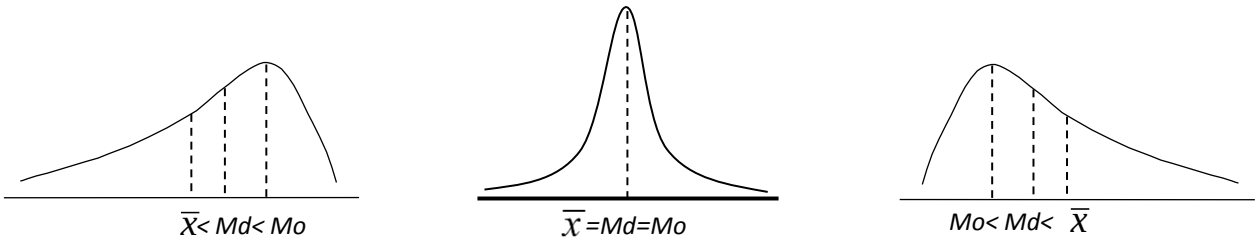
Este indicador se basa en la relación existente entre la media y la mediana:

$$ASP = \frac{3(\bar{x} - Md)}{S_*}$$

Obsérvese que si la distribución es:

- Simétrica => ASP = 0, ya que en este caso $\bar{x} = Md$
- Asimétrica por la derecha => ASP > 0, debido a que $\bar{x} > Md$
- Asimétrica por la izquierda => ASP < 0, porque $\bar{x} < Md$

El coeficiente de asimetría de Pearson toma valores en el intervalo (-3, 3)



b. ***Coeficiente de Asimetría de Fisher***

Se denota por γ_1 y viene dado por la siguiente fórmula:

- Datos no agrupados:

$$\gamma_1 = \frac{\left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n} \right)}{S_*^3}$$

- Datos agrupados:

$$\gamma_1 = \frac{\left(\frac{\sum_{i=1}^{k} (m_i - \bar{x})^3 f_i}{n} \right)}{S_*^3}$$

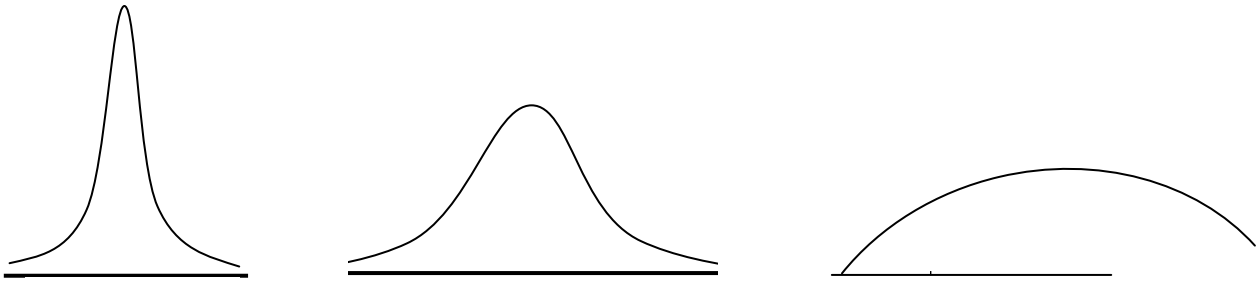
El coeficiente γ_1 está basado en la media aritmética e indica de que lado las diferencias respecto de éstas son mayores.

Su interpretación es similar a la del coeficiente de Asimetría de Pearson

Ejercicio:
Calcule e interprete los coeficientes de asimetría ASP y γ_1 para las distribuciones de frecuencias correspondientes a las variables peso, estatura, ingreso y número de hermanos.

2. **Medidas de Apuntamiento o Curtosis**

Estas medidas indican el grado de apuntamiento o achatamiento del gráfico. La medición del apuntamiento de un gráfico se hace tomando como referencia la curva normal (curva de campana o curva de Gauss). Por tanto, el gráfico al que se le desee medir el apuntamiento, debe ser al menos aproximadamente simétrico. A la curva normal se le llama mesocúrtica, si es más puntiaguda se le llama leptocúrtica y si es más achatada platicúrtica.



Leptocúrtica **Mesocúrtica** **Platicúrtica**

Nótese que los indicadores de curtosis, miden el nivel de concentración de datos en la región central.

Coeficiente de Pearson

El coeficiente β_2 de Pearson es el más utilizado de las medidas de apuntamiento y viene dado por:

- Datos no agrupados:

$$\beta_2 = \frac{\left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n} \right)}{S_*^4}$$

- Datos agrupados:

$$\beta_2 = \frac{\left(\frac{\sum_{i=1}^{k} (m_i - \bar{x})^4 f_i}{n} \right)}{S_*^4}$$

- Si la curva es normal (mesocúrtica) , $\beta_2 = 3$
- Si la curva es leptocúrtica , $\beta_2 > 3$
- Si la curva es platicúrtica , $\beta_2 < 3$

Ejercicio:

Calcule e interprete el coeficiente β_2 de Pearson para las distribuciones de frecuencia correspondientes a las variables peso, estatura y número de hermanos.

DIAGRAMAS DE CAJA

El diagrama de tallo y hoja y el histograma proporcionan una impresión visual general del conjunto de datos, mientras que las cantidades numéricas tales como $\bar{X}$ o S brindan información sobre una sola característica de los datos. El ***diagrama de caja*** es una presentación visual que describe al mismo tiempo varias características importantes de un conjunto de datos, tales como el centro, la dispersión, la simetría o asimetría y la identificación de observaciones atípicas.

El diagrama de caja representa los tres cuartiles, y los valores mínimo y máximo de los datos sobre un rectángulo (caja), alineado horizontal o verticalmente.

Construcción:

1. El rectángulo delimita el rango intercuartílico con la arista izquierda (o inferior) ubicada en el primer cuartil Q_1, y la arista derecha (o superior) en el tercer cuartil Q_3.
2. Se dibuja una línea a través del rectángulo en la posición que corresponde al segundo cuartil (que es igual al percentil 50 o a la mediana), Q_2 = Md.
3. De cualquiera de las aristas del rectángulo se extiende una línea, o *bigote*, que va hacia los valores extremos (valor mínimo y valor máximo). Estas son observaciones que se encuentran entre cero y 1.5 veces el rango intercuartílico a partir de las aristas del rectángulo.
4. Las observaciones que están entre 1.5 y 3 veces el rango intercuartílico a partir de las aristas del rectángulo reciben el nombre de *valores atípicos*. Las observaciones que están más allá de tres veces el rango intercuartílico a partir de las aristas del rectángulo se conocen como *valores atípicos extremos*. En ocasiones se emplean diferentes símbolos (como círculos vacíos o llenos), para identificar los dos tipos de valores atípicos.

A veces, los diagramas de caja reciben el nombre de *diagramas de caja y bigotes.* Nótese que el rectángulo o caja representa el 50% de los datos que particularmente están ubicados en la zona central de la distribución. La caja representa el cuerpo de la distribución y los bigotes sus colas.

La Figura 6 presenta esquemáticamente un diagrama de caja indicando sus partes. Del diagrama se interpreta que la distribución de los datos es asimétrica por la derecha, ya que la longitud de los rectángulos por debajo y por encima de la mediana así como los bigotes indican que los datos están más agrupados en sus valores inferiores que en los superiores y además se observa que $\bar{X} > Md$. También destaca la existencia de dos valores atípicos en el extremo superior de los datos.

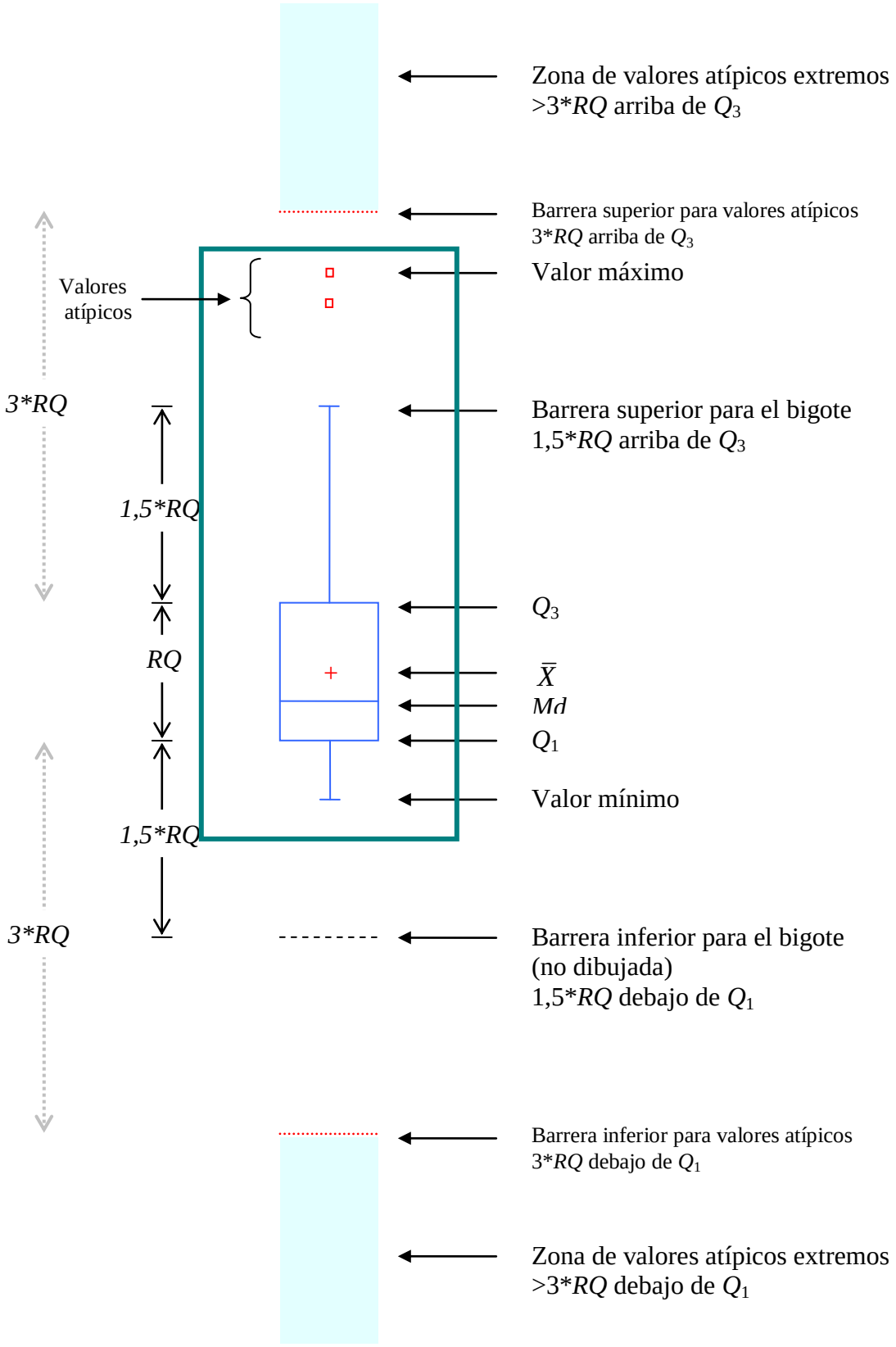


Figura 6. Partes de un diagrama de Caja

En la Figura 7, se muestra el diagrama de caja para la variable *peso* de los estudiantes de Métodos Estadísticos I, sección 02 cursantes del semestre A-2012. Analizando este diagrama se observa que la distribución de los pesos es asimétrica por la derecha, no existen valores atípicos y que por debajo del primer cuartil se encuentra aproximadamente la misma cantidad de datos que por arriba del tercer cuartil. Asimismo, se nota que la mitad de los pesos correspondientes a la parte central de su distribución, se encuentran entre un valor cercano a los 60 kilos y un valor cercano a los 80 kg. También se puede observar que el rango de los pesos varía entre un valor mínimo cercano a los 40 kg y un valor máximo cercano a los 100 kg. Este diagrama de caja fue generado mediante el uso del software estadístico R.

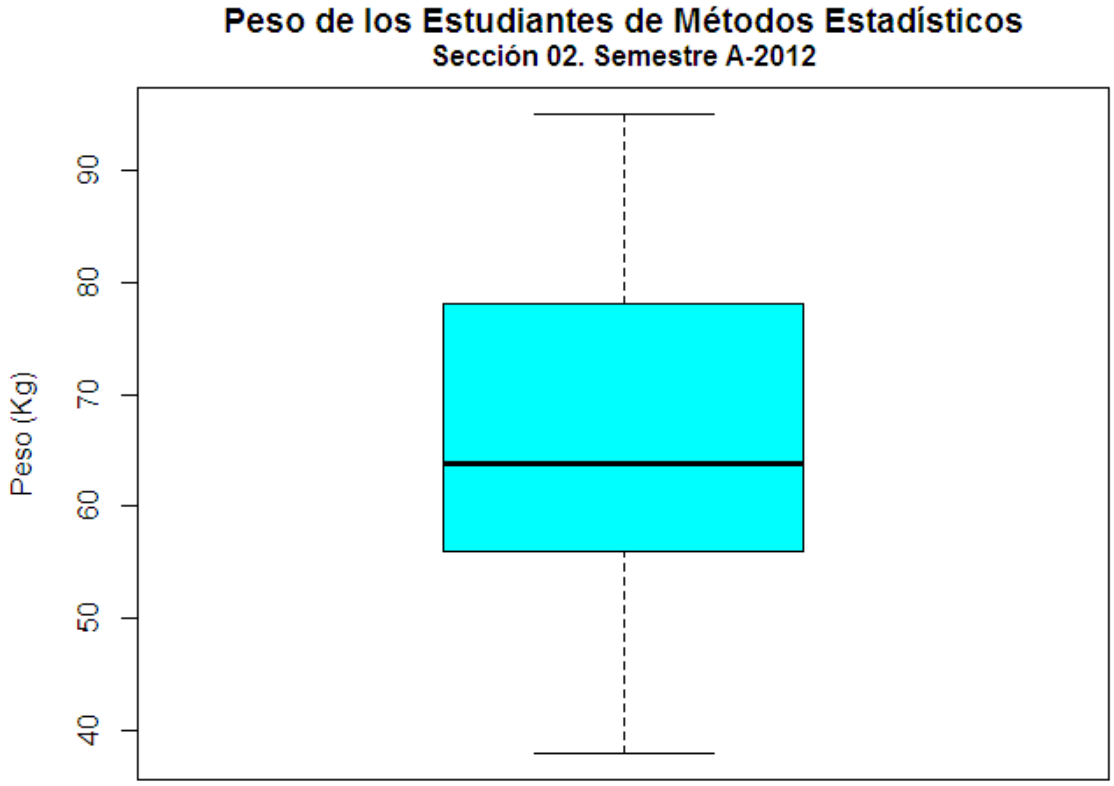


Figura 7. Diagrama de caja (vertical) para los datos de peso de los estudiante de Métodos Estadísticos I Sem-A2012

Nótese en la Figura 8 el mismo diagrama de caja para los datos de pesos, pero ahora dibujado de forma horizontal. Los análisis obtenidos con el diagrama de caja vertical son los mismos que se obtendrían al analizar el diagrama orientado horizontalmente.

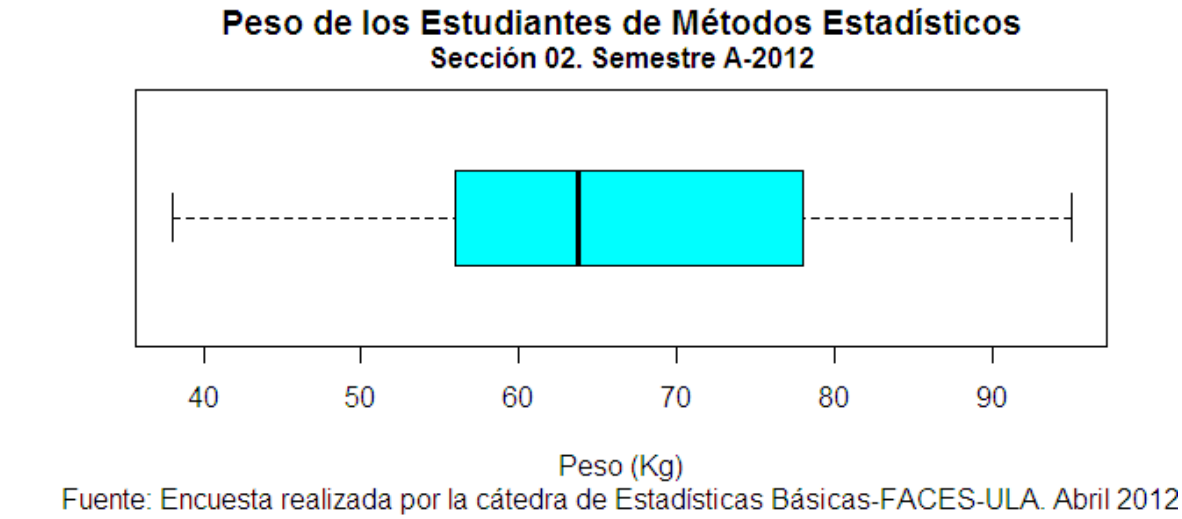


Figura 8. Diagrama de caja (horizontal) para los datos de peso de los estudiante de Métodos Estadísticos I Sem-A2012

Los diagramas de caja son muy útiles al hacer comparaciones gráficas entre conjuntos de datos, ya que tienen un gran impacto visual y son fáciles de comprender. Por ejemplo, la Figura 9 presenta los diagramas de caja comparativos para la variable *peso* de los estudiantes de Métodos Estadísticos I clasificados por *sexo*. El examen de este diagrama revela que el peso de los varones es mayor que el de las hembras. También se observa que la variabilidad de los pesos de las hembras es mayor a la de los varones. Sin embargo, la variabilidad en la parte central de la distribución de los pesos tanto de las féminas como de los masculinos es muy similar. Se nota la existencia de un valor atípico en la distribución de las mujeres, que es un peso muy alto (el valor máximo de todos los pesos) en comparación a los pesos del resto de las muchachas. La distribución del peso de los varones es asimétrico por la izquierda mientras que las hembras presentan una distribución asimétrica por la derecha influenciada por el valor atípico.

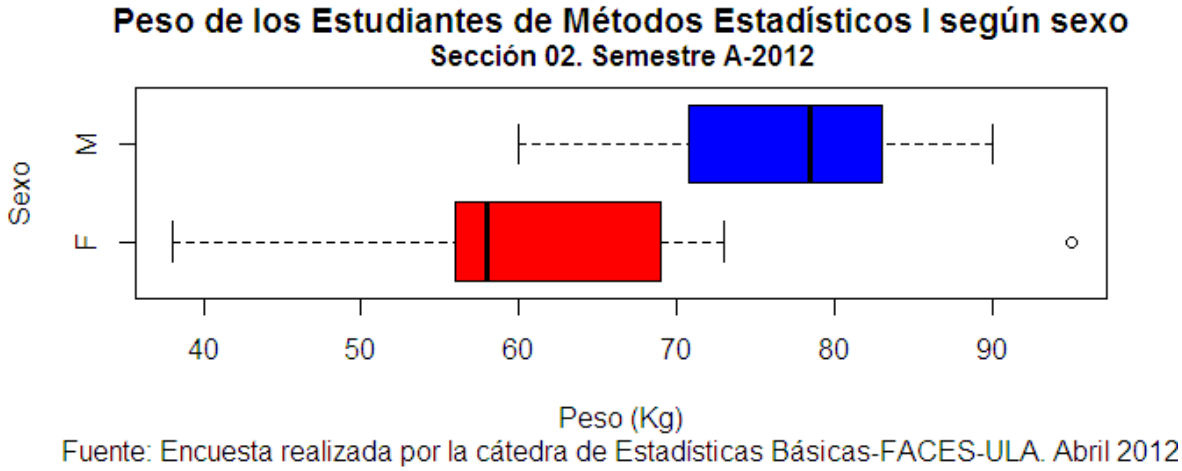


Figura 9 Diagramas de Caja comparativos para la variable Índice Académico clasificado por Sexo

Los datos de la Figura 10, muestran diferentes tipos de distribuciones. Se colocan de manera comparativa los diagramas de caja con los histogramas del mismo conjunto de datos.

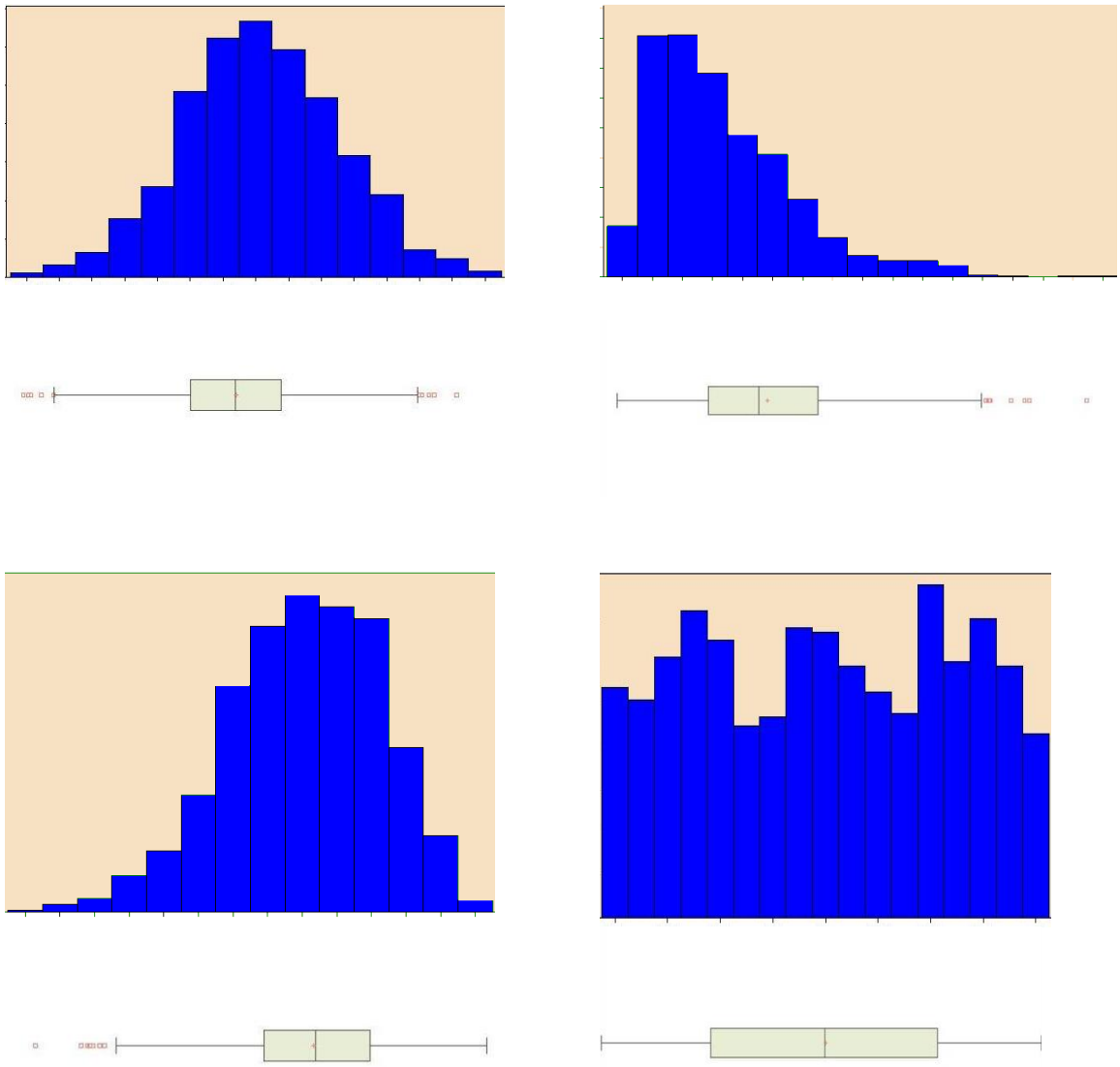


Figura 10. Histogramas y Diagramas de Caja para 1000 observaciones de cuatro distribuciones