Matemáticas para Ingenieros

ESTADÍSTICA DESCRIPTIVA

Ing. Adriana M. Apaza

La Estadística es la Ciencia de la

Descriptiva

- **Sistematización, recogida, ordenación y presentación** de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico, con objeto de

Probabilidad

- **deducir las leyes** que rigen esos fenómenos,

Inferencia

- y poder de esa forma hacer previsiones sobre los mismos, tomar **decisiones** u obtener **conclusiones**.

- **Población:**

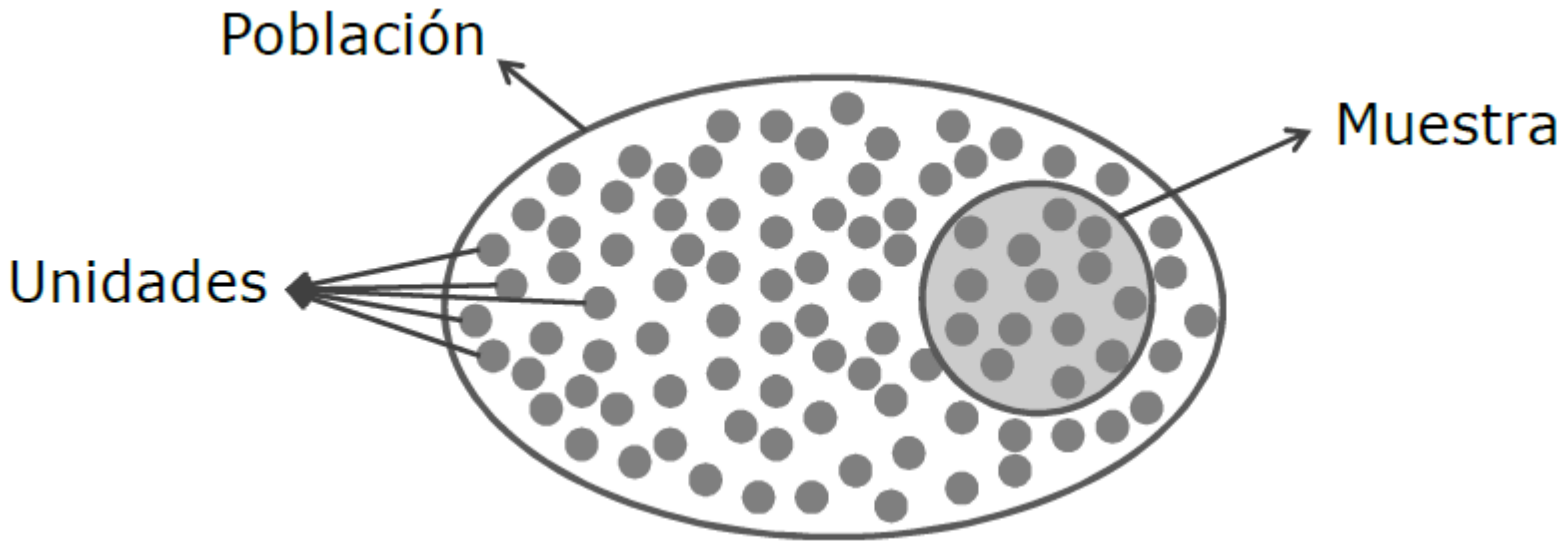
  Conjunto definido de TODOS los INDIVIDUOS, de donde se observa cierta característica.

  Al número de observaciones en la población se llama **tamaño de la población** y se representa con la letra **N.**

- **Muestra:** Subconjunto de una población, que intenta reflejar las características de la población lo mejor posible.

  El número de observaciones en la muestra, llamado **tamaño de la muestra** se representa con la letra **n**.

- **Individuo:** Es el elemento de la población o de la muestra que aporta información sobre lo que se estudia.



- **Variable:**

  Característica o propiedad de los individuos que se desea estudiar y se puede medir o calificar; cambia o varía con el tiempo en un individuo dado, o cambia o varía de elemento a elemento.

Ej. Edad, peso, sexo, estado civil, número de hijos, etc.

- **Dato:**

  Valor que se obtiene al realizar la medición de la característica de la variable en estudio.

  Los datos pueden ser **datos cuantitativos** o **datos cualitativos**.

CLASIFICACIÓN DE LAS VARIABLES

- Cualitativos: No toman valores numéricos y describen cualidades o atributos, clasificándolos en una de varias categorías, es decir, no son valores numéricos. Ej:
  - Sexo: f/m.
  - Hábito de fumar: Fumador/No fumador
  - Color de ojos: negro, azul, marrón, …
  - Religión: católica, evangélica, …
  - Estado civil: soltero, casado, divorciado,…

- Cuantitativas: son variables que pueden medirse, cuantificarse o expresarse numéricamente. Ejemplos:

  - Peso
  - Edad
  - Estatura
  - Presión
  - Humedad
  - Intensidad de un sismo
  - Cantidad de hermanos

- **Tipos de variables cuantitativas**:

  - **Discretas**: toman valores que surgen de un proceso de conteo.

    Ejemplo: cantidad de hermanos.

  - **Continuas**: es la variable que puede tomar cualquier valor en una escala continua.

    Ejemplo: cantidad de líquido contenido en un recipiente.

Escalas de medición:

Los datos recopilados pueden también describirse de acuerdo al nivel de medición que se logre.

Una medición es establecer números o categorías o códigos a las observaciones mediante escalas adecuadas. Las escalas se diferencian por propiedades de orden y distancia.

Escalas de Medición

- Escala Nominal
- Escala Ordinal

} Variables Cualitativas

- Escala de Intervalo
- Escala de Razón

} Variables Cuantitativas

- **Escala Nominal:**

Los datos se pueden agrupar en categorías que no mantienen una relación de orden entre sí.

Ejemplo: sexo, estado civil, color de ojos, deporte favorito, carrera a estudiar, etc.

- **Escala Ordinal:**

  Los valores de la variable tienen un ORDEN con un nivel específico, pero no se pueden hacer operaciones aritméticas entre ellas.

  Ejemplo:

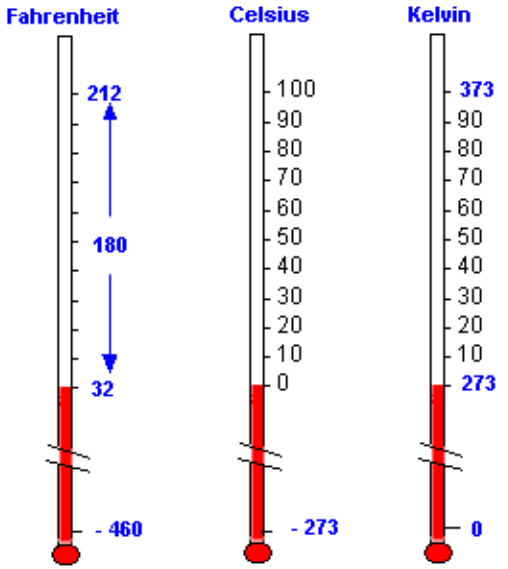
  estudios realizados: Primaria – Secundaria – Terciario- Universitario

- **Escala de Intervalo:**

  En ella existe un orden entre los valores de la variable y además una NOCIÓN DE DISTANCIA

  El cero o punto de inicio no es único, es más bien un punto de referencia.

  Ejemplo: Escalas de temperatura, la edad de la Tierra, la línea del tiempo de la humanidad.



- **Escala de Razón:**

  La magnitud tiene SENTIDO FÍSICO, existe el cero absoluto, existe orden, se puede determinar cuántas veces es mayor uno que otro.

  Ejemplo: peso, estatura, edad, distancia, dinero, etc.

Organización de los datos

- Una vez que se ha realizado la recolección de los datos, se obtienen datos en bruto, los cuales rara vez son significativos sin una organización y tabulación.

31	18	10	27	14	31	18	10	27	14
12	24	16	28	20	12	24	16	28	20
13	30	22	9	26	13	30	22	9	26
19	11	23	15	32	19	11	23	15	32
25	17	29	21	8	25	17	29	21	8
31	18	10	27	14	31	18	10	27	14
12	24	16	28	20	12	24	16	28	20
13	30	22	9	26	13	30	22	9	26
19	11	23	15	32	19	11	23	15	32
25	17	29	21	8	25	17	29	21	8


Presentación ordenada de datos
Tablas de frecuencias
Representaciones gráficas

ESTADÍSTICA DESCRIPTIVA: Datos cualitativos

Ordenando la Información

Al ordenar datos es usual agruparlos en *clases* o *categorías*. Al determinar cuántos pertenecen a cada clase, establecemos la *frecuencia.* Construimos así una tabla de datos llamada tabla de frecuencias.

¿Para qué se construyen las tablas de frecuencias?

1- Ordenar
2- Agrupar
3- Resumir información

El formato general de una tabla estadística, llamada también **TABLA DE FRECUENCIAS O TABLA DE DISTRIBUCIÓN DE FRECUENCIAS** es la siguiente:

Nombre de la variable	**Frecuencia**
Categoría 1	Frecuencias observadas
Categoría 2	Frecuencias observadas
TOTAL	n

Una variable estadística X puede tomar distintos valores X1,X2,…,Xk, pero cada uno de éstos puede aparecer repetido más de una vez. El número total de observaciones es n (frecuencia total)

TIPOS DE FRECUENCIAS

a) **Frecuencia absoluta**: Es el número de veces que se presenta un valor o categoría de una variable. Se representa por f.

b) **Frecuencia relativa**: Es la razón entre la frecuencia absoluta y el total de los datos. Se representa por $f_{r.}$

c) **Frecuencia relativa porcentual**: si a las frecuencias relativas multiplicamos por cien. Se representa por 100 f_r(%)

Variable cualitativa

Ejemplo: Dentro de los procesos industriales de gran importancia para el Ingeniero Químico, están los procesos de tratamiento de aguas. Un laboratorio determinó la dureza del agua de 10 muestras obteniendo los siguientes resultados.

Muestra	Dureza
1	Agua blanda
2	Agua blanda
3	Agua dura
4	Agua muy dura
5	Agua muy dura
6	Agua extremadamente dura
7	Agua blanda
8	Agua blanda
9	Agua dura
10	Agua muy dura

Construir la tabla de distribución de frecuencias relativas para la variable X =”Dureza del agua”.

Dureza del agua	f (frecuencia absoluta)	fr frecuencia relativa)	fr % (fr porcentual)
Agua blanda	4	0.4	40
Agua dura	2	0.2	20
Agua muy dura	3	0.3	30
Agua extremadamente dura	1	0.1	10
TOTAL	10	1.0	100

4/10

2/10

0,4*100

Siempre es el número total

Siempre es 1

Siempre es 100

GRÁFICOS

El propósito de los gráficos es mostrar la información contenida en las tablas de frecuencias en forma precisa y clara.

Algunos requisitos recomendables al construir un gráfico son:

- Evitar distorsiones por escalas exageradas.
- Elección adecuada del tipo de gráfico, según los objetivos y tamaño de recorrido de las variables.
- Sencillez y autoexplicación.

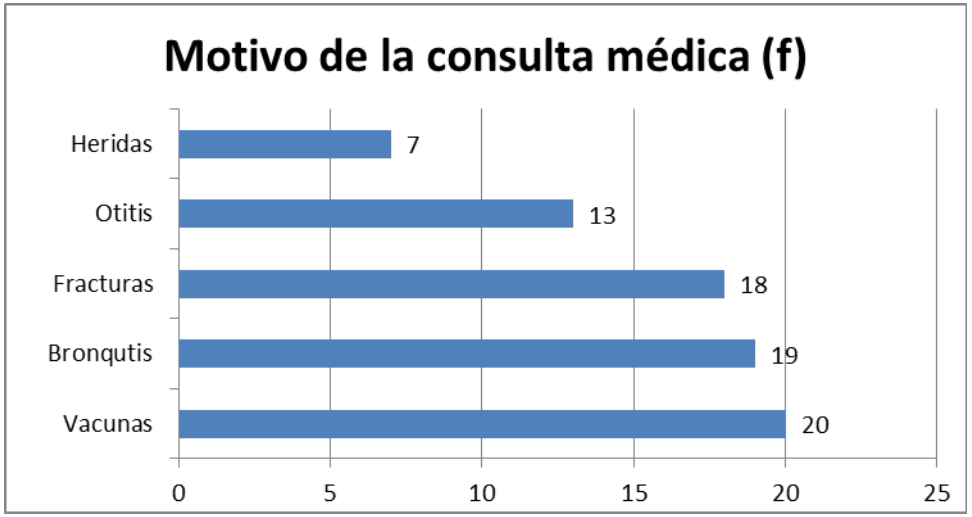
En función de la naturaleza de los datos y de la forma en que estos se presentan existen distintos tipos de representaciones

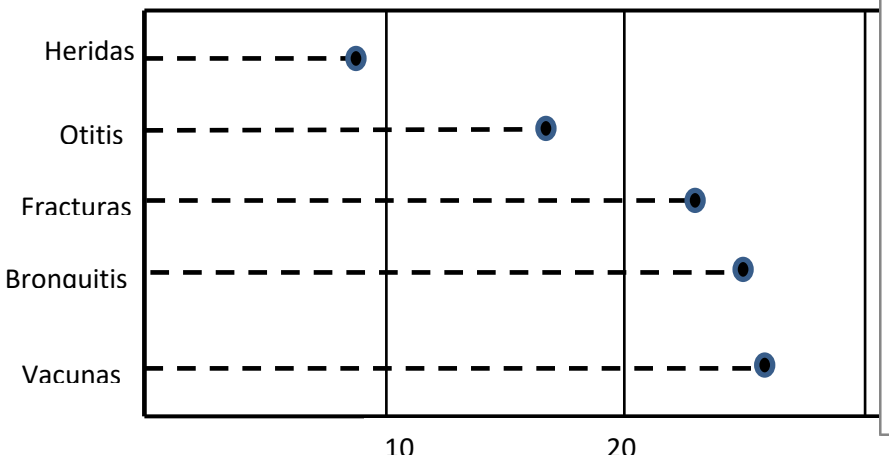
- **Gráfico de barras**

- **Se usa para representar distribuciones de frecuencias de una variable categórica.**
- **Las barras deben ser de igual ancho y estar igualmente espaciadas**
- **Alturas proporcionales a las frecuencias (absolutas o relativas)**
- **Para respuestas categóricas o cualitativas las barras se disponen en forma horizontal, mientras que para respuestas numéricas se disponen en forma vertical.**

Tabla 1: Se presenta el MOTIVO DE LA CONSULTA MÉDICA, durante una semana.

Motivo consulta	Número de pacientes (f)
Bronquitis	19
Otitis	13
Heridas	7
Fracturas	18
Vacunas	20
TOTAL	77





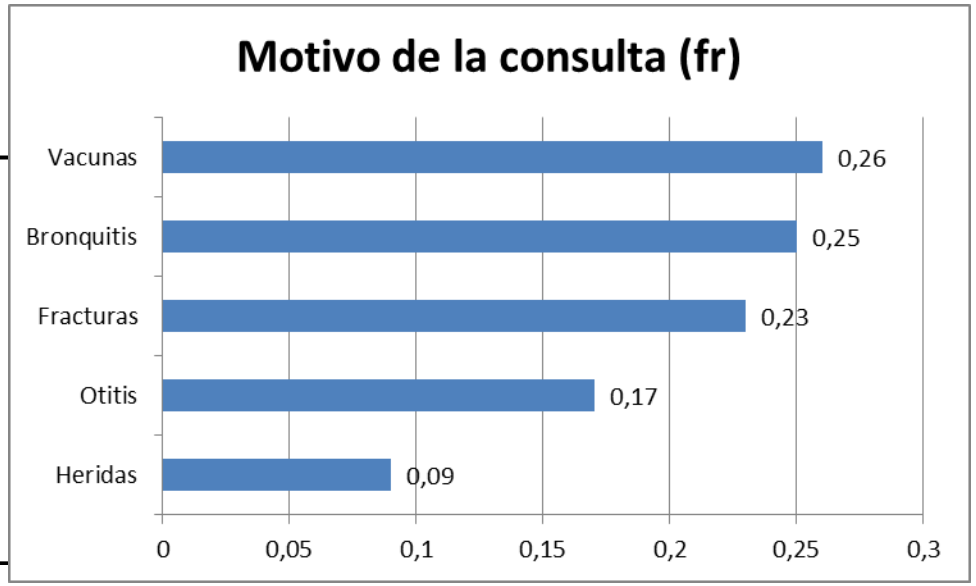
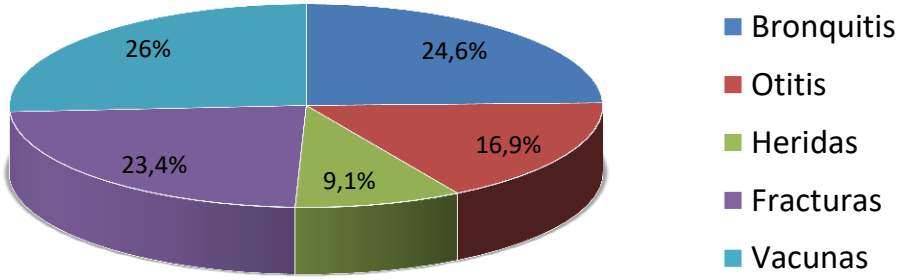


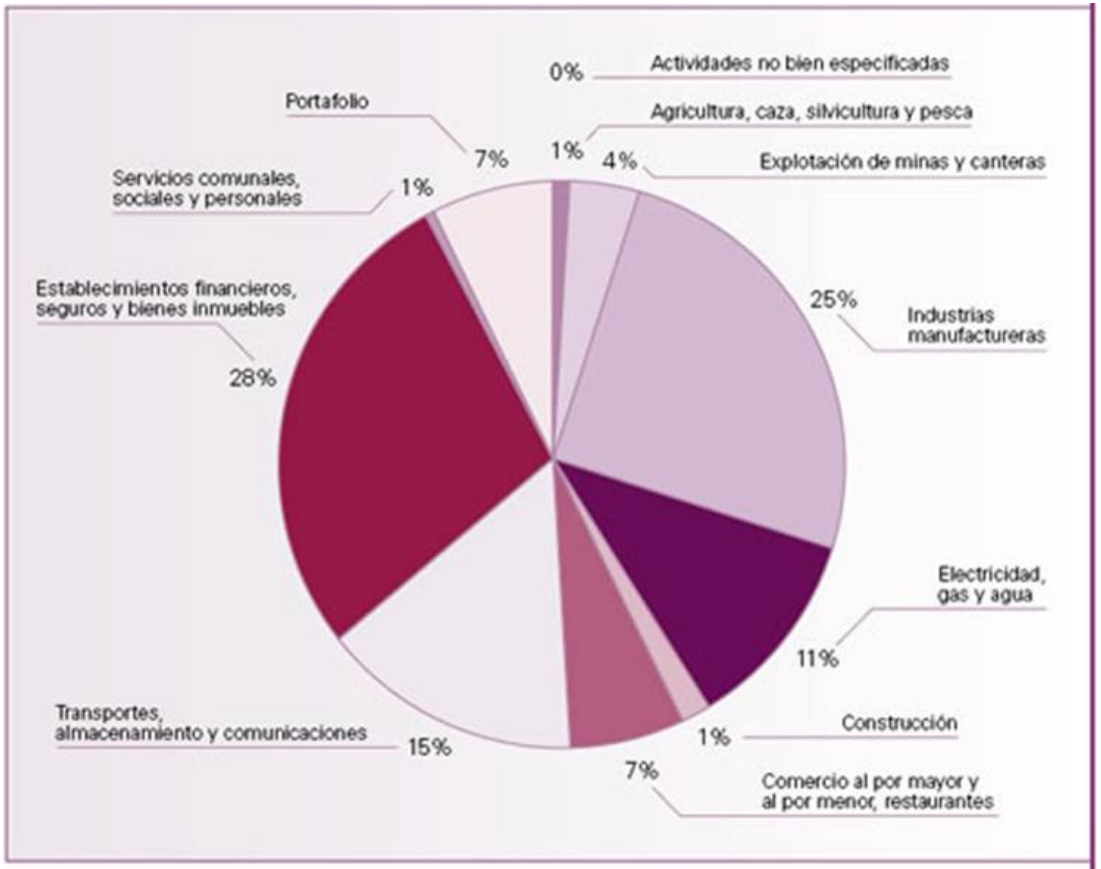
Diagrama circular dividido en sectores

Presenta las proporciones o frecuencias relativas de cada uno de los valores de la variable. Este tipo de diagrama es particularmente útil si se quiere hacer hincapié en los tamaños relativos de las componentes de los datos Es un gráfico que se basa en una proporcionalidad entre la frecuencia y el ángulo central de una circunferencia, de tal manera que a la frecuencia total le corresponde el ángulo central de 360º.

0.25 x 360°=90°

0.26 x 360°=93,6°





Fuente: Banco de la República y Departamento Nacional de Planeación.

Diagrama de Tallos y Hojas

Permite observar al conjunto de datos como un todo y destacar algunas características, tales como:

- La simetría del conjunto de datos
- La variabilidad de los datos
- La presencia o no de “outliers” (datos atípicos)
- Concentración de los datos
- Brechas en el conjunto de los datos

Procedimiento

- Cada dato se divide en dos partes, una conocida como **tallo, que se pone en una** primera columna, y la otra que se denomina **hoja, que se pone en fila en frente al** tallo correspondiente.

Para datos con 2 dígitos, por ejemplo 65 se escribirá

6 | 5.

Para datos de 3 dígitos, por ejemplo 265 se escribirá

26 | 5

- Cada tallo define una clase. El número de hojas la frecuencia de dicha clase.

- Cuando se observan muchas hojas en cada línea, existe la posibilidad de dividir las líneas repitiendo los tallos.
  Se pueden considerar dos líneas por cada tallo:

  Primera línea "*": 0, 1, 2, 3, 4

  Segunda línea "•": 5, 6, 7, 8, 9

Otra opción es considerar cinco líneas por cada tallo.
La notación propuesta es:

1* se ubican las hojas 0 y 1

t se ubican las hojas 2 y 3 (two, three)

f se ubican las hojas 4 y 5 (four, five)

s se ubican las hojas 6 y 7 (six, seven)

1· Se ubican las hojas 8 y 9

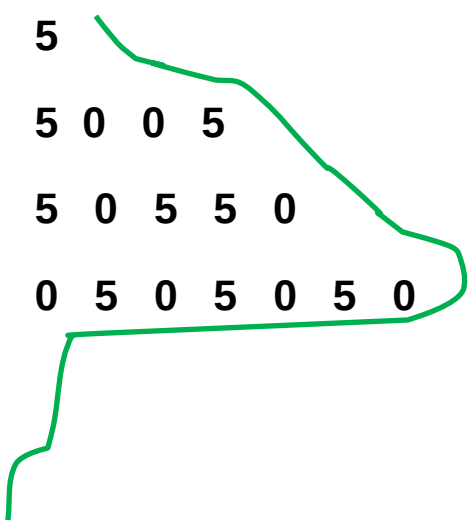
Ejemplo:

En un programa para la detección de hipertensión en una muestra de 30 hombres en edades entre 30 y 40 años, la distribución de la presión diastólica (mínima) en mm Hg fue la siguiente:

70	**85**	**85**	**75**	**65**	**90**	**110**	**95**	**90**	**70**
60	75	80	120	85	95	90	70	100	65
80	90	95	90	95	110	100	85	80	75

Diagrama de tallo y hojas: Presión diastólica [mm de Hg]

Profundidades o Frecuencia	Hojas	(Unidad = 5)
3	6	0 5 5
6	7	0 5 5 0 0 5
7	8	0 5 5 0 5 5 0
9	9	0 5 0 5 0 5 0 5 0
2	10	0 0
2	11	0 0
1	12	0
n = 30		



60	65	65	70	70	70	75	75	75	80
80	80	85	85	85	85	90	90	90	90
90	95	95	95	95	100	100	110	110	120

Los siguientes datos corresponden a tiempos de falla de cables Kevlar 49/epoxy sometidos a una presión del 90%:

0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.04	0.05	0.06
0.07	0.07	0.08	0.09	0.09	0.10	0.10	0.11	0.11	0.12
0.13	0.18	0.19	0.20	0.23	0.80	0.80	0.83	0.85	0.90
0.92	0.95	0.99	1.00	1.01	1.02	1.03	1.05	1.10	1.10
1.11	1.15	1.18	1.20	1.29	1.31	1.33	1.34	1.40	1.43
1.45	1.50	1.51	1.52	1.53	1.54	1.54	1.55	1.58	1.60
1.63	1.64	1.80	1.80	1.81	2.02	2.05	2.14	2.17	2.33
3.03	3.03	3.24	4.20	4.69	7.89				

En nuestro caso, cada dato lo separamos en su parte entera (tallo) y su parte decimal (hoja)

Se pueden considerar dos líneas por cada tallo:
Primera línea “*”: 0, 1, 2, 3, 4
Segunda línea “•”: 5, 6, 7, 8, 9

Diagrama de tallo y hojas: tiempo de falla

Profundidades		(Unidad = 0.1)
25	**0***	**0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 2 2**
8	**0·**	**8 8 8 8 9 9 9 9**
18	**1***	**0 0 0 0 0 1 1 1 1 1 2 2 3 3 3 4 4 4**
14	**1·**	5 5 5 5 5 5 5 5 6 6 6 8 8 8
5	**2***	**0 0 1 1 3**
	2·	
3	**3***	0 0 2
	3·	
1	**4***	2
1	**4·**	6
	5*	
	5·	
	6*	
	6·	
	7*	
1	**7·**	8
n = 76		

Variable cuantitativa

Cuando el número de datos es pequeño (<10), pueden presentarse mediante una Tabla de Distribución de Frecuencias de la Variable.

¿Qué crees que deberíamos hacer si tenemos muchos datos?

Se agrupan los datos en CLASES o CATEGORÍAS → **TABLA DE FRECUENCIAS DE DATOS AGRUPADOS**

Las clases o categorías deben formar un sistema exhaustivo y excluyente

-Exhaustivo: no podemos olvidar ningún posible valor de la variable.

- Excluyente nadie puede presentar dos valores simultáneos de la variable

TABLA DE FRECUENCIAS DE DATOS AGRUPADOS

- Se recomienda su uso cuando se tienen grandes cantidades de datos
- Comenzamos por determinar el número de clases o categorías a considerar. Para definir la cantidad de intervalos de clase (k) se puede usar

✓ la regla de Sturges:

$$k = 1 + 3.3\log(n)$$, con n número de datos

✓Tomar el número de clases igual al entero más próximo a $2\sqrt{n}$, siendo n el número de datos.

✓ Tomar un número de intervalo de clase entre 5 y 20 dependiendo de los datos.

- ✓ La amplitud de todas las clases deberá ser la misma. Para determinar la amplitud, h de cada intervalo hacemos h = OSCILACIÓN/k,

  k es el número de categorías

  OSCILACIÓN o RANGO es la distancia entre el valor máximo y el mínimo.

- ✓ Marca de clase: $\frac{límite\ superior + límite\ inferior}{2}$

Ejemplo:

En un programa para la detección de hipertensión en una muestra de 30 hombres en edades entre 30 y 40 años, la distribución de la presión diastólica (mínima) en mm Hg fue la siguiente:

70	**85**	**85**	**75**	**65**	**90**	**110**	**95**	**90**	**70**
60	75	80	120	85	95	90	70	100	65
80	90	95	90	95	110	100	85	80	75

La variable en estudio es: **Presión diastólica (medida en mm de Hg),** una variable numérica continua.

1. Calcule el rango (R) o recorrido

$$R = x_{máx.} - x_{mín.} \Rightarrow R = 120 - 60 = 60$$

2. Determine el número de intervalos (K).

$$K = 1 + 3.3\ Log(n) = 1+3.3Log(30)$$

$K = 5.875 \approx 6$ (siempre es un número entero, se aproxima por exceso)

3. Determine la amplitud del intervalo de clase (h).

$$h = R/K = 10$$

En este caso , entonces, la tabla de frecuencias tendrá aproximadamente 6 clases de amplitud 10 unidades en cada clase.

TABLA DE FRECUENCIAS
PRESIÓN DIASTÓLICA

Intervalo de Clase	Marca de clase	F Frecuencia absoluta	fa Frecuencia acumulada	fr Frecuencia relativa	fr (%) Frec. Rel. porcentual	fa (%) Frec. Ac. porcentual
60-70	65	3	3	0.1	10	10
70-80	75	6	9	0.2	20	30
80-90	85	7	16	0.23	23	53
90-100	95	9	25	0.3	30	83
100-110	105	2	27	0.07	7	90
110-120	115	2	29	0.07	7	97
120-130	125	1	30	0.03	3	100
Total		30		1.00	100	

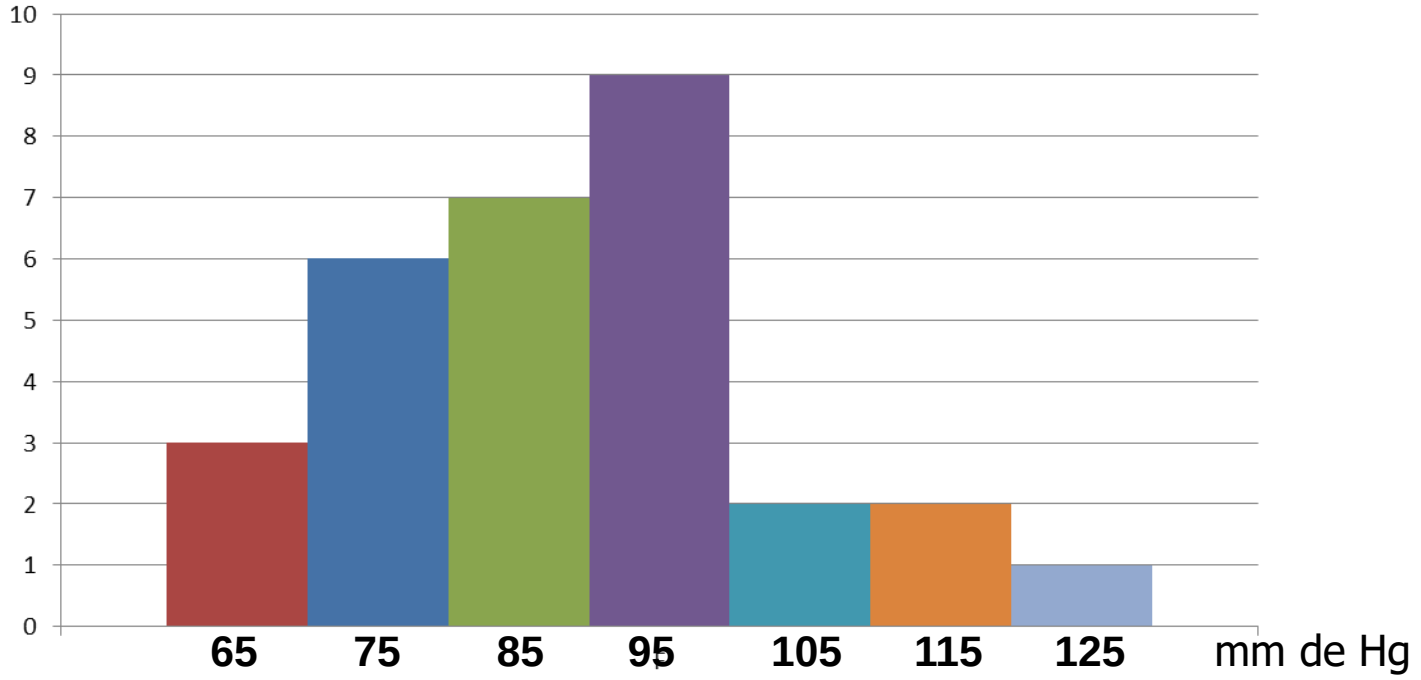
Para datos cuantitativos agrupados en clases, comúnmente se utilizan tres gráficos:

- Histogramas.
- Polígono de frecuencias.
- Ojiva o Polígono de frecuencias acumuladas

HISTOGRAMA

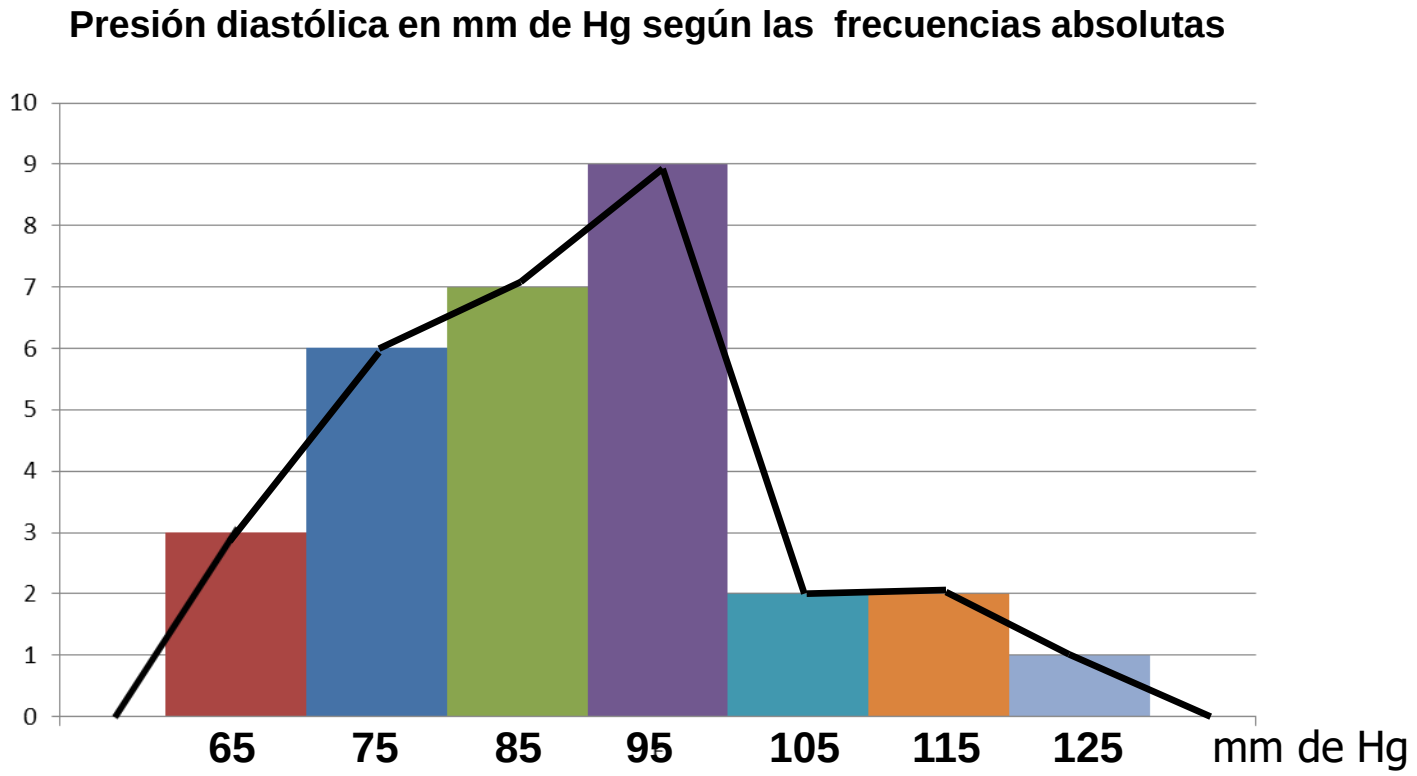
- Está formado por rectángulos, cuyas bases corresponden con los intervalos de clase y sus áreas son iguales o proporcionales a sus frecuencias.
- Las categorías se dibujan a lo largo del eje horizontal, sobre el eje vertical puedo indicar frecuencias, frecuencias relativas o frecuencias relativas porcentuales. Se parece a un diagrama de barras sólo que no hay espacio entre las barras.

Histograma de la distribución de presión diastólica en mm de Hg según las frecuencias absolutas:



Polígono de Frecuencias

Es una línea poligonal que une los puntos medios de las bases superiores de los rectángulos de un histograma..



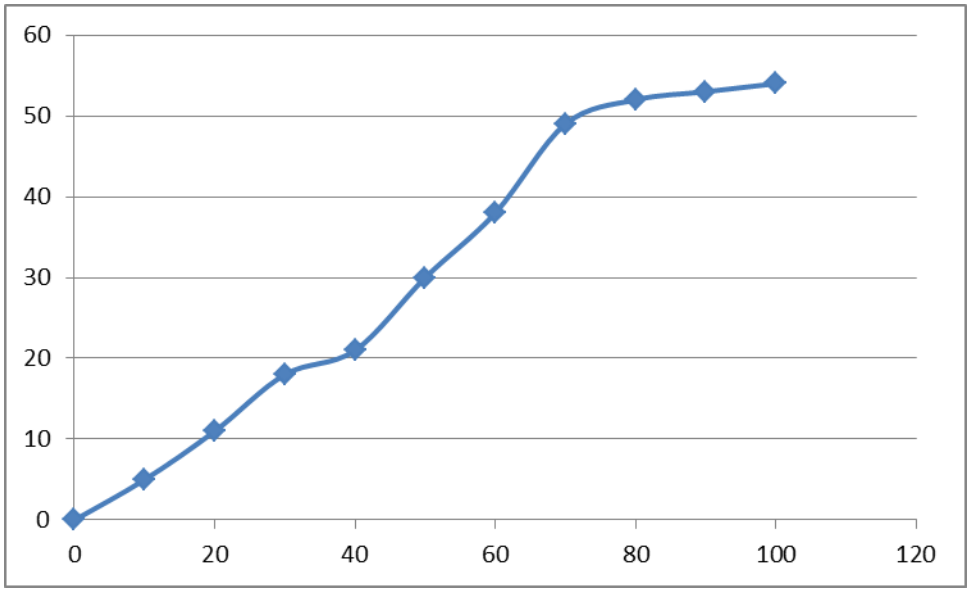
Ojiva o Polígono Porcentual Acumulado

En este tipo de gráfico, el objetivo es representar distribuciones de frecuencias de variables cuantitativas continuas, pero sólo para **frecuencias acumuladas porcentuales**.

El fenómeno de interés se representa sobre el eje horizontal, en tanto que los porcentajes sobre el eje vertical

Tabla de frecuencias – Calificaciones

Intervalo de clase	fabs	facum
0-10	5	5
10-20	6	11
20-30	7	18
30-40	3	21
40-50	9	30
50-60	8	38
60-70	11	49
70-80	3	52
80-90	1	53
90-100	1	54

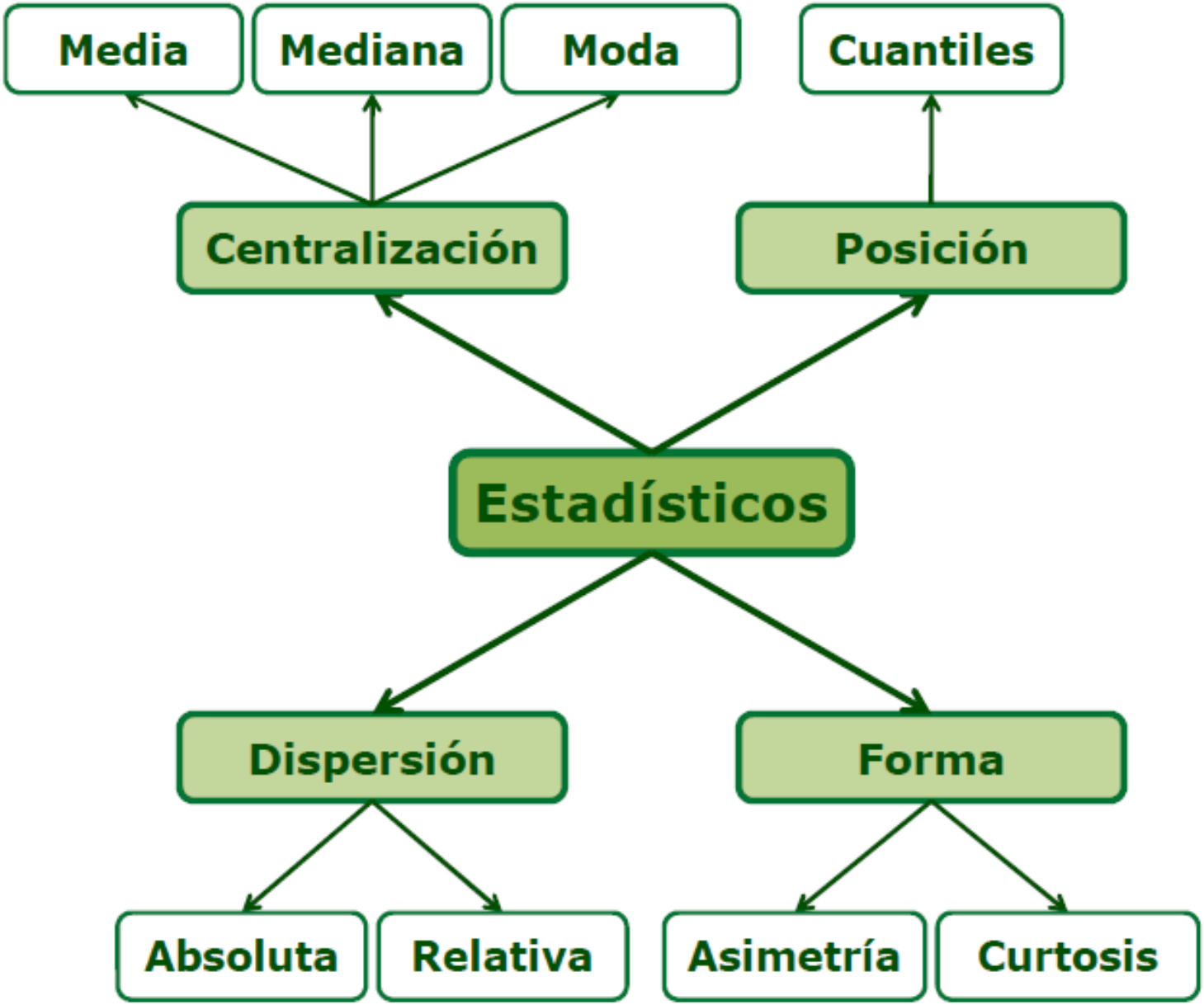


Nota

Parámetro: cualquier característica medible de una población, por ejemplo, la media de la población (μ).

Estadístico: cualquier característica medible de una muestra, por ejemplo, la media muestral ($\bar{X}$).

Símbolos	Población Parámetro	Muestra Estadístico
Tamaño de la muestra	N	n
Media aritmética	μ	$\bar{X}$
Varianza	σ^2	S^2
Desviación estándar	σ	S


Media
Mediana
Moda
Cuantiles
Centralización
Posición
Estadísticos
Dispersión
Forma
Absoluta
Relativa
Asimetría
Curtosis

PROPIEDADES DE LOS DATOS CUANTITATIVOS

Posición

Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos:

Cuantiles: percentiles, cuartiles, deciles,…

Medidas de tendencia central

Indican valores con respecto a los que los datos parecen agruparse:

Media aritmética, mediana, moda, rango medio u oscilación media, eje medio

Dispersión

Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización: **Desviación típica, coeficiente de variación, rango, varianza**

Forma

- Asimetría
- Apuntamiento o curtosis

Medidas de Posición

Se define el **cuantil** de orden α como un valor de la variable por debajo del cual se encuentra una frecuencia acumulada α.

Casos particulares son los percentiles, cuartiles, deciles, quintiles,...

Percentiles son los valores de la variables que dividen al conjunto de datos ordenados en cien partes iguales

Percentil de orden k = cuantil de orden k/100

La **mediana** es el percentil 50.

El percentil de orden 15 deja por debajo al 15% de las observaciones. Por encima queda el 85%.

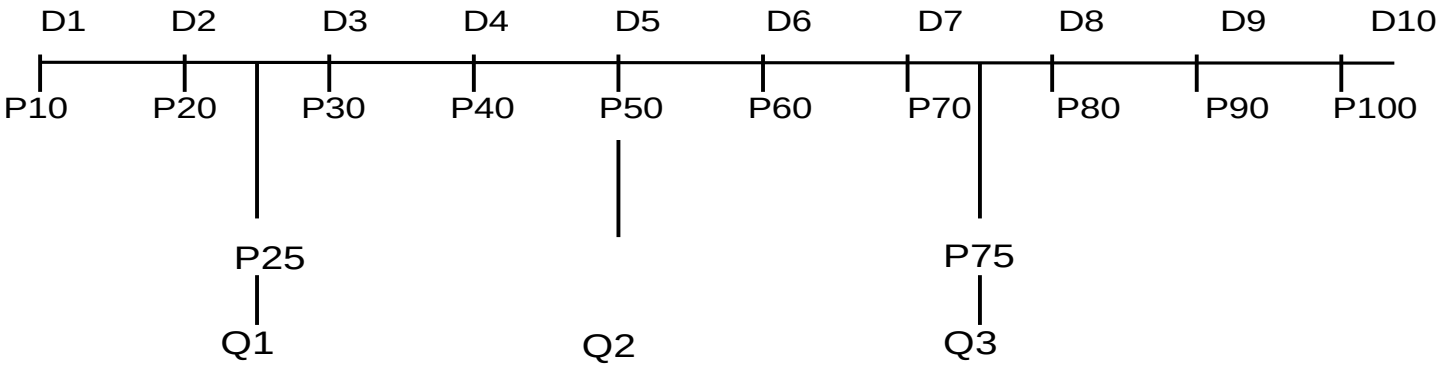
Cuartiles: Dividen a la muestra en 4 grupos con frecuencias similares.

Primer cuartil = Percentil 25 = Cuantil 0,25.

Segundo cuartil = Percentil 50 = Cuantil 0,5 = **mediana.**

Tercer cuartil = Percentil 75 = cuantil 0,75.

- Requisitos
  - Variables cuantitativas
  - Los resultados están ordenados de menor a mayor
- Equivalencias
  - Como todas las medidas se refieren al mismo grupo de datos, se pueden hacer equivalentes entre sí



Medidas de tendencia central

Añaden unos cuantos casos particulares a las medidas de posición. Son medidas que buscan posiciones (valores) con respecto a los que los datos muestran tendencia a agruparse.

1. **Media Aritmética:** Es la media aritmética (promedio) de los valores de una variable.

• Conveniente cuando los datos se concentran simétricamente con respecto a ese valor. Muy sensible a valores extremos.

• Centro de gravedad de los datos.

• Se calcula como la suma de los valores dividido por el tamaño muestral.

Cálculo de la Media Aritmética

- Para datos no agrupados:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Para datos agrupados en intervalos

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n}$$

Donde: x_i: punto medio de la clase i

f_i: frecuencia absoluta de la clase i

k: cantidad de clases

2. Mediana

- Es el valor que ocupa la posición central de un conjunto de observaciones, una vez que han sido ordenados en forma ascendente o descendente.
- Divide al conjunto de datos en dos partes iguales.
- Es conveniente cuando los datos son asimétricos.
- No es sensible a valores extremos.

Cálculo de la mediana

Para datos no agrupados:

- Si n es impar: posición donde se ubica la mediana es igual a (n+1)/2.

  Mediana de 1, 2, 4, **5**, 6, 6, 8 es 5

- Si n es par: mediana será igual al promedio de las dos posiciones centrales.

  Mediana de 1, 2, 4, **5**, **6**, 6, 8, 9 es (5+6)/2 = 5,5

Para datos agrupados

Clase mediana es la que contiene a la observación que ocupa la posición n/2.

$$M_E = L_2 + \frac{\frac{n}{2} - (\sum f)_2}{f_{ME}} C$$

Donde: L_2 : límite inferior de la clase mediana

n: número de datos

$(\sum f)_2$: suma de todas las frecuencias inferiores a la mediana o frecuencia acumulada de la clase anterior a la clase mediana

f_{ME} : frecuencia de la clase mediana

C: tamaño del intervalo de la clase mediana

3. Moda

- Observación o clase que tiene la mayor frecuencia en un conjunto de observaciones.

- Puede no existir y en caso de existir no ser única.

- Es la única medida de tendencia central que se puede determinar para datos de tipo cualitativo.

- Es más variable para distintas muestras que las demás medidas de tendencia central.

- A diferencia de la media aritmética, la moda no se afecta ante la ocurrencia de valores extremos.

Cálculo de la moda

Para datos no agrupados

Es simplemente la observación que más se repite.

Para datos agrupados

- Como el punto medio o marca de clase del intervalo modal
- A través de la siguiente fórmula de interpolación

$$Mo = L_1 + \frac{f_1}{f_1 + f_2} C$$

Donde:

L_1 : límite inferior del intervalo modal

f_1 : diferencia entre la frecuencia de la clase modal y la anterior

f_2 : diferencia entre la frecuencia de la clase modal y la posterior

C: Amplitud de la clase modal

4- Rango Medio u Oscilación Media

El Rango Medio (R.M.) es el promedio de las observaciones mayor y menor de un conjunto de datos

$$R.M. = \frac{x_{mín} + x_{máx}}{2}$$

Si hay observaciones extremas se distorsiona como medida de tendencia central

5- Eje Medio (E.M.)

El E.M. es el promedio del primer y Tercer Cuartil de una serie de datos.

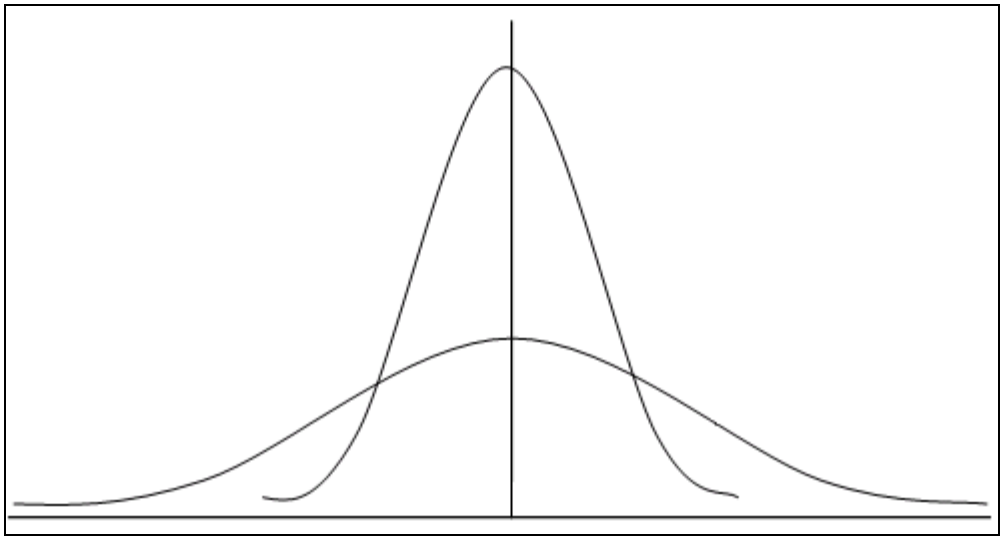
No se ve afectada por las observaciones extremas

$$E.M. = \frac{Q_1 + Q_3}{2}$$

Medidas de dispersión, variación o variabilidad.

• Son valores numéricos que indican o describen la forma en que las observaciones están dispersas o diseminadas, con respecto al valor central.

• Son importantes debido a que dos muestras de observaciones con el mismo valor central pueden tener una variabilidad muy distinta.



Medidas de dispersión, variación o variabilidad.

- Rango.
- Rango Intercuartílico
- Varianza Muestral.
- Desviación Estándar Muestral.
- Coeficiente de variación.

Rango (R)

El R es la diferencia entre las observaciones mayor y menor de un conjunto de datos

$$R = x_{máx} - x_{mín}$$

Mide la dispersión total del conjunto de datos
No toma en consideración la forma en que se distribuyen los datos entre los valores más pequeños y los más grandes

Rango Intercuartílico(RI)

El RI o propagación media es la diferencia entre el tercer y primer cuartil en una serie de datos

$$R = Q_3 - Q_1$$

Varianza y Desviación Estándar Muestral

Es un valor numérico que mide el grado de dispersión relativa porque depende de la posición de los datos $x_1, x_2, \ldots, x_n$ con respecto a la media.

Es el promedio al cuadrado de las desviaciones de cada observación con respecto a la media.

Notación: s^2, σ^2, var(X)

Cálculo de la varianza

Para datos no agrupados

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

Fórmula abreviada

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \left[\left(\sum_{i=1}^{n} x_i^2\right)^2 / n\right]^2}{n-1}$$

Para datos agrupados

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

Fórmula abreviada

$$s^2 = \frac{\sum_{i=1}^{k} f_i x_i^2 - \left[\left(\sum_{i=1}^{k} f_i x_i\right)^2 / n\right]^2}{n-1}$$

La desviación estándar

- Es la raíz cuadrada de la varianza.
- Notación: s, σ.

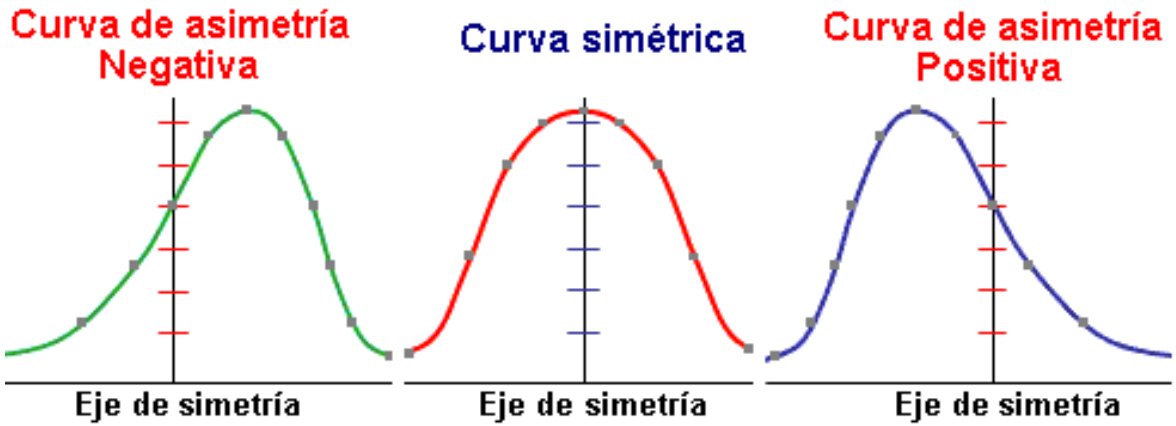
Coeficiente de variación

$$CV = \frac{s}{\bar{x}} \times 100\%$$

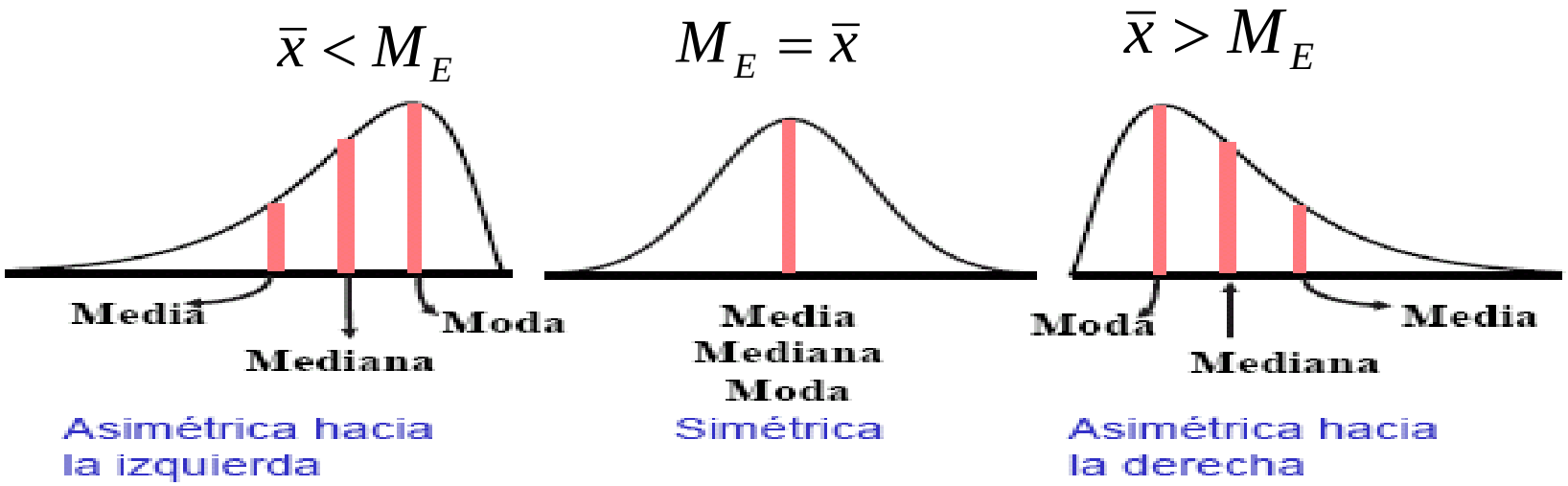
- Es una medida de dispersión relativa, mide la dispersión de los datos respecto a la media.
- No tiene dimensiones.
- Permite comparar la variabilidad de dos o más muestras de datos expresados en diferentes unidades de medición
- Notación: CV

Forma

La forma, es la propiedad que describe la manera en que se distribuyen los datos. Una distribución de datos puede ser simétrica o insesgada o sesgada o asimétrica.



Para describir la forma, lo que se requiere, es comparar la media y la mediana



En la mayor parte de los conjuntos de datos, gran parte de las observaciones tienden a aglutinarse en torno a la mediana del siguiente modo:

El aglutinamiento ocurre hacia valores mayores que la mediana

Las observaciones tienden a distribuirse en forma homogénea alrededor de la mediana o la media

El aglutinamiento ocurre hacia valores menores que la mediana

La desviación estándar y la forma

Cuando no se da un sesgo extremo, se puede utilizar la siguiente regla empírica:

Se encontrará que aproximadamente dos de cada tres observaciones, es decir el 67 %, están comprendidas dentro de una distancia de una desviación estándar en torno a la media, en el intervalo $(\bar{x} - S, \bar{x} + S)$ y que aproximadamente entre el 90 y el 95 % de las observaciones están comprendidas en una distancia de dos desviaciones estándar en torno a la media, en el intervalo $(\bar{x} - 2S, \bar{x} + 2S)$

Regla de Chebyschev para una distribución

Sin importar cómo se distribuye un conjunto de datos, el porcentaje de observaciones que están contenidas dentro de K desviaciones estándar en torno a la media, debe ser, cuando menos:

$$[1-(1/k^2)]100\ \%$$

Para una distribución de datos resulta:

Por lo menos el $[1-(1/2^2)]100\ \% = 75\ \%$ de las observaciones debe estar contenidas dentro de 2 desviaciones estándar en torno a la media (K = 2), es decir estarán en el intervalo $(\bar{x}-2S, \bar{x}+2S)$

Por lo menos el $[1-(1/3^2)]100\ \% = 88.89\ \%$ de las observaciones debe estar en el intervalo $(\bar{x}-3S, \bar{x}+3S)$

Por lo menos el $[1-(1/4^2)]100\ \% = 93.75\ \%$ de las observaciones debe estar en el intervalo $(\bar{x}-4S, \bar{x}+4S)$

Diagrama de cajas o box-plot

En 1977, Tukey presentó un simple método gráfico-cuantitativo que resume varias de las características más destacadas de un conjunto de datos. Tal método se conoce con el nombre de **gráfico de caja o box-plot**.

Una **gráfica de caja es una herramienta útil para mostrar la forma y la dispersión de los** datos

Las características de los datos incorporadas por este gráfico son:

a) centro o posición del valor más representativo,
b) dispersión,
c) naturaleza y magnitud de cualquier desviación de la simetría
d) identificación de los puntos no usuales o atípicos, o sea puntos marcadamente alejados de la masa principal de datos.

La presencia de datos atípicos producen cambios drásticos en la media muestral ($\overline{x}$) y la desviación estándar muestral (s), no así en otras medidas que son más *resistentes o robustas*, como lo son la mediana muestral y una medida de dispersión llamada *rango intercuartil* (RIQ).

Pasos a seguir para la construcción del *box plot* :

Paso 1: Ordenar los datos de menor a mayor.

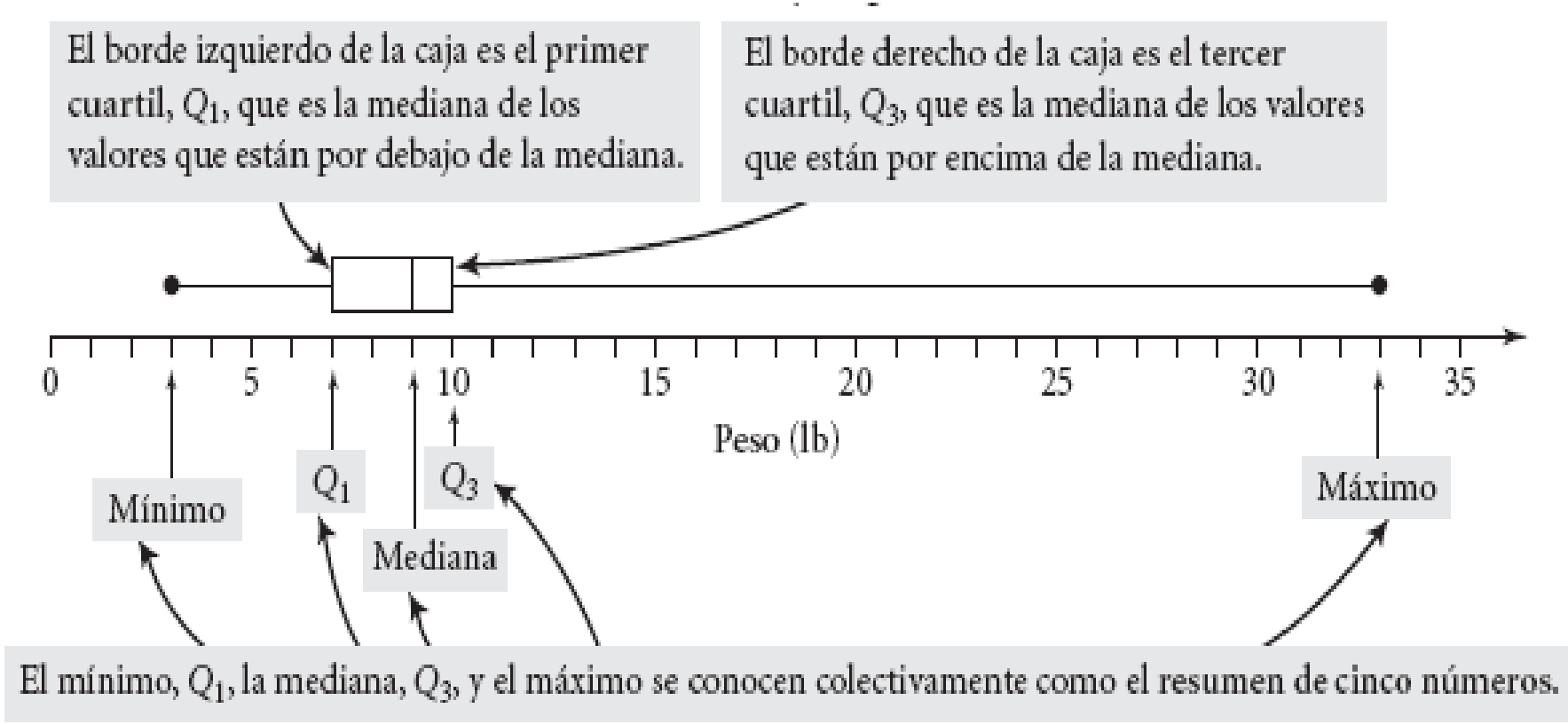
Paso 2: Calcular la mediana muestral, el primer cuartil (Q1), el tercer cuartil (Q3), y el RIQ.

Paso 3: Sobre un eje horizontal dibujar una caja cuyo borde izquierdo sea el cuartil Q1 y el borde derecho el cuartil Q3.

Paso 4: Dentro de la caja marcar con un punto la posición de la mediana y trazar un segmento perpendicular cuya posición corresponde al valor de la mediana.

Paso 5: Trazar segmentos desde cada extremo de la caja hasta las observaciones más alejadas, que no superen (1.5 RIQ) de los bordes correspondientes.

Paso 6: Si existen observaciones que superen (1.5 RIQ) entonces marcarlos con circunferencias aquellos puntos comprendidos entre (1.5 RIQ) y (3 RIQ) respecto del borde más cercano, estos puntos se llaman valores alejados, y con asteriscos aquellos puntos que superen los (3 RIQ) respecto de los bordes más cercanos, estos puntos se llaman *puntos muy alejados*.


El borde izquierdo de la caja es el primer cuartil, Q_1, que es la mediana de los valores que están por debajo de la mediana.
El borde derecho de la caja es el tercer cuartil, Q_3, que es la mediana de los valores que están por encima de la mediana.
0
5
10
15
20
25
30
35
Peso (lb)
Mínimo
Q_1
Mediana
Q_3
Máximo
El mínimo, Q_1, la mediana, Q_3, y el máximo se conocen colectivamente como el resumen de cinco números.