TEMA 1
ESTADÍSTICA DESCRIPTIVA

1. Introducción

En sus orígenes la Estadística se basaba en el simple conteo de personas y sus bienes. Desde la época más antigua existen todo tipo de censos.
La Estadística se utiliza hoy en día como tecnología al servicio de las ciencias donde la variabilidad y la incertidumbre forman parte de su naturaleza.

Estadística

La ciencia se desarrolla observando hechos, formulando leyes que los explican y realizando experimentos para validar o rechazar dichas leyes. Los modelos que crea son de tipo determinista o aleatorio (estocástico).
Estadística es la ciencia que sistematiza, recolecta y ordena los datos referentes a fenómenos que presentan variabilidad e incertidumbre, para su estudio metódico [DESCRIPTIVA].
Con objeto de *deducir las leyes* que rigen esos fenómenos [PROBABILIDAD], y poder hacer *previsiones* sobre los mismos, *tomar decisiones* u obtener *conclusiones* [INFERENCIA].

- ***Población***: Conjunto de individuos, sujetos u observaciones que poseen una característica en común que se desea analizar.

La población puede ser finita, esto es, la cantidad de elementos que la componen se pueden contar (por ejemplo los alumnos de una escuela). O infinita, es decir, cuando la cantidad de elementos que la constituyen es infinito, no se puede contar o lo suficientemente grande como para ser considerada finita (por ejemplo, cantidad de personas que se atienden en los hospitales públicos del país). Al conjunto de cosas o personas que constituyen la población se las denomina elementos.

- ***Muestra***: Es un subconjunto de la población, que comparten una característica en común.

La muestra debe ser representativa de la población, para que la información que nos proporcione permita inferir resultados válidos para la población. Hay diversas maneras

de obtener muestras de la población, éstas se denominan "*técnicas de muestreo"*. El "*tamaño de la muestra*" es la cantidad de elementos que forman la misma y se designa con la letra *n.*

- ***Individuo o unidad experimental:*** Es la unidad mínima que se estudia. Es la persona, objeto, etc. sobre la cual se harán las mediciones dado que posee la característica o cualidad que nos interesa estudiar.
- ***Variable***: Una variable es una característica observable *que varía entre los diferentes individuos* de una población.

La información que disponemos de cada individuo es resumida en **variables.**

Ejemplo 1. En los individuos de la provincia son variables:

- Número de hijos → *0, 1, 2, 3, 4, 5 o más*
- Nivel de estudios → *Primario, Secundario, Universitario*
- Consumo de energía eléctrica mensual → *Más de 0w*
- Equipo de futbol preferido → *River, San Martín, Boca, Racing, etc*

Clasificación de las variables.

TIPO DE VARIABLES		***EJEMPLOS***
Variables Cualitativas Toman valores que no se pueden asociar naturalmente a un número.	**Nominal** Si toma valores que no se pueden ordenar.	▪ *Color de cables* ▪ *Religión* ▪ *Estado civil*
	Ordinal Si toma valores que se pueden ordenar.	▪ *Nivel de Educación* ▪ *Grado académico* ▪ *Intensidad de dolor*
Variables Cuantitativas Toman valores que se pueden asociar naturalmente a un número.	**Discreta** Si toma valores enteros.	▪ *Número de hijos* ▪ *Número de accidentes* ▪ *Cantidad de glóbulos rojos*
	Continua *Si toma valores enteros y decimales.*	▪ *Consumo de energía eléctrica mensual* ▪ *Tiempo de vida útil de un artefacto eléctrico.* ▪ *Altura, Peso*

El análisis exploratorio de los datos de una muestra se realiza para describir toda la información que aporta la misma (o en el caso de poblaciones finitas) y se designa comúnmente con el nombre de *Estadística Descriptiva*. La descripción de los datos se

realiza mediante tablas y gráficos, como así también a través de ciertas medidas que resumen la información.

2. Distribución de Frecuencias y Gráficos

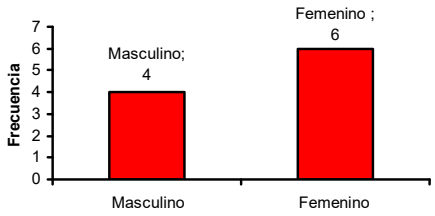
En términos generales, *frecuencia* es la cantidad de veces que se observa el valor de una variable. Hay varias formas de expresar esta frecuencia según se muestra en los ejemplos.

Ejemplo 2. En una muestra de 10 empleados de una fábrica se considera la variable "Género", la información se presenta de la siguiente forma:

Tabla 1. Información de los 10 empleados.

Variable "Género" **M** Masculino - **F** Femenino										
Observación	M	F	F	F	M	F	M	M	F	F

Género	**Frecuencia**
Masculino	4
Femenino	6



Las tablas y el gráfico son una representación equivalente de los datos de la muestra, donde se indica que hay cuatro individuos de sexo masculino y seis de sexo femenino. La tabla se denomina "distribución de frecuencias" porque en ella se indica la cantidad de veces que se repite la variable (frecuencia absoluta), pero también se pueden considerar otro tipo de frecuencias que se definen a continuación.

La presentación de los datos debe ser ordenada mediante una tabla donde se indican distintos tipos de frecuencias.

- Frecuencia absoluta **(f_i):** número de veces que está presente un valor dado de la variable en el conjunto de datos que se desea describir
- Frecuencia relativa: $fr_i = \frac{f_i}{n}$ donde **n** es el tamaño de la muestra.

  Al multiplicar por 100, la frecuencia relativa, representa el porcentaje en que aparece el dato "x_i ".
- Frecuencia acumulada (F_i): cantidad de datos que se poseen hasta alcanzar el valor medido para el cual se calcula la frecuencia acumulada, es decir, $F_i = f_1 + f_2 + \cdots + f_i$

- Frecuencia acumulada relativa: $\boldsymbol{Fr_i = \frac{F_i}{n}}$

Ejemplo 3. Sea la variable X: "Número de hijos por familia" Los resultados obtenidos en una muestra de 8 familias (n = 8) se presentan en la Tabla 2. La misma información se muestra en dos gráficos, considerando la frecuencia absoluta y la acumulada, respectivamente.

Tabla 2. Distribución de frecuencias para "Número de hijos"

Número de hijos $\boldsymbol{X_i}$	$\boldsymbol{f_i}$ **frecuencia absoluta**	***fr*** **frecuencia relativa**	$\boldsymbol{F_i}$ **frecuencia acumulada**	$\boldsymbol{Fr_i}$ **Frecuencia acumulada relativa**
0	1	1/8	1	1/8 = 0.125
1	3	3/8	4	4/8 = 0.50
2	3	3/8	7	7/8 = 0.875
3	1	1/8	8	8/8 = 1

Gráfico 2.1. Gráfico de bastones

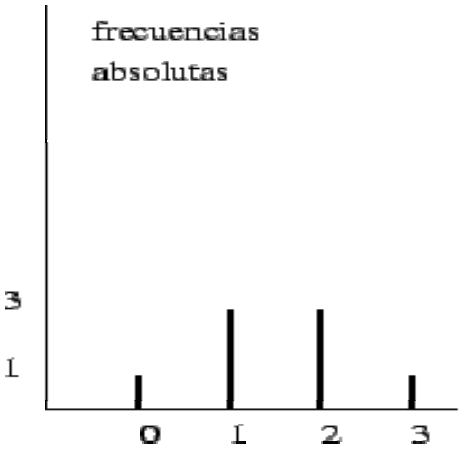
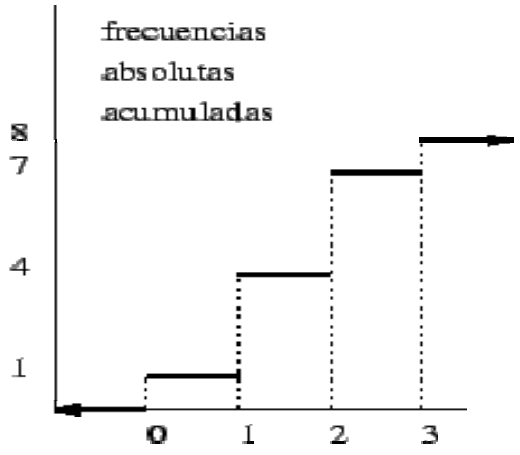


Gráfico 2.2. Gráfico de frecuencias acumuladas



Observación. Las distribuciones de frecuencias, en general, cuentan con los siguientes elementos:

- Título de la tabla, indica la variable a la que se refiere (si se conoce también lugar y tiempo)
- Títulos de las filas y columnas.
- El cuerpo de la tabla, que corresponden a las frecuencias calculadas.

- Total general, representa el tamaño de la muestra.
- Fuente de información, en caso de que la tabla haya sido obtenida de otro servicio u organismo investigador.

Ejemplo 4. Sea la variable X: "Consumo semanal de bebidas gaseosas por familia", en litros. Se considera una muestra de 12 familias.
Los datos en la Tabla 3.

Tabla 3. Distribución de datos agrupados para "Consumo de bebidas"

Consumo de bebidas gaseosas [ )	***m_i (marca de clase)***	***f_i frecuencia absoluta***	***fr frecuencia relativa***	***fr%***	***F_i frecuencia acumulada***
0 – 2	1	2	0.1667	16.67%	2
2 - 4	3	1	0.0833	8.33%	3
4 - 6	5	4	0.333	33.3%	7
6 - 8	7	3	0.25	25%	10
8 – 10	9	2	0.1667	16.67%	12
Total		12	1	100 %	

La diferencia entre las distribuciones de frecuencias presentadas en las tablas 2 y 3 radica en la forma de considerar las observaciones (valores que toma la variable). En la Tabla 3 se ha agrupado en intervalos los valores observados de la variable (datos). Este tipo de distribución se denomina *"distribución de datos agrupados".*

Para muestras grandes es conveniente agrupar los datos en intervalos. Se debe determinar el número de intervalos a formar y la amplitud o ancho de cada uno. El número óptimo de intervalos (o clases), está entre 8 y 20. Esto se debe a que en cualquier agrupamiento en intervalos se produce una pérdida de información ya que ahora se posee rango de valores de los datos y no su valor específico. El agrupamiento de la información permite manipular mejor la tabla, así se gana en practicidad y formando una cantidad adecuada de intervalos, la pérdida de información no será considerable.

Cálculo del *número de intervalos* → ***k***

La fórmula de *Sturges* indica ***k = 1 + 3,3 log n.***

Forma alternativa es considerar la cota $\mathbf{2^K \geq n}$, de donde $k \approx \frac{\log\ n}{\log\ 2}$ **.**

Amplitud de los intervalos → **A** (k es la cantidad de intervalos).

$$A = \frac{x_M - x_m}{k}$$

Donde x_M es el mayor valor de la muestra y x_m el menor.

Una vez obtenida la cantidad de intervalos y la amplitud, se procede a formar la tabla tomando, en general, intervalos semi-abiertos a derecha [a ; b).

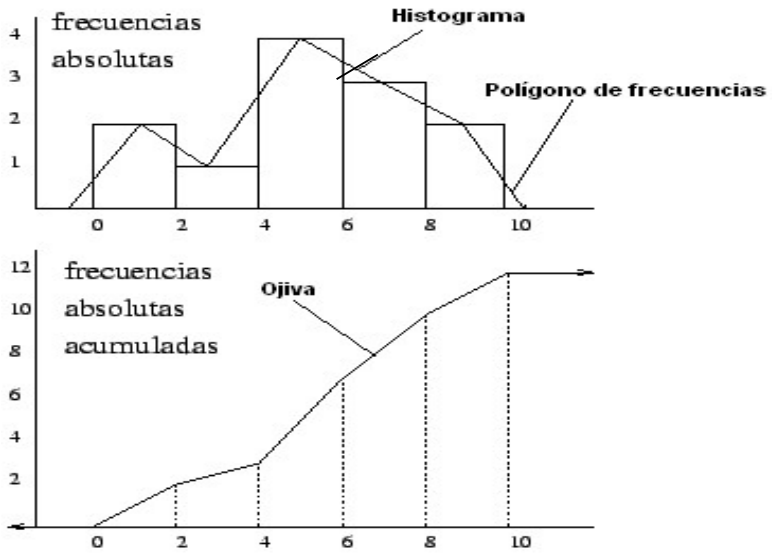
El primer intervalo toma como límite inferior al menor valor registrado y el límite superior será el resultado que se obtenga de sumarle al mínimo valor la amplitud del intervalo. El límite superior del primer intervalo constituye también el límite inferior del segundo. Para obtener el límite superior del segundo intervalo, a su límite inferior se le suma la amplitud. Y así sucesivamente, se sigue hasta formar la cantidad especificada de intervalos.

Posteriormente se obtienen las frecuencias absolutas de cada intervalo contando la cantidad de datos que tienen un valor mayor o igual al límite inferior y menor al límite superior del intervalo.

Las frecuencias que se presentan en la Tabla 3 son análogas a las obtenidas en la Tabla 2, pero se agregan las marcas de clase o puntos medios de cada intervalo. Las marcas de clase representan el promedio de los límites de cada intervalo y se usan en el gráfico del polígono de frecuencias.

El Gráfico 3 muestra una información equivalente a la dada en la Tabla 3.

Grafico 3. Gráficos para Distribuciones de Frecuencias Agrupadas



El primer gráfico corresponde al histograma y polígono de frecuencias, y el segundo es la ojiva o polígono de frecuencias acumuladas.

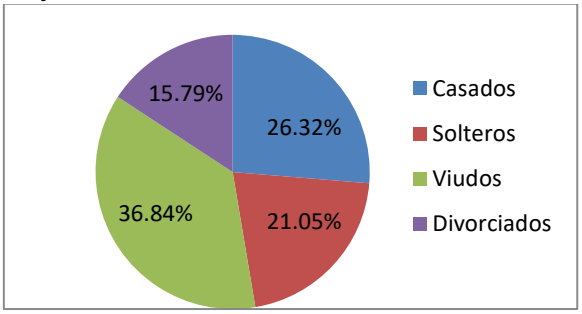
Los intervalos también pueden ser semi-abiertos a la izquierda o ser intervalos con diferentes amplitudes. Esta última opción se toma cuando, si se emplearan intervalos de igual amplitud, quedaran algunos con muchos datos y otros con casi ninguno, lo que provocaría una pérdida de información excesiva. En general esto no sucede y se construyen intervalos de igual amplitud.

Ejemplo 5. Sea la variable X: "Estado civil de los empleados de una empresa". Los datos registrados en 19 personas se presentan en una distribución de frecuencias y en un gráfico.

Tabla 4. Estado Civil de Empleados

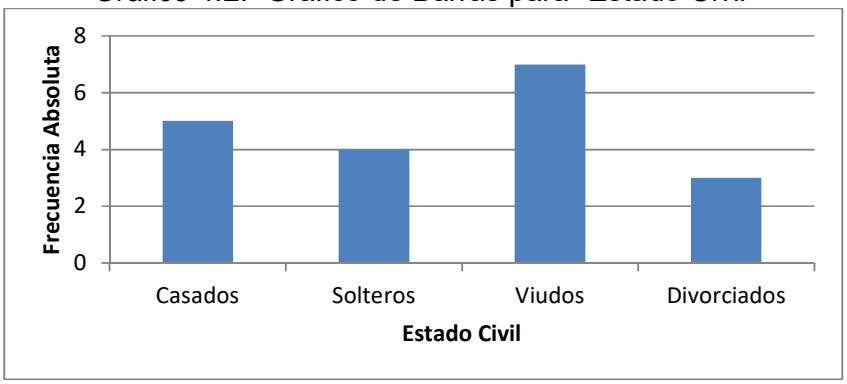
Estado Civil X	f_i	fr_i	$fr_i\%$
Casado	5	5/19	26.31%
Soltero	4	4/19	21.05%
Viudo	7	7/19	36.84%
Divorciado	3	3/19	15.79%

Gráfico 4.1. Sector circular "Estado Civil"



Observación. Este tipo de gráfico se llama sector circular o gráfico de torta. El tamaño de la porción correspondiente a cada categoría se calcula como $\alpha^\circ = \frac{360^\circ \times f_i}{n}$

Gráfico 4.2. Gráfico de Barras para "Estado Civil"



Conclusión. El tipo de variable determina la clase de distribución de frecuencias y el gráfico que es conveniente realizar, de acuerdo a las observaciones.

Tipo de variable		Diagrama de Frecuencias y Gráficos
Cuantitativa	**Discreta**	Diagrama simple. Gráfico de bastones.
	Continua	Diagrama agrupado. Histograma. Polígono de frecuencias. Ojiva.
Cualitativa	**Nominal** **Ordinal**	Diagrama simple. Sector circular. Barras.

3. Medidas de resumen

Las distribuciones de frecuencias y los gráficos dan una idea rápida e intuitiva del comportamiento de los datos. Sin embargo, existe otro tipo de medidas que resumen y describen la información obtenida en la muestra. Se puede considerar que hay dos tipos de medidas, a saber:

- Las que ayudan a encontrar el centro de la distribución de frecuencia de los datos (medidas de tendencia central).
- Las que miden la dispersión de los datos (medidas de dispersión o variabilidad).

Es muy frecuente llamar "*estadísticas" (estadígrafos)* a las medidas descriptivas numéricas calculadas a partir de datos de muestras. En contraste, las medidas descriptivas numéricas de la población se denominan "*parámetros"*. Los valores de los parámetros por lo general se desconocen (ya que es común que no se pueda registrar la población completa).

Medidas de Tendencia Central

3.1. Media Aritmética

Es la medida de posición más frecuentemente usada y también se conoce con el nombre de promedio. La media aritmética de un conjunto *n* de mediciones es el promedio de dichas mediciones.

Para calcular la media aritmética o promedio de un conjunto de observaciones, se suman todos los valores y se divide por el número total de observaciones.

Definición. Dada una muestra aleatoria de *n* observaciones, denotadas por X_1, X_2, ..., X_n, se define la media muestral, cuyo símbolo es $\bar{x}$, como,

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

Ejemplo 6. $X_1 = 10 \quad X_2 = 14 \quad X_3 = 12 \quad X_4 = 11 \quad X_5 = 12 \quad X_6 = 13$

Luego, n = 6 y haciendo un cálculo sencillo (con la calculadora!!) se obtiene,

$$\bar{x} = 12$$

Observación. Si los datos se repiten más de una vez y se presentan en una distribución de frecuencias (Ejemplo 3) el cálculo es,

$$\bar{x} = \frac{1}{n}\sum_i x_i \, f_i$$

Ejemplo 7. En un estudio sobre la edad de los hijos de familias en cierto barrio, la información se presenta agrupada en intervalos como sigue,

Tabla 5. Edades de los hijos de un grupo de familias

Edad [)	Frecuencia	Marca de clase
0 - 5	5	2,5
5 - 10	10	7,5
10 - 15	16	12,5
15 - 20	6	17,5
20 - 25	13	22,5
Total	50	

Si se trata de datos agrupados en intervalos la media se calcula como,

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

Donde *k* es la cantidad de intervalos y $m_i \rightarrow$ ***marca de clase*** en el intervalo ***i*** (punto medio) tal que,

$$m_i = \frac{L_{inf} + L_{sup}}{2}$$

Luego, la edad promedio de los hijos de ese grupo de familias es de 13,7 años, o se puede decir que tienen 14 años de edad, aproximadamente.

Características y propiedades de la media

a. Se usa para datos numéricos.

b. Representa el centro de gravedad o el punto de equilibrio de los datos. Se puede imaginar a los datos como un sistema físico, en el que cada dato tiene una "masa" unitaria y lo ubicamos sobre una barra en la posición correspondiente a su valor.

La media representa la posición en que deberíamos ubicar el punto de apoyo para que el sistema esté en equilibrio.

c. La suma de las distancias de los datos a la media es cero. Esta propiedad está relacionada con el hecho que la media es el centro de gravedad de los datos.

En la tabla siguiente comprobamos esta propiedad para los datos del ejemplo 6.

Tabla 6. Cálculo de desvíos respecto a la media

Variable x_i	10	14	12	11	12	13	Total
$x_i - \overline{x}$	-2	2	0	-1	0	1	0

d. Es muy sensible a la presencia de datos atípicos ("outliers").

Si se modifica un dato en el ejemplo anterior, $X_2 = 14$ por $X_2 = 26$, entonces, ahora

$$\bar{x} = 14$$

Con sólo modificar un dato, la media se desplazó tanto, que ya no se encuentra entre la mayoría de los datos. Se puede decir que en este caso la media no es una buena medida de posición de los datos. En consecuencia, la media es una buena medida del centro de la distribución cuando ésta es simétrica.

Ventajas de la media:

- Emplea toda la información disponible.
- Tiene una fórmula fija de cálculo.

Desventajas de la media:

- No puede calcularse para variables cualitativas.
- Su valor es sensible a datos extremos (muy grandes o chicos).

3.2. Mediana

La mediana es el número que divide la muestra de datos *ordenados,* en dos partes iguales en tamaño (cantidad de datos). Símbolo: **Me**.

Si la variable es cuantitativa y no están agrupadas en intervalos los datos, para el cálculo de la mediana *en primer lugar* se deben ordenar los datos (en general de menor a mayor).

Luego se considera si el número de observaciones (n) es par o impar. Si hay datos repetidos deben ser incluidos en el ordenamiento.

i. Número impar de datos, n impar, la mediana es el dato ubicado en la posición **(n+1) / 2**.

Ejemplo 8. $X_1 = 10$ $X_2 = 14$ $X_3 = 12$ $X_4 = 11$ $X_5 = 13$

Ordenados → 10 11 **12** 13 14 n = 5

Posición 3ª ↑ → **Me = 12**

ii. Número par de datos, n par, la mediana es el valor promedio entre los datos ubicados en la posición **n / 2** y **(n+2) / 2**.

Ejemplo 9. $X_1 = 10$ $X_2 = 14$ $X_3 = 17$ $X_4 = 11$ $X_5 = 12$ $X_6 = 13$

Ordenados → 10 11 **12** **13** 14 17 n = 6

Posición entre 3ª y 4º ↑ ↑ → **Me = 12.5**

Mediana para datos agrupados

En caso de distribución de frecuencias de datos agrupados la mediana está ubicada en el primer intervalo que contiene el 50% de las observaciones acumuladas. Este intervalo se denomina *clase mediana*. Se tomará como valor (aproximado) de la mediana para datos agrupados a la marca de clase (punto medio) de la clase mediana.

Ejemplo 10. Considerando los datos del ejemplo 7, se tiene que:

Tabla 7. Edades de los hijos de un grupo de familias

Edad [)	Frecuencia	Marca de clase	F. Acumulada
0 - 5	5	2,5	5
5 - 10	10	7,5	**15**
10 - 15	**16**	12,5	31
15 - 20	6	17,5	37
20 - 25	13	22,5	50
Total	**50**		

Si hay 50 observaciones la mitad de la fecuencia observada es 25. El primer intervalo que contiene las 25 observaciones acumuladas es [10,15) luego este intervalo es la clase mediana. Por lo tanto, el valor aproximado de la mediana es 12.5.

Es decir que la **edad** de los hijos de *la mitad* (el 50%) de las familias del barrio citado llega ***hasta 12.5*** años (o hasta 13 años), aproximadamente.

Propiedades de la mediana

a. La mediana puede ser usada no sólo para datos numéricos, sino además para datos cualitativos ordinales, ya que para calcularla sólo es necesario establecer un orden en los datos.

b. Si la distribución de los datos es aproximadamente simétrica, la media y la mediana serán aproximadamente iguales.

c. La mediana es una medida de posición robusta. No se ve afectada por la presencia de outliers (datos atípicos), salvo que modifiquemos casi el 50% de los datos menores o mayores de la muestra (la proporción de datos que debemos modificar para modificar la mediana depende del número de datos de la muestra).

Ejemplo 11.

I. 10 11 12 12 13 14 $\rightarrow$ $\bar{x} = 12$ Me = 12

II. 10 11 12 12 13 26 $\rightarrow$ $\bar{x} = 14$ Me = 12

d. La mediana es insensible a la distancia de las observaciones al centro, ya que solamente depende del orden de los datos. Esta característica que la hace robusta, es una desventaja de la mediana.

Ejemplo 12. Los siguientes conjuntos de datos tienen mediana 12.

I.	10	11	12	13	14
II.	10	11	12	13	100
III.	10	11	12	12	12
IV.	10	11	12	100	100

Ventajas de la mediana:

- Su valor no es sensible a datos extremos (muy grandes o chicos).

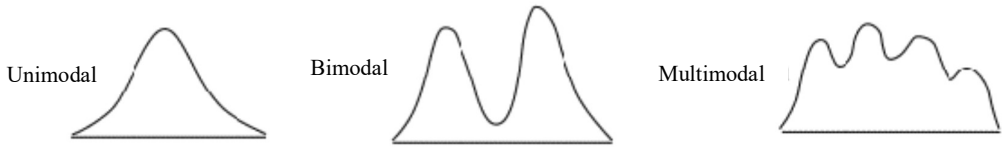
Desventajas de la mediana:

- No puede calcularse para variables cualitativas nominales.
- Su fórmula de cálculo depende del tipo de variable.

3.3. Modo

El modo (o moda) es el o los datos que se presentan con mayor frecuencia absoluta (los que más se repiten). Símbolo: **Mo**

Una muestra puede no tener moda (frecuencias absolutas iguales para todos los valores posibles de la variable), o tener más de un modo.



Para el caso de datos agrupados en intervalos se llama *clase modal* al intervalo (o los) de mayor frecuencia absoluta. Así se considera el valor del modo (aproximado) a la marca de clase de la clase modal.

Ejemplo 13. Considerando los datos del ejemplo 10, la mayor frecuencia absoluta es 16 y la *clase modal* es [10,15). Podría considerarse que el modo es 12.5, pero cuidado pues esto nos podría llevar a una conclusión errónea.

Ventajas de la moda:

- Se puede calcular para todo tipo de variables.

Desventajas de la moda:

- No suele dar valores que sean representativos de la mayoría de los datos.
- Su fórmula de cálculo depende del tipo de variable.

De las tres medidas, la media es la que más se utiliza, aunque es muy sensible a valores muy pequeños o muy grandes que se puedan presentar en la muestra, por lo que puede resultar engañosa en algunos casos. No sucede lo mismo con la mediana, la cual no es sensible a valores extremos, por lo que representa mejor el "centro" de la distribución de los datos para los casos en que la distribución de dichos datos es marcadamente sesgada.
El modo sólo se utiliza si se desea saber cuál es el valor de la variable bajo estudio que más frecuentemente se presenta.

3.4. Otras medidas de posición

Percentiles

El percentil ***p%,*** de un conjunto de datos ordenados, es la observación que deja a lo sumo p% de las observaciones debajo de él y a lo sumo (1 - p)% encima de él. Son 99 valores que dividen en cien porciones iguales al conjunto de datos ordenados.

Ejemplo 14. Se considera la distribución de la producción de tomate en un mes determinado.

Se informa que,

- percentil 10% es 24500 kg

Significa que el 10% de la producción de tomate de ese mes es de 24500 kg o menos (así el 90% producen más que 24500 kg).

- percentil 90% $\rightarrow$ 33700 kg,

Significa que el 90% de la producción de tomate es menor o igual que 33700 kg (el 10% con producción mayor que 33700 kg).

Cuartiles

Son los tres valores que dividen al conjunto de datos ordenados en cuatro porciones iguales, son un caso particular de los percentiles, correspondiendo al 25%, 50% y 75%.

- El primer cuartil $\mathbf{Q_1}$ es el valor de la variable que deja a la izquierda el 25% de la distribución.
- El segundo cuartil $\mathbf{Q_2}$ (la mediana), es el valor de la variable que deja a la izquierda el 50% de la distribución.
- El tercer cuartil $\mathbf{Q_3}$ es el valor de la variable que deja a la izquierda el 75% de la distribución.

Cálculo de los cuartiles de una muestra de n observaciones:

1. Ordenar los datos de menor a mayor.
2. El cuartil inferior $\mathbf{Q_1}$ es el dato que ocupa la posición **(n+1)/4** en la muestra ordenada.
3. El cuartil superior $\mathbf{Q_3}$ es el dato que ocupa la posición **3(n+1)/4** en la muestra ordenada.

Si la posición resulta ser un número decimal, promediamos los datos que se encuentran a izquierda y derecha de la posición obtenida.

Ejemplo 15. Consideremos los siguientes datos ordenados (n = 13).

Q_1 = 140 **Q_2 = Me= 170** **Q_3 = 320**

Datos	104	112	134	146	155	168	170	195	246	302	338	412	678
Posición	1	2	3	4	5	6	7	8	9	10	11	12	13

Posición de Q_1 = (13+1) / 4 = 3.5, entonces $Q_1 = \frac{134+146}{2} = 140$

Posición de la mediana = (13+1) / 2 = 7, luego $Q_2 = Me = 170$

Posición de Q_3 = 3.(13+1) / 4 = 10.5, entonces $Q_3 = \frac{302+3}{2} = 320$

Observación: Los paquetes estadísticos calculan los percentiles usando diferentes métodos y criterios para interpolar. Cuando el conjunto de datos es grande, los distintos métodos tienden a producir el mismo valor para el percentil, pero para conjuntos pequeños pueden diferir ligeramente.

4. Medidas de dispersión

Las medidas de posición dan una idea de dónde se encuentra el centro de la distribución, pero no indican cuán disperso es el conjunto de datos.

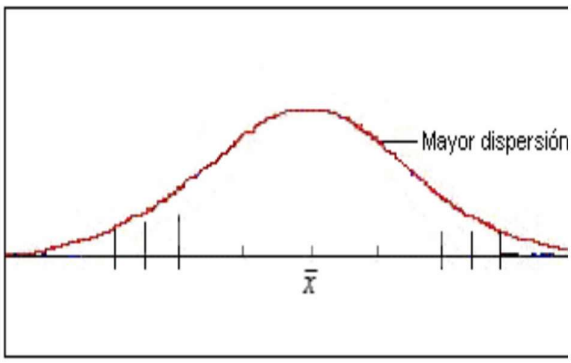
Se consideran los siguientes conjuntos de datos:

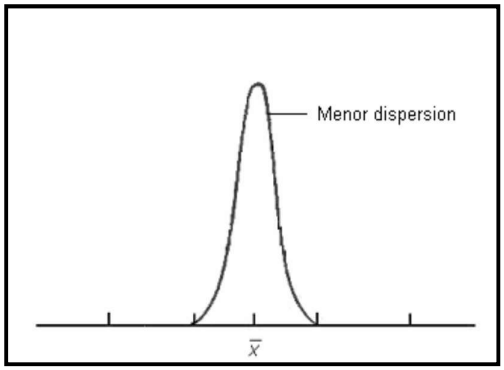
Muestra A:	55	55	55	55	55	55	55
Muestra B:	47	51	53	55	57	59	63
Muestra C:	39	47	53	55	57	63	71

En todos ellos $\bar{x} = Me = 55$, pero las muestras son totalmente diferentes.

Las medidas de dispersión o variabilidad describen cuán cercanos se encuentran los datos entre ellos, o cuán cerca se encuentran de alguna medida de posición.

En caso de datos continuos:





4.1. Rango

El rango de n observaciones X_1, X_2, ..., X_n es la diferencia entre la observación más grande y la más pequeña, $R = X_{máx} - X_{mín}$.

Ejemplo 16.

Muestra A $\rightarrow$ Rango = 55 – 55 = 0

Muestra B $\rightarrow$ Rango = 63 – 47 = 16

Muestra C $\rightarrow$ Rango = 71 – 39 = 32

Características y propiedades del rango

a. Es muy simple de calcular.

b. Es extremadamente sensible a la presencia de datos atípicos. Si hay datos atípicos, éstos estarán en los extremos, que son los datos que se usan para calcular el rango.

c. Ignora la mayoría de los datos.

En consecuencia, reportar el rango o el máximo y el mínimo de un conjunto de datos, no informa demasiado sobre las características de los datos.

4.2. Varianza y Desviación Estándar

Una forma de medir la variabilidad de los datos de una muestra es tomar algún valor central, por ejemplo la media, y calcular el promedio de las distancias a ella. Mientras mayor sea este promedio, más dispersión deberían presentar los datos.

Sin embargo, esta idea no resulta útil, ya que las observaciones que se encuentran a la derecha de la media tendrán distancias (o desviaciones) positivas, en tanto que las observaciones menores que la media tendrán distancias negativas y por ende la suma de las distancias a la media será inevitablemente igual a cero. Un modo de evitar este inconveniente es elevar las distancias al cuadrado y así tener todos sumandos positivos.

Definición. Dada una muestra aleatoria de ***n*** observaciones, denotadas por X_1, X_2, ..., X_n, se define la varianza como,

$$S^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1} = \frac{\sum_{ii=1}^{n}(x_i-\bar{x})^2}{n-1}$$

Si se emplean las unidades de medida de la variable bajo estudio, en los cálculos, la varianza tiene dichas unidades elevadas al cuadrado. Pero para una mejor interpretación de la información se necesita un valor en las mismas unidades de medida que las de la variable.

Así se define la ***desviación estándar*** como la raíz cuadrada (positiva) de la varianza. En símbolos,

$$S = +\sqrt{Varianza}$$

La razón para usar *(n – 1)* y no n en el denominador de la varianza muestral es debido a que el valor de S^2 obtenido en una muestra, se usa para *estimar* la varianza poblacional σ^2. Definida con *(n – 1)* en el denominador, la varianza muestral resulta ser insesgada, es decir, en promedio no subestima ni sobrestima el valor de la varianza poblacional.

Ejemplo 17.

Muestra A: 55 55 55 55 55 55 55 → **$S^2 = 0$** → **$S_A = 0$**

Muestra B: 47 51 53 55 57 59 63 → **$S^2 = 28$** → **$S_B = 5{,}29$**

Muestra C: 39 47 53 55 57 63 71 → **$S^2 = 108$** → **$S_C = 10{,}39$**

Comparando las desviaciones estándar entre las tres muestras, se observa que **$S_A < S_B < S_C$**. En la muestra A, como todas las observaciones toman el mismo valor, la desviación estándar es cero.

Un valor elevado de varianza (o desviación estándar) indica que los datos son muy variables respecto a la media, por lo tanto están muy dispersos entre sí. Mientras que si su valor es pequeño indica que los datos están concentrados alrededor del valor de la media.

Interpretación del valor de la desviación estándar

La desviación estándar **S** es útil para comparar la variabilidad de dos conjuntos de datos en los que la variable ha sido medida en las mismas unidades. Si en una muestra S = 5.3 y en otra S = 10.4, podemos asegurar que los datos de la segunda muestra están más dispersos que los de la primera. Pero ¿cómo interpretamos el valor S = 5.3?

La desviación estándar nos da idea de la distancia promedio de los datos a la media (aunque estrictamente hablando no es el promedio). Pero la interpretación de la desviación estándar requiere algún conocimiento de la distribución de los datos.

Propiedades de la desviación estándar

a. Mide la dispersión alrededor de la media, por lo tanto es natural elegir esta medida de dispersión cuando se usa la media como medida de posición.

b. **S = 0** solamente si todos los datos son iguales, de otro modo **S > 0**.

c. Es una medida de dispersión muy sensible a la presencia de datos atípicos. De hecho, es más sensible que la media ya que las distancias están elevadas al cuadrado.

4.4. Coeficiente de Variación.

El coeficiente de variación sirve para comparar distintas muestras en las que la variable ha sido medida en diferentes unidades y determinar cuál es la más variable, en término de homogeneidad (menor dispersión) y heterogeneidad (mayor dispersión). El coeficiente de variación es una medida adimensional y se define como:

$$CV = \frac{S}{\bar{x}}$$

Coeficiente de variación porcentual, CV% = (S / $\bar{x}$) * 100%

El *coeficiente de variación* permite eliminar la dimensionalidad de las variables y tiene en cuenta la proporción existente entre media y desviación estándar.

Ejemplo 18. El coeficiente de variación para los datos del ejemplo 13 es,

Muestra A → CV% = 0%

Muestra B → CV% = 9,62%

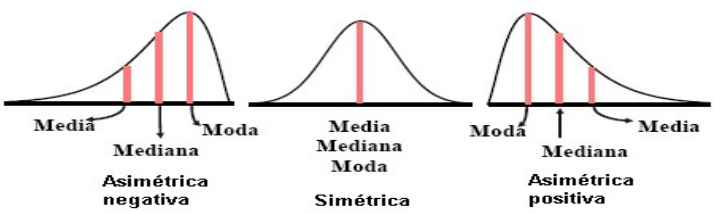
Muestra C → CV% = 18,90%.

Es claro que la mayor variación de los datos es en la muestra C (18,90%).

5. Otras medidas.

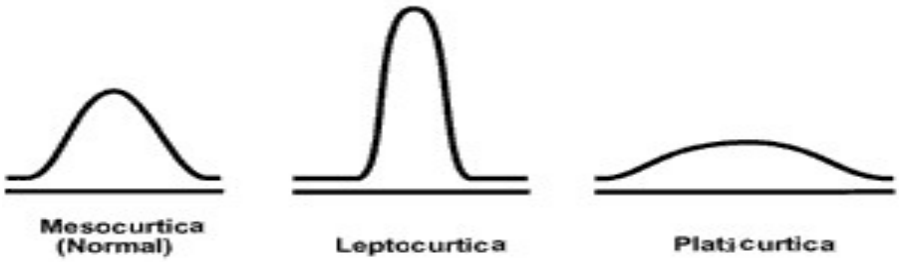
- **Distribución respecto a un eje de simetría**

Las distribuciones pueden tener diferentes formas, y una manera de caracterizar la forma es observar la simetría. Una distribución de frecuencias puede ser simétrica o asimétrica. Para saber si es simétrica tenemos que tomar una referencia, es decir, ver respecto a qué es simétrica. El gráfico de las distribuciones según el tipo de asimetría se muestra en la siguiente figura.



- **Distribución según la forma**

Otros tipos de situaciones de una distribución son los que se presentan en los gráficos siguientes.



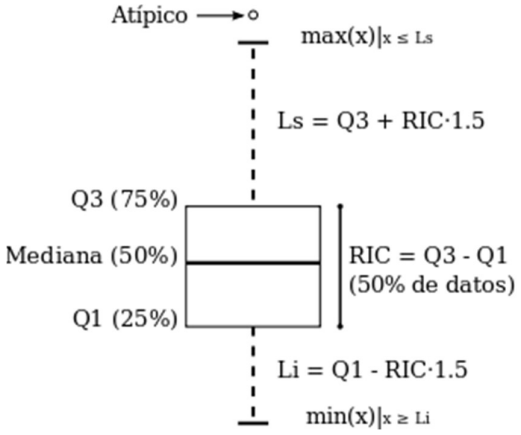
El grado de "aplastamiento" de la distribución se mide con el coeficiente de curtosis.

6. Gráfico de caja (Box-Plot)

Este gráfico fue propuesto por Tukey para presentar datos numéricos, especialmente útil para comparar distribuciones de distintas variables.

Para realizar un Box-Plot se deben tener en cuenta los siguientes puntos:

1. Ordenar los datos de menor a mayor.
2. Calcular la mediana, el cuartil inferior Q_1, el cuartil superior Q_3 y el rango de la distribución.
3. Dibujar una escala que cubra el rango de variación de los datos y marcar la mediana y los cuartiles. Dibujar una caja que se extienda entre los cuartiles y marcar en ella la posición de la mediana.



El rango inter-cuartil (RIC) se define como la diferencia entre el tercer y primer cuartil.
El gráfico Box-Plot indica las siguientes características de una distribución de datos:

- Muestra las medidas de posición.
- Permite estudiar la simetría de la distribución.

- Da un criterio de detección de datos atípicos (outliers).

Los distintos software estadísticos dibujan Box-Plots que no siempre se basan en los criterios que hemos detallado aquí, algunos cambian el modo de calcular los cuartiles, otros por ejemplo, ofrecen opciones de indicar la media y no la mediana en la caja.

Ejemplo 16. Dados los datos del ejemplo 15:

			Q_1 = 140				**Me= 170**			**Q_3 = 320**			
Datos	104	112	134	146	155	168	170	195	246	302	338	412	678
Posición	1	2	3	4	5	6	7	8	9	10	11	12	13

Valor mínimo = $x_m = 104$

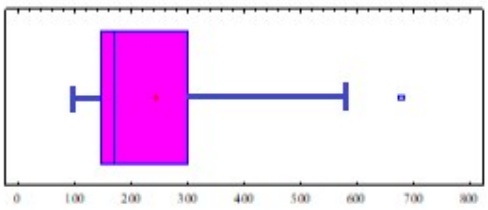
Valor máximo = $x_M = 678$

RIC = $Q_3 - Q_1 = 320 - 140 = 180$

$L_i = Q_1 - RIC.1{,}5 = 140 - 180.1{,}5 = -130$, como $x_m = 104$ es mayor que L_i, la primera línea del gráfico comienza en el valor mínimo 104.

$L_s = Q_3 + RIC.1{,}5 = 320 + 180.1{,}5 = 590$, como el valor $x_M = 678$ es mayor que L_s, la segunda línea del gráfico finaliza en el valor $L_s = 590$ y no en el valor máximo. Siendo así, el dato 678 un dato atípico.

La caja se extiende entre el primer cuartil y el tercero, en ella se marca la mediana.

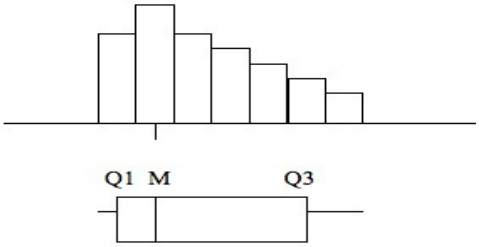


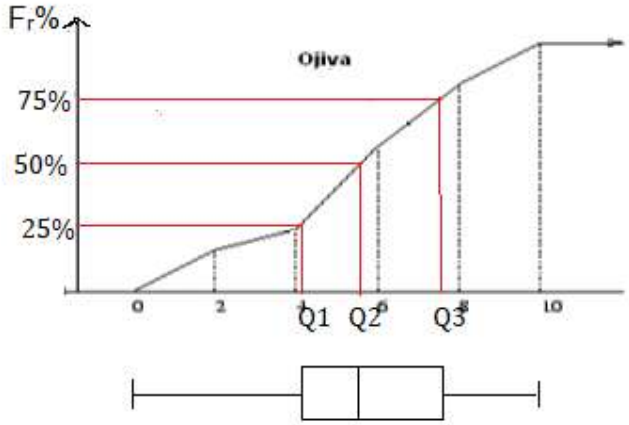
Se observa:

- Un dato atípico (outlier).
- La distribución de los datos es asimétrica positiva, la mitad inferior de los datos se distribuye en un rango mucho menor que la mitad superior.

Relación entre gráficos

Histograma – Gráfico de caja - Ojiva





7. Ejercicios

Se adjunta una Guía de Ejercicios para resolver.