Estadística aplicada a la Investigación

Docente: BC. Aníbal Espínola Cano

CLASIFICACIÓN DE LA ESTADÍSTICA

ESTADÍSTICA DESCRIPTIVA

Ordenando la Información

Al ordenar datos muy numerosos, 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
Categorías o Recorrido de la variable	Frecuencias Observadas
TOTAL	n

VARIABLES NOMINALES Y ORDINALES
TABLA DE FRECUENCIAS

FRECUENCIAS	DEFINICIÓN	OBSERVACIÓN
Absoluta	f_a	N° de veces que se repite el valor x_i
Relativa	$f_r = f_a / n$	Proporción de unidades de observación que toman el valor x_i
Porcentual	$p_i = f_r * 100$	Proporción porcentual.
Acumulada Absoluta	F_a	Frecuencia absoluta acumulada hasta el valor x_i de la variable.
Acumulada relativa	F_r	Proporción de unidades de observación hasta el valor x_i de la variable.
Acumulada porcentual	P_i	Proporción acumulada porcentual.

TIPOS DE FRECUENCIAS

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

b) Frecuencia Relativa: La frecuencia relativa se puede expresar en términos de porcentaje o de proporción y se representa por fr. (Es la razón entre la frecuencia absoluta y el total de datos)

En la siguiente tabla se presenta el motivo de la consulta médica, durante una semana.

Motivo Consulta	Número de pacientes
Bronquitis	19
Otitis	13
Heridas	7
Fracturas	18
Vacunas	20

Ejemplo

Los siguientes datos corresponden a las notas obtenidas por un curso de **24 alumnos** en un trabajo de matemáticas:

3,2 4,2 5,6 6,0 2,8 3,9 4,2 4,2 5,0
5,0 3,9 3,9 3,2 3,2 4,2 5,6 6,0 6,0
3,2 6,0 4,2 5,0 5,6 5,0

Ordenando estos datos en una tabla:
Nombre de variable: Notas
Frecuencia Absoluta
Frecuencia relativa (ambas)

3,2 4,2 5,6 6,0 2,8 3,9 4,2 4,2 5,0
5,0 3,9 3,9 3,2 3,2 4,2 5,6 6,0 6,0
3,2 6,0 4,2 5,0 5,6 5,0

Nota	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Relativa Porcentual (%)
2,8	1	0,041	4,166
3,2	4	0,166	16,666
3,9	3	0,125	12,500
4,2	5	0,208	20,833
5,0	4	0,166	16,666
5,6	3	0,125	12,500
6,0	4	0,166	16,666

Distribución de frecuencia de pacientes con HTA según estado de enfermedad al ingresar al estudio.

Estado de HTA (OMS)	Frecuencia	Frecuencia acumulada	Frecuencia relativa (%)	Frecuencia relativa acumulada (%)
I	631	631	56,9	56,9
II	325	956	29,3	86,3
III	152	1108	13,7	100,0
Total	1108		100,0	

MEDIDAS DE RESUMEN

¿Qué medida o estadístico usar en una situación determinada?

Dependerá de los **objetivos del estudio** y **del nivel de medición** de la variable.

Estadísticos: reciben este nombre las medidas o valores estadísticos que proceden o son obtenidos a partir de las muestras.

Parámetros: son medidas o valores estadísticos que caracterizan una población.

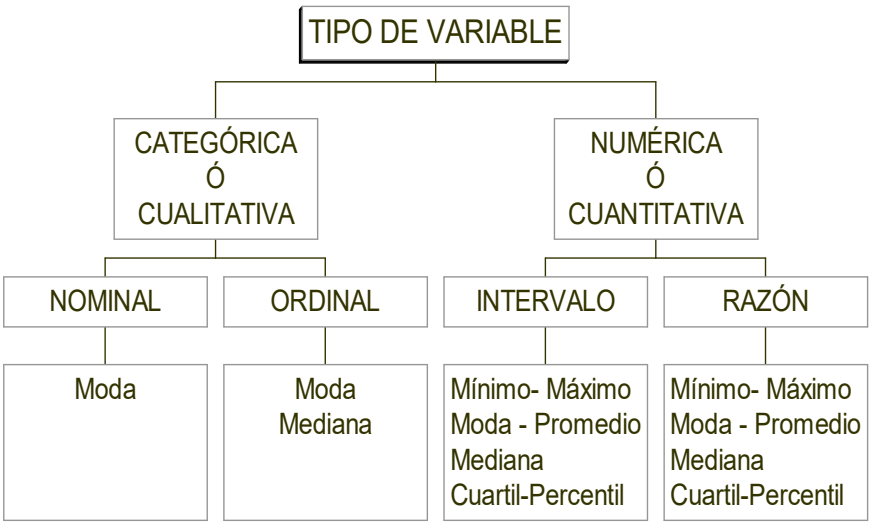
Los estadísticos o parámetros descriptivos se clasifican en:

- **Medidas de posición o de tendencia central**
- **Medidas de dispersión**
- **Medidas de forma**

Un brevísimo resumen sobre estadísticos

- Posición
  - Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.
    - Cuantiles, percentiles, cuartiles, deciles,...
- Centralización
  - Indican valores con respecto a los que los datos parecen agruparse.
    - Media, mediana y moda
- Dispersión
  - Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.
    - Desviación típica o estándar, coeficiente de variación, rango, varianza
- Forma
  - Asimetría
  - Apuntamiento o curtosis

Selección de una medida de posición adecuada



MEDIDAS DE POSICIÓN y TENDENCIA CENTRAL

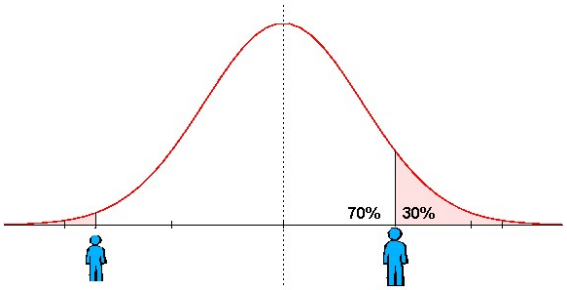
Medidas	Cálculos
Moda	Es aquel valor de la variable con mayor frecuencia.
Media	$\bar{x} = (\sum x_i \cdot f_i) / n$
Mediana	n impar: $M_d = x_{(n+1/2)}$ n par: $M_d = (x_{n/2} + x_{(n+1/2)}) / 2$
Cuartil	$Q_p = x_{(n.\,p\,/\,4)}$
Percentil	$P_p = x_{(n.\,p\,/\,100)}$

DESCRIPCIÓN DE DATOS

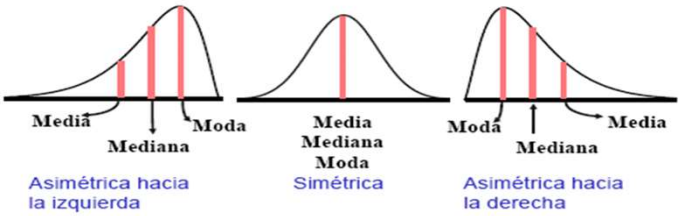
- **Posición:**

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,...



- **Tendencia central** (media, mediana, moda): indica entorno a que valor se agrupan los datos



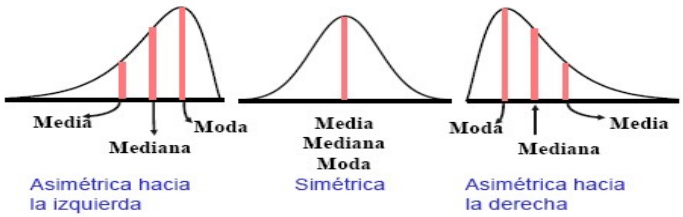
Distribución normal:
media aritmética, geométrica, armónica.

Distribución asimétrica (sesgada):
mediana

Distribución no agrupada en torno a un valor: moda

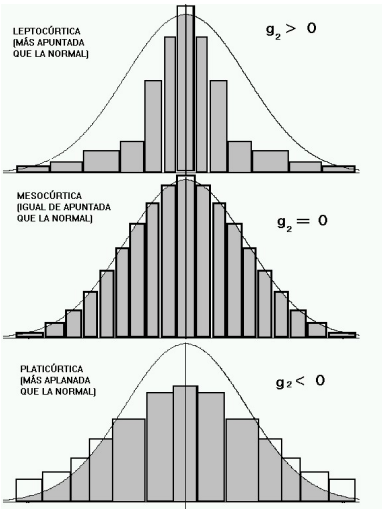
- **Asimetría**

  El concepto de asimetría se refiere a si la curva que forman los valores de la serie presenta la misma forma a izquierda y derecha de un valor central (media aritmética)



- **Curtosis**

  Nos indica el grado de apuntamiento (aplastamiento) de una distribución con respecto a la distribución normal o gaussiana. Es adimensional.



Interpretación de las medidas de posición: Peso

Peso	Moda	Mediana	Media
Total	68	72	72,36 ≅ 72
F	68	68	68,63 ≅ 69
M	76	76	76,84 ≅ 77

Media: el promedio de puntaje obtenido por los pacientes es 72 kg. Analizando según sexo se tiene un promedio de 69 kg para las mujeres y de 77 kg en los varones

Mediana: la mitad de los pacientes en estudio han obtenido 72 kg. de peso o menos, según sexo es de 68 kg en las mujeres y 76 kg o menos en los hombres.

Moda: 68 kg. es el peso que más se repite en general en los pacientes investigados, coincide con el sexo femenino, en cambio en los hombre predomina 76 kg.

DISCUSIÓN

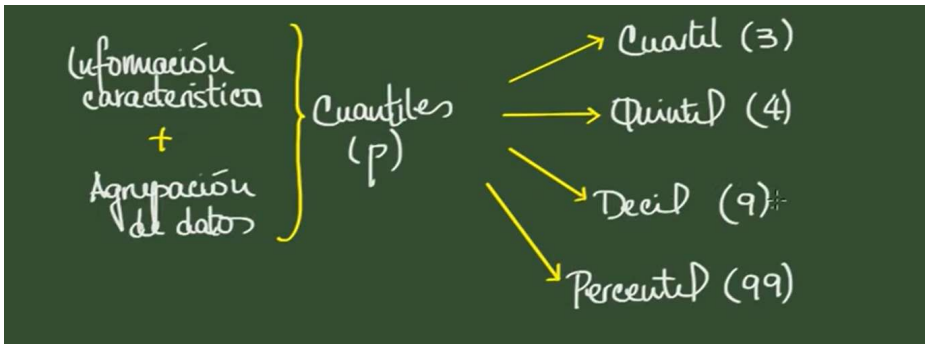
Tener en cuenta que:

La MEDIA: Se descentra cuando hay algún dato numérico muy distante del resto.

La MODA: Se descentra si hay pocos datos e incluso en ese caso es fácil que aparezcan varias modas.

La MEDIANA: Obtiene el centro posicional y no tiene en cuenta los valores, salvo el central o los dos centrales.

Medida de posición no centrales



https://www.youtube.com/watch?v=dB-QwndRdDc

- **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

Ejercicio de ejemplo

- Se muestra el tiempo en minutos logrados de 20 alumnos en una prueba de 20 metros planos
- Hallar los valores correspondientes a Q1, Q2 y Q3
- Tiene sentido buscar percentiles en esta serie de datos?

2,0	1,9	2,6	2,0
1,3	2,4	1,2	1,6
1,8	2,2	2,8	2,3
1,9	2,6	2,3	2,8
1,7	1,5	1,7	2,5

Pasos

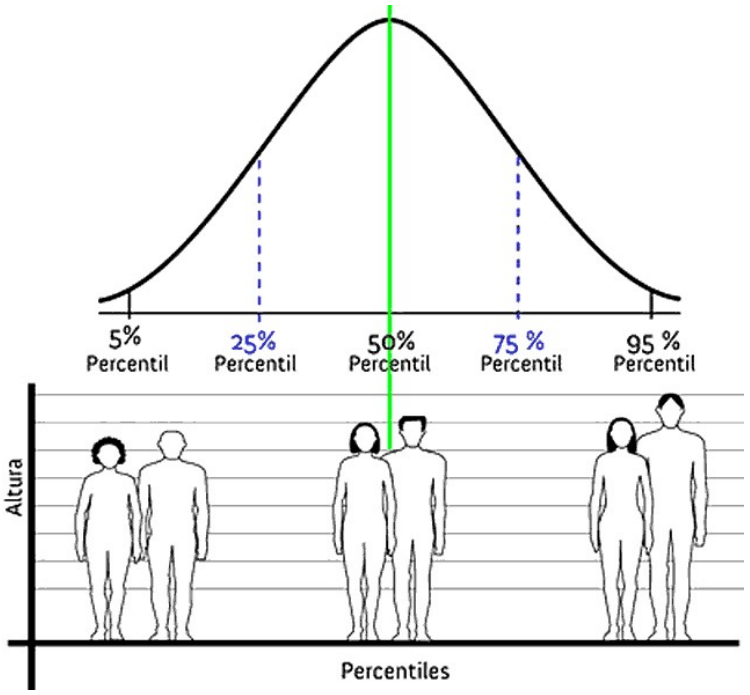
- 1-Ordenar los valores
- 2-Numero par de datos .
- 3-Calcular las medianas

1,2
1,3
1,5
1,6
1,7 (1,7+1,7)/2=1,7 Q1 25%
1,7
1,8
1,9
1,9
2 (2+2)/2=2 Q2 50%
2
2,2
2,3
2,3
2,4 (2,4+2,5)/2 =2,4 5 Q3 75%
2,5
2,6
2,6
2,8
2,8

Resultados

- **Los cuartiles**
- Q1=1,7 lo que significa que el 25% de los estudiantes estuvieron por debajo de los 1,7 min
- Q2=2 el 50% de los estudiantes estuvieron por debajo o igual a 2 min
- Q3= 2,45 el 75 % de los estudiantes estuvieron por debajo de 2,45 min.

Tiene sentido buscar percentiles en esta serie de datos ?
Se podría buscar los percentiles 25,50 y 75% pero esto no tendría sentido por en numero de muestra que solo es de 20



Percentil

Percentil q, en que q es un número entero entre 1 y 99

El valor del percentil 95 indica el valor por debajo del cual queda el 95% de los valores de la serie medida.
La mediana representa el percentil 50.

INDICADORES

- **Razón**: cociente entre dos cantidades de diferente naturaleza.

  $r = A / B$

- **Proporción**: cociente entre dos cantidades de igual naturaleza.

  $P = A / B \quad ; A \subset B$

- **Porcentaje**: proporción multiplicada por 100.

  $p\% = P * 100$

TASAS:

- **Una tasa es un cociente formado por tres elementos:**

- **Numerador**: es la frecuencia de ocurrencia de un hecho, en un periodo de tiempo dado y un área determinada.
- **Denominador**: es la población expuesta al riesgo de que le suceda el hecho que aparece en el numerador.
- **Constante**: es un número por la cual se multiplica el cociente (k = 100, 1000 ó 10000).

$$T = (A / B) * k$$

" Es necesario que en una tasa haya concordancia entre el numerador y el denominador en tres aspectos:

La naturaleza del hecho, la zona geográfica y el período de tiempo dentro del cual ocurre el hecho".

TIPOS DE TASAS:

- Se pueden distinguir dos tipos de tasas:

a) ***Tasas crudas o brutas:*** en el denominador figura el total de la población.

b) ***Tasas específicas:*** en el denominador sólo se usa un sector de la población (según sexo, grupo de edad, etc.).

Las tasas se aplican en diferentes áreas pero con mayor frecuencia en Salud Pública se usan las siguientes:

- ✓ ***Tasa bruta de mortalidad.***
- ✓ ***Tasa bruta de natalidad.***
- ✓ ***Tasa de morbilidad (frecuencia, duración y gravedad de una enfermedad).***

Ejemplo de Indicadores en Salud

Se presenta una amplia gama entre los que se mencionan:

Tasa de incidencia =

$$\frac{\text{Número de casos nuevos en el período}}{\text{Población a mitad del período}} * 100.000$$

Tasa de prevalencia=

$$\frac{\text{Número de casos existentes en un momento dado}}{\text{Población en riesgo en ese momento}} * 100.000$$

MEDIDAS DE DISPERSIÓN

Indica el grado de dispersión de un serie de valores

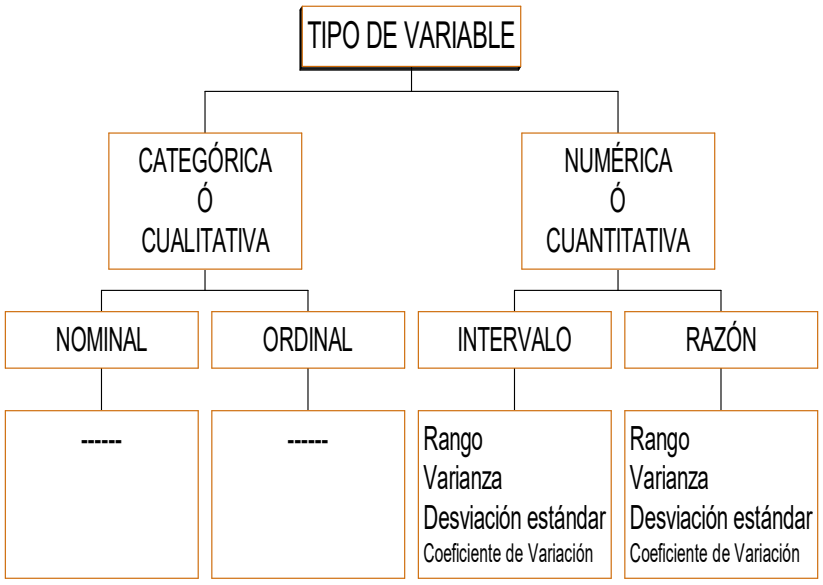
- **Dispersión :** DE, varianza, intervalo min-max, CV:

  Distribución normal: desvío estándar

  Distribución asimétrica: intervalo intercuartil (q3-q1)

El intervalo de confianza (Ej. IC_{95})conjuga tendencia central con dispersión.

Elección de la medida de DISPERSIÓN adecuada



Existen diversas medidas estadísticas de dispersión, pero muchos autores coinciden en que las principales son:

- Rango
- Rango intercuartílico
- Varianza
- Desviación estándar
- Coeficiente de variación

RANGO

Mide la amplitud de los valores de la muestra y se calcula por diferencia entre el valor más elevado (Límite superior) y el valor más bajo (Límite inferior).

FÓRMULA

$$Rango = X_{MAX} - X_{MIN}$$

Ejemplo 1.

Ante la pregunta sobre número de hijos por familia, una muestra de 12 hogares, marcó las siguientes respuestas:

2	1	2	4	1	3
2	3	2	0	5	1

Calcula el rango de la variable

Solución.

$$Rango = 5 - 0 = 5$$

Ejemplo 2.

Hay dos conjuntos sobre la cantidad de lluvia (mm) en Taipei y Seúl en un año.

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Taipei	86	135	178	170	231	290	231	305	244	122	66	71
Seúl	40	77	83	89	147	168	184	252	209	101	32	13

Calcula el rango en cada una de las ciudades.

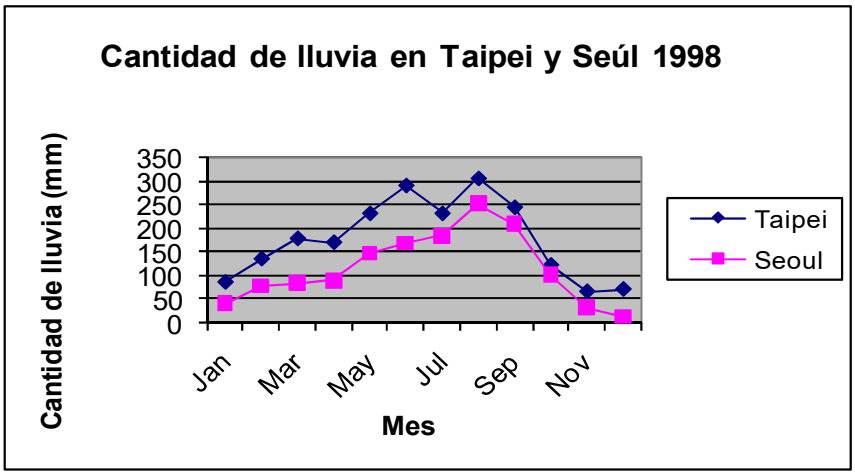
Solución.

Aplicando la fórmula correspondiente tenemos:

Taipei $Rango = 305mm - 66mm = 239mm$

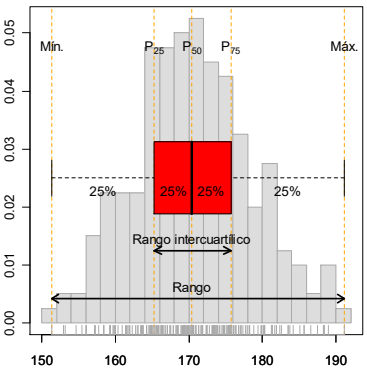
Seúl $Rango = 252mm - 13mm = 239mm$

En este caso se puede observar que el rango es el mismo para ambos casos aunque las cantidades sean diferentes.



Rango intercuartílico

- Es la distancia entre primer y tercer cuartil.
  Rango intercuartílico = $P_{75} - P_{25}$
- Parecida al rango, pero eliminando las observaciones más extremas inferiores y superiores.
- No es tan sensible a valores extremos.



VARIANZA

Mide la distancia existente entre los valores de la serie y la media. Se calcula como sumatoria de las diferencias al cuadrado entre cada valor y la media, multiplicadas por el número de veces que se ha repetido cada valor. La sumatoria obtenida se divide por el tamaño de la muestra. Es sensible a valores extremos (alejados de la media).

FÓRMULA

Muestral →

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

Poblacional →

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

La varianza siempre será mayor que cero. Mientras más se aproxima a cero, más concentrados están los valores de la serie alrededor de la media. Por el contrario, mientras mayor sea la varianza, más dispersos están.

Ejemplo 1.

Calcula la varianza para los siguientes datos

2 1 2 4 1 3 2 3 2 0 5 1

Solución.

Primero es necesario obtener la media. En este caso $\bar{x} = 2.16$

Ahora aplicamos la fórmula correspondiente

$$s^2 = \frac{(2-2.16)^2+(1-2.16)^2+(2-2.16)^2+(4-2.16)^2+(1-2.16)^2+(3-2.16)^2+(2-2.16)^2+(3-2.16)^2+(2-2.16)^2+(0-2.16)^2+(5-2.16)^2+(1-2.16)^2}{12-1}$$

$$s^2 = \frac{21.6672}{11} = 1.9697$$

DESVIACIÓN ESTÁNDAR

También llamada desviación típica, es una medida de dispersión usada en estadística que nos dice cuánto tienden a alejarse los valores puntuales del promedio en una distribución.
Específicamente, la desviación estándar es "el promedio de la distancia de cada punto respecto del promedio". Se suele representar por una S o con la letra sigma,σ, según se calcule en una muestra o en la población.
Una desviación estándar grande indica que los puntos están lejos de la media, y una desviación pequeña indica que los datos están agrupados cerca de la media.

FÓRMULA

Muestral

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Poblacional

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

COEFICIENTE DE VARIACIÓN

Es una medida de dispersión que se utiliza para poder comparar las desviaciones estándar de poblaciones con diferentes medias y se calcula como cociente entre la desviación típica y la media.

FÓRMULA

Muestral

$$CV = \frac{S}{\overline{x}} \cdot 100\%$$

Poblacional

$$CV = \frac{\sigma}{\mu} \cdot 100\%$$

RESUMEN DE MEDIDAS DE DISPERSION

	Serie Simple	Serie de Frecuencia	Intervalo de clase
RANGO O AMPLITUD	$R = x_{max} - x_{min}$		
VARIANZA	$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$	$S^2 = \frac{\sum (x_i - \bar{x})^2 . f_a}{n-1}$	$S^2 = \frac{\sum (x_i' - \bar{x})^2 . f_a}{n-1}$
DESVIACIÓN ESTÁNDAR	$DE = \sqrt{S^2}$		
COEFICIENTE DE VARIACIÓN	$C.V\% = \frac{DE}{\bar{x}} . 100$		

Nota: en el caso de poblaciones se sustituye S por σ y n - 1 por n.

REPRESENTACIONES GRÁFICAS

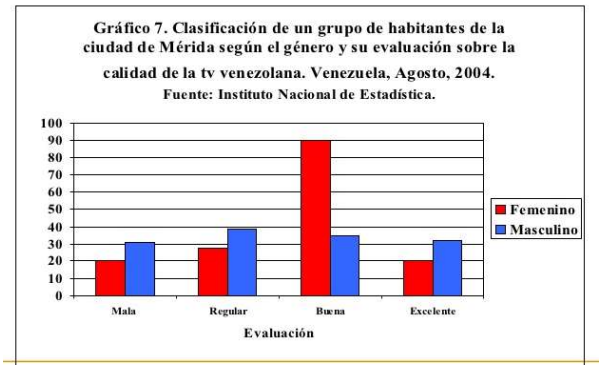
Diferentes conjuntos de datos son particularmente aptos para ciertos tipos de gráficos.

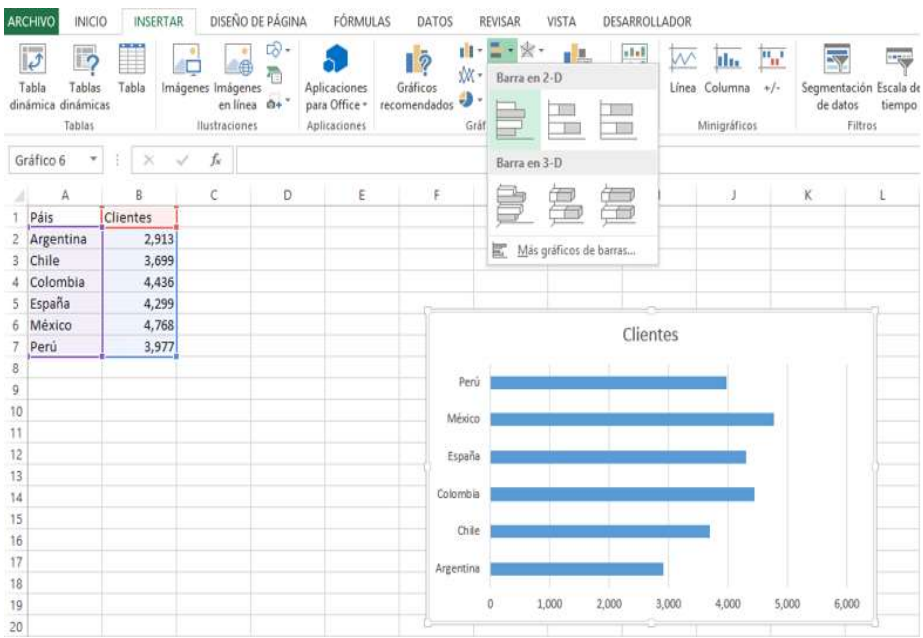
Cuadro 1.1: Principales diagramas según el tipo de variable.

Tipo de variable	Diagrama
V. Cualitativa	Barras, sectores, pictogramas
V. Discreta	Diferencial (barras) Integral (en escalera)
V. Continua	Diferencial (histograma, polígono de frecuencias) Integral (diagramas acumulados)

Barra

- El gráfico de Barras también conocido como gráfico de Columnas es una herramienta excelente para presentar o comparar varios conjuntos de datos.
- Relaciona datos de frecuencia absolutas





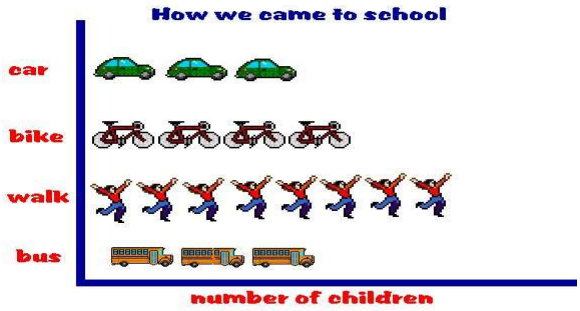
	A	B
1	Páis	Clientes
2	Argentina	2,913
3	Chile	3,699
4	Colombia	4,436
5	España	4,299
6	México	4,768
7	Perú	3,977

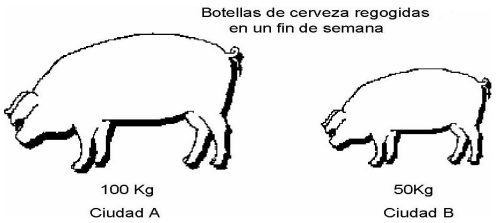
Gráfico circular

- Un gráfico circular muestra los datos como un círculo dividido en secciones de colores o diseños. Este tipo de gráfico se usa solamente con un grupo de datos .
- Relaciona la frecuencia relativa

Pictogramas

- Fáciles de entender.
- El área de cada modalidad debe ser proporcional a la frecuencia.

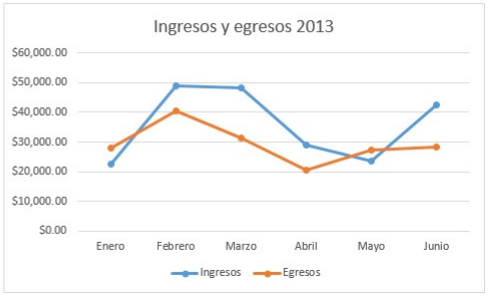




• GRÁFICO LINEAL

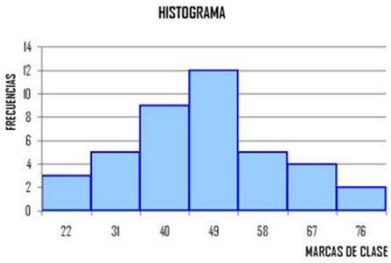
- Este gráfico presenta sus datos como una serie de puntos conectados por una línea. El gráfico de línea se usa mejor con los datos de un gran número de grupos

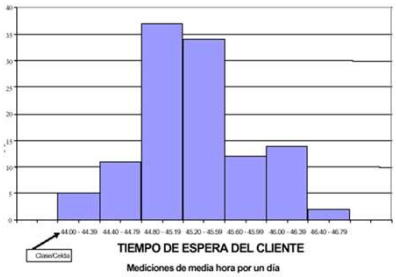
	A	B	C
1	Mes	Ingresos	Egresos
2	Enero	$22,591.00	$28,119.00
3	Febrero	$49,000.00	$40,610.00
4	Marzo	$48,362.00	$31,515.00
5	Abril	$29,127.00	$20,591.00
6	Mayo	$23,657.00	$27,190.00
7	Junio	$42,575.00	$28,363.00



- ***Histograma***

- ***Es la gráfica de la tabla de distribución de frecuencias*** para datos agrupados, consiste de barras cuyas bases son los intervalos de clases y cuyas alturas son proporcionales a las frecuencias absolutas (o relativas) de los correspondientes intervalos.
- **Un histograma permite ver la forma de la distribución de los datos, en particular, se puede ver si hay simetría con respecto al centro de la distribución**





CONSTRUCCIÓN DE UN HISTOGRAMA

Ejemplo: Datos sobre la cantidad exacta de cafe contenido en paquetes de 250 gramos (120 unidades medidas)

Paso 1 Preparación de los datos.

Paso 2 Determinar los valores extremos de los datos.

Recorrido total= Vmax-Vmin= 258 grm – 243 grm =15 grm

Paso 3 Definir las Clases

Número de datos	Número de clases recomendado
20* - 50	6
51 - 100	7
101 - 200	8
201 - 500	9
501 - 1000	10
Más de 1000	11 - 20

Amplitud de clase= recorrido / # de clase
15 grm / 8 = 1,875

257	255	249	248	258	251	252	249	251	249
248	254	250	249	248	250	252	253	252	250
243	251	247	249	246	250	247	243	250	251
249	250	255	250	254	249	245	249	256	246
250	252	253	251	256	247	255	250	243	244
251	252	246	248	247	252	251	252	246	255
248	247	249	250	252	253	252	248	249	249
247	256	251	252	252	251	251	250	257	246
245	254	252	252	250	248	248	251	248	257
249	246	250	253	251	251	254	251	244	245
250	248	250	247	254	250	253	253	251	252
251	251	247	250	255	250	251	249	247	250

CONSTRUCCIÓN DE UN HISTOGRAMA

Paso 4 Construir las clases anotando los limites de cada una de ellas

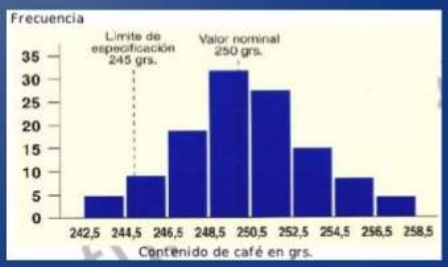
Como el valor menor es 243 grs empezamos el primer intervalo en 242,5 grs y construimos 8 clases con 2 grs de amplitud

Clase	Intervalo	Clase	Intervalo
1	De 242,5 a 244,5	5	De 250,5 a 252,5
2	De 244,5 a 246,5	6	De 252,5 a 254,5
3	De 246,5 a 248,5	7	De 254,5 a 256,5
4	De 248,5 a 250,5	8	De 256,5 a 258,5

Paso 5 Calcular la frecuencia de clase

Límites de la clase	Recuento	Total
242,5 - 244,5	𝍸	5
244,5 - 246,5	𝍸 ///	9
246,5 - 248,5	𝍸 𝍸 ///	19
248,5 - 250,5	𝍸 𝍸 𝍸 𝍸 //	32
250,5 - 252,5	𝍸 𝍸 𝍸 𝍸 /	28
252,5 - 254,5	𝍸 𝍸	15
254,5 - 256,5	𝍸 //	8
256,5 - 258,5	///	4
		120

Paso 6 Dibujar el Gráfico



- **POLÍGONO DE FRECUENCIA**

- Un polígono de frecuencias es la gráfica que se obtiene al unir en forma consecutiva con segmentos los puntos de intersección entre los puntos medios de cada clase y su frecuencia, incluyendo el punto medio anterior a la primera clase y el punto medio posterior a la última clase

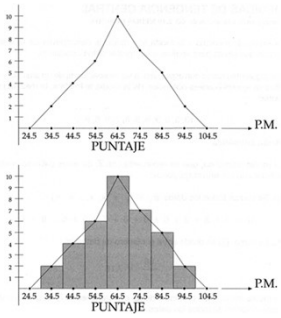
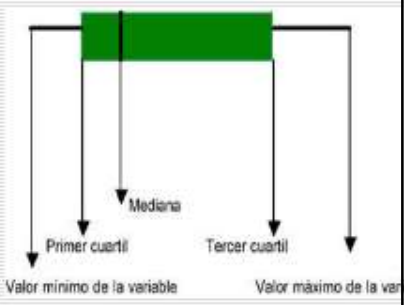


DIAGRAMA DE CAJA

Es una forma rápida de obtener una representación visual ilustrativa del conjunto de datos

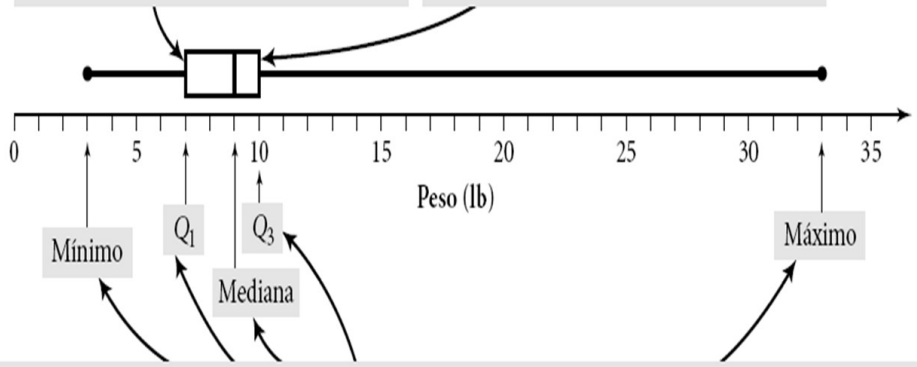
El diagrama de caja y bigote es un gráfico basado en cuartiles para representar un conjunto de datos basándose en los cuartiles Q1 y Q3, la mediana, el valor mínimo y el valor máximo de una conjuntos de datos Alaminos, (1993).



Cajas y bigotes

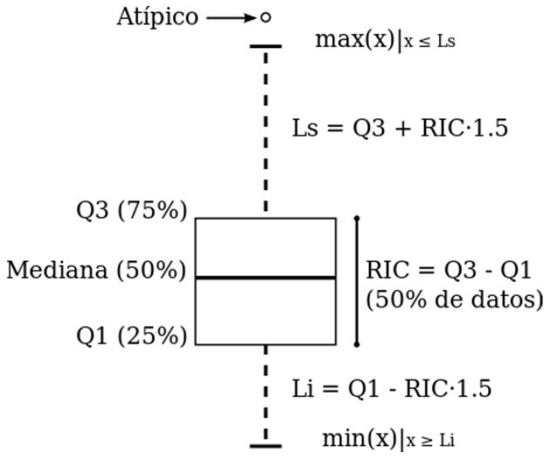
El borde izquierdo de la caja es el **primer cuartil**, Q_1, que es la mediana de los valores que están por debajo de la mediana.

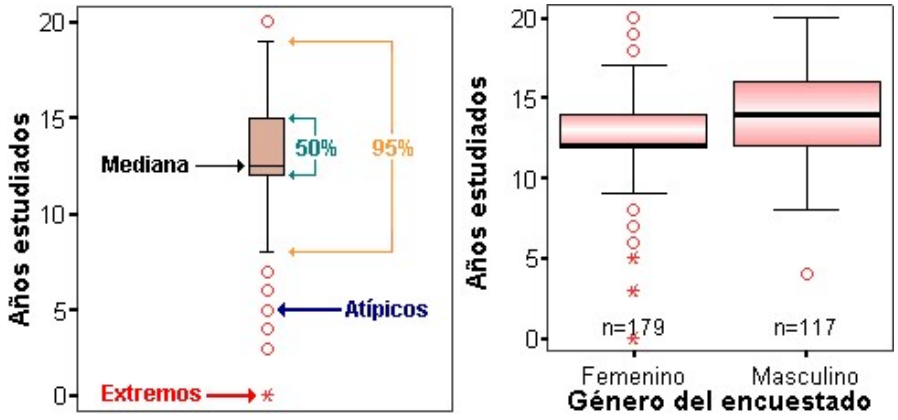
El borde derecho de la caja es el **tercer cuartil**, Q_3, que es la mediana de los valores que están por encima de la mediana.



El mínimo, Q_1, la mediana, Q_3, y el máximo se conocen colectivamente como el **resumen de cinco números.**

Construcción





Útil para determinar valores outliers o valores atípicos , antes de iniciar estudios estadísticos mas detallados . Estos valores podrían ser errores en la recogida de los datos.