ESTADÍSTICA BÁSICA

DOCENTE: JOHANNA TROCHEZ

INGENIERA INDUSTRIAL

ESPECIALISTA Y MAGÍSTER EN CIENCIAS ESTADÍSTICA

Temas a ver en este curso:

Estadística descriptiva

Métodos gráficos

Regresión lineal

Probabilidad

Bibliografía

- Guerrero P, A., Buitrago C, M. V., & Curieses P, M. d. (2010). Estadística Básica (2da ed.). Medellín: Fondo Editorial ITM.
- BERENSON, Marck L. y LEVINE David. Estadística básica en administración conceptos y aplicaciones; 6aed. México: Prentice-Hall, 1996, 943 p
- WALPOLE, Ronald y MYERS Raymon. Probabilidad y Estadística 4aed; México: Mc Graw Hill 1992, 797 p.
- CANAVOS, George. Probabilidad y estadística: aplicaciones y métodos. México: McGraw-Hill, 1988, 651 p.

¿QUÉ ES LA ESTADÍSTICA?

Estadística es la ciencia de describir o hacer inferencias sobre el mundo desde una muestra de datos

Ciencia que proporciona metodologías para recolectar organizar, resumir, presentar y analizar datos y hacer inferencias a partir de ellos

¿Por qué es importante la estadística?

Confiabilidad

Control de producción

Análisis financieros

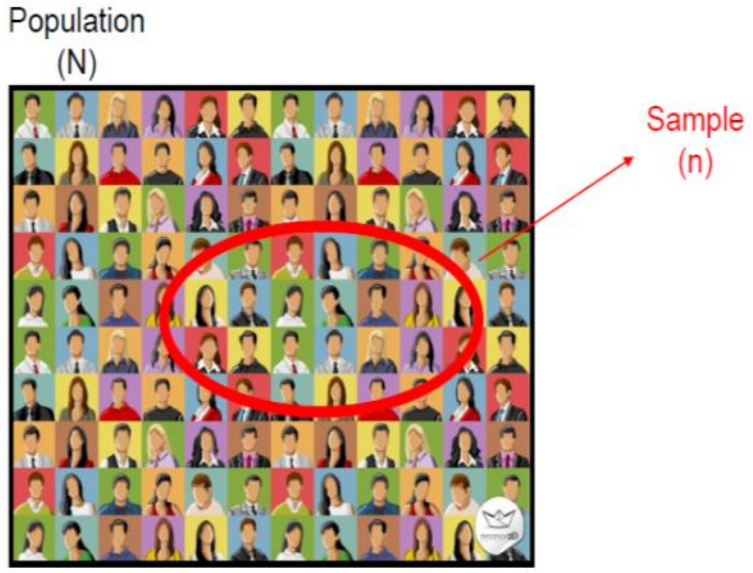
Tendencias a través del tiempo

Supervivencia: Probabilidad de que la vida útil de una batería falle antes de un año

DEFINICIONES

Población: conjunto de elementos sobre los que queremos hacer afirmaciones

Muestra: subconjunto de la población que se extrae para ser estudiado



DEFINICIONES

Parámetro: Valor descriptivo de la población

Estadístico: Valor descriptivo para una muestra

DEFINICIONES

Dato: es una observación sobre la variable medida

Datos: es una colección de observaciones

ESTADÍSTICA

DESCRIPTIVA INFERENCIAL

ESTADÍSTICA DESCRIPTIVA

Métodos para organizar y resumir los datos

ESTADÍSTICA INFERENCIAL

Métodos que sirven para obtener conclusiones de la población a partir de una muestra, esto ocurre cuando es imposible censar toda la población.

La meta de la estadística es ayudar a los investigadores a organizar e interpretar los datos

"Data don't make any sense,
we will have to resort to statistics."

¿QUÉ ES UN MUESTREO?

Selección de un conjunto de personas o cosas que se consideran representativos del grupo al que pertenecen, con la finalidad de estudiar o determinar las características del grupo.

"para hacer una buena encuesta se necesita antes hacer un buen muestreo"

¡¿Por qué es necesario muestrear?

- Medir todas las unidades es prácticamente imposible
- Muestrear unas pocas unidades ahorra dinero y tiempo
- Algunas medidas son destructivas

Variable

Es una característica o condición que puede tomar diferentes valores en una muestra. Ejm:

Presión sanguínea

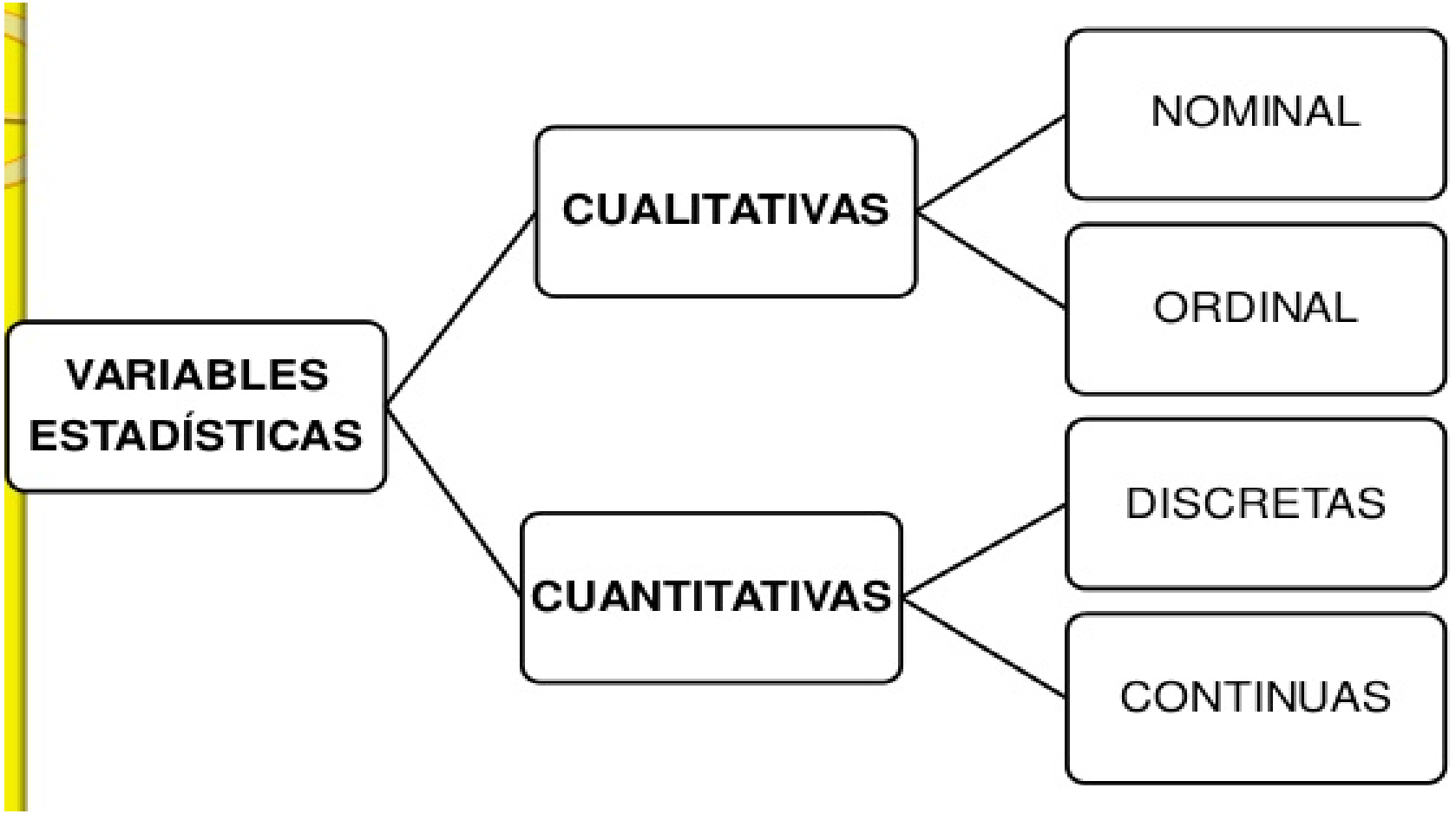
Masa de los niños

Frecuencia cardiaca

Estatura del grupo

Edad de los pacientes de un medico

TIPOS DE VARIABLES



VARIABLES CUALITATIVAS

Aquellas variables estadísticas que clasifican el conjunto de elementos de la muestra o población en categorías. Ejm:

Estado civil

Nacionalidad

Nivel educativo

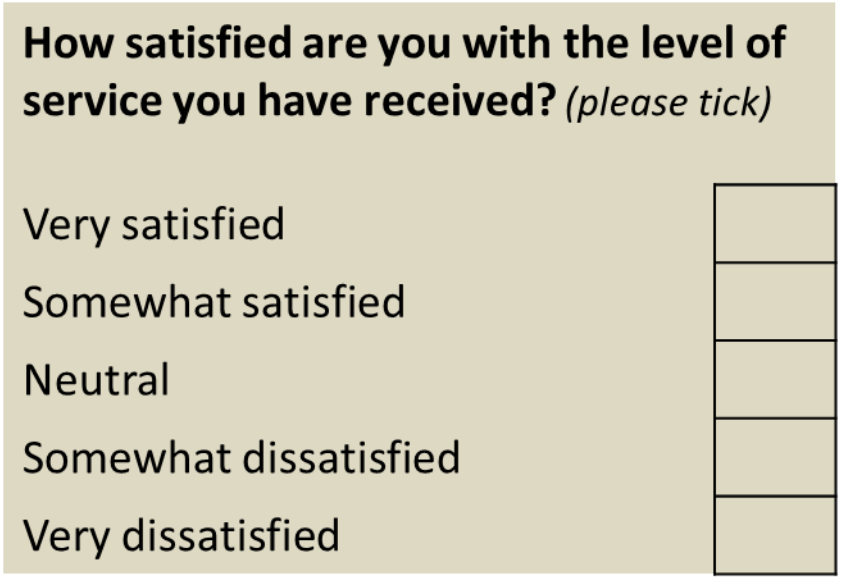
What is your gender?
(please tick)

Male ☐

Female ☐

Variable cualitativa ordinal

Variable que comprende un orden



Variable cualitativa nominal

En esta variable los valores no pueden ser sometidos a un criterio de orden, como por ejemplo los colores, el genero.

Variable cuantitativa

Son las variables que toman como argumento cantidades numéricas, son variables matemáticas. Ejm:

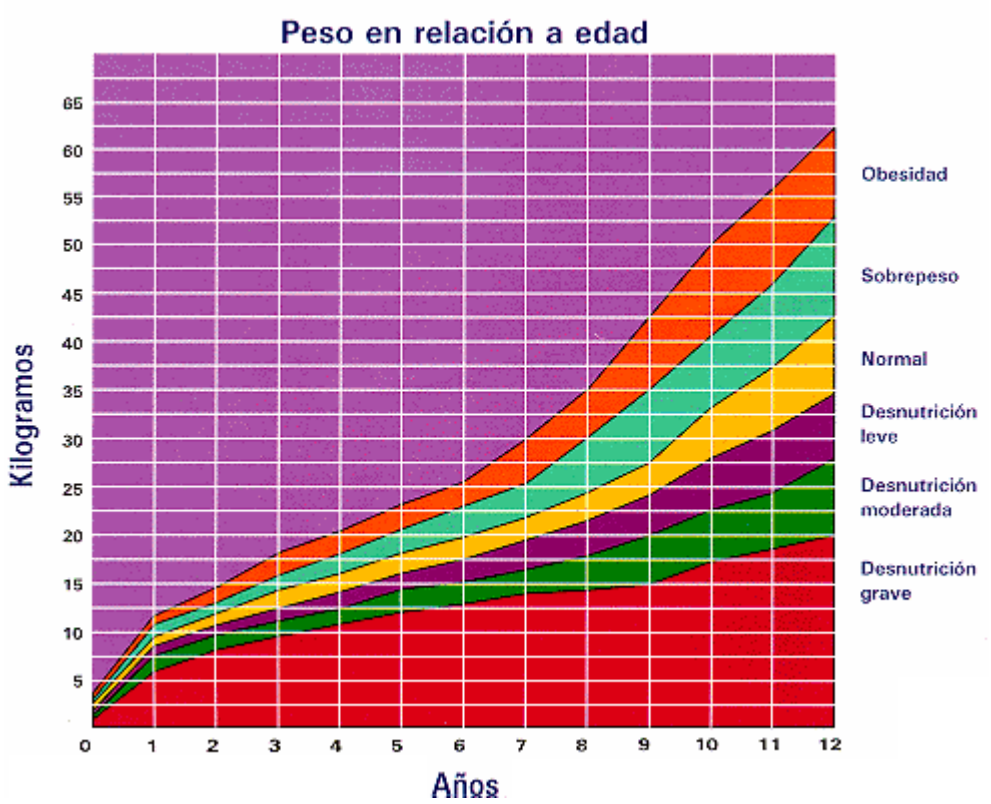
Peso, estatura, longitud

Variable cuantitativa continua

Puede adquirir cualquier valor dentro de un intervalo especificado de valores. Ejm:

la masa (2,3 kg, 2,4 kg, 2,5 kg,...)

la altura (1,64 m, 1,65 m, 1,66 m,...)



Variable cuantitativa discreta

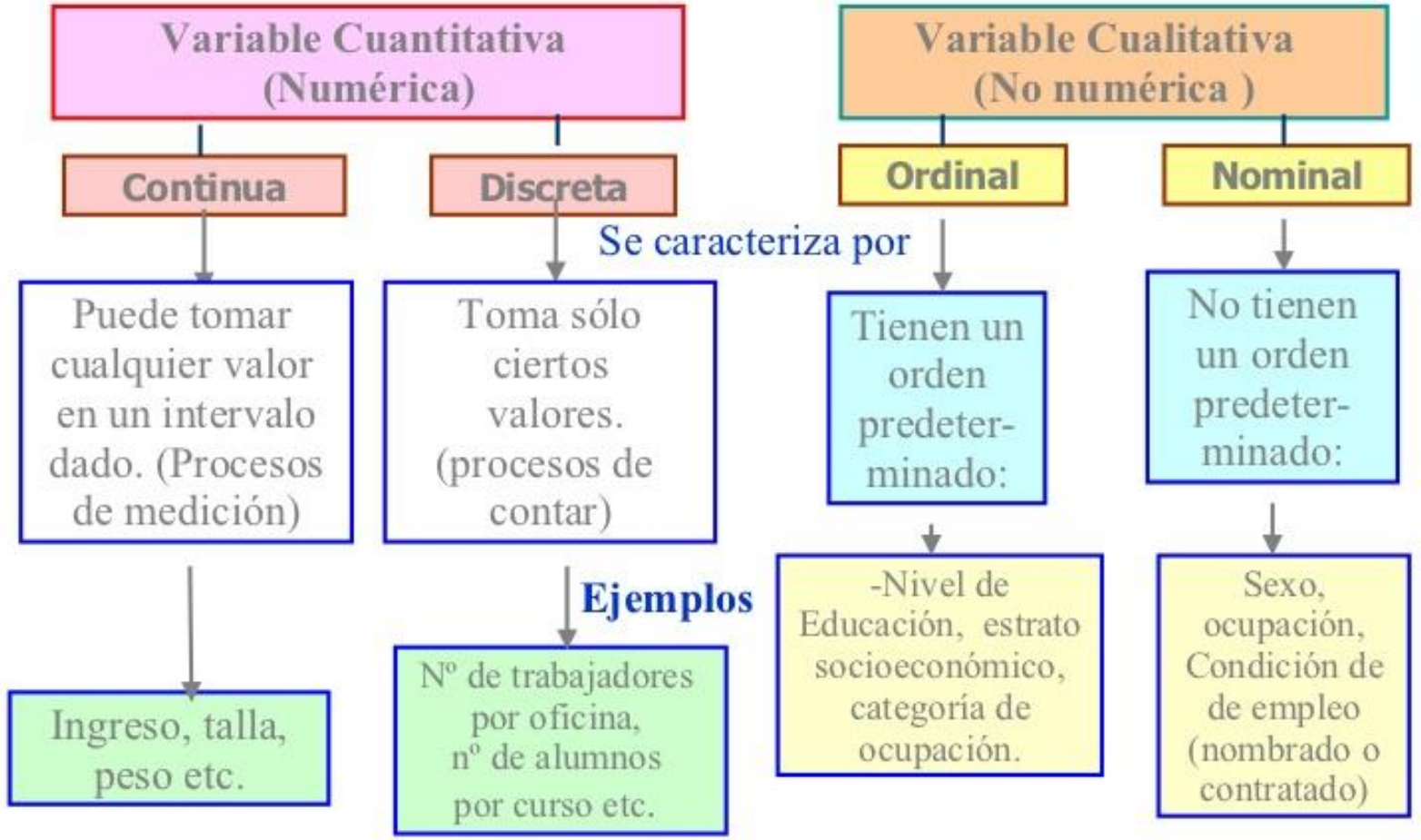
Es aquella variable cuantitativa que solo está expresada por números enteros.

Ejemplo:

El número de hijos (1, 2, 3, 4, 5)

Numero de estudiantes por curso

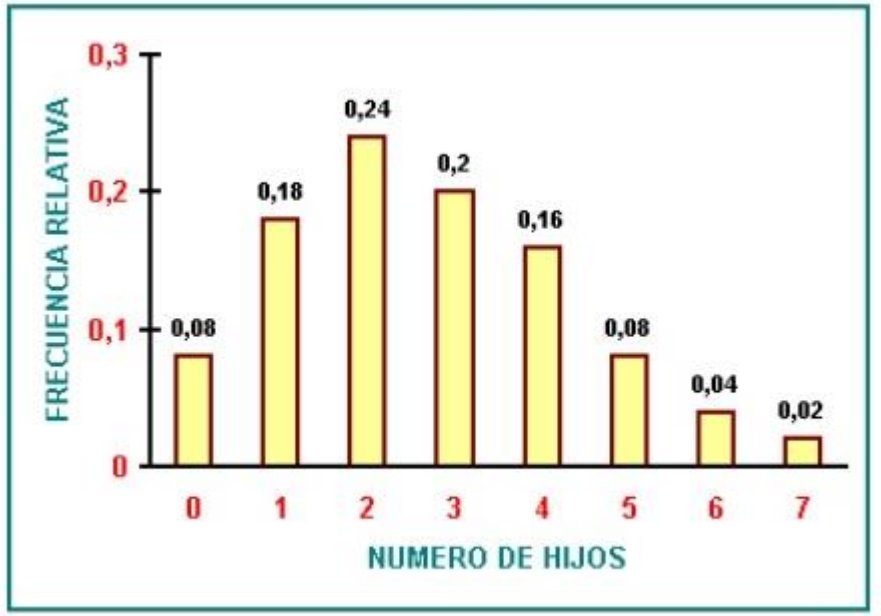
Clasificación de Variables



Herramientas estadísticas

FRECUENCIA ESTADÍSTICA

En estadística, la frecuencia de un evento i, es el número de veces en que dicho evento se repite durante un experimento. Comúnmente, la distribución de la frecuencia suele visualizarse con el uso de histogramas.



Tablas

- Tablas de frecuencia absoluta (número)
- Tablas de frecuencia relativa (porcentaje)

Tablas de frecuencia para variables cuantitativas discretas

Numero de hijos	Frecuencia absoluta	Frecuencia relativa
0	2	0,02
1	23	0,25
2	41	0,44
3	18	0,19
5	8	0,09
7	1	0,01
Total	**93**	**1**

→ = 2/93

→ = 23/93

Tablas de frecuencia para variables cuantitativas continuas

Edad	Frecuencia Absoluta	Frecuencia Relativa
10-14	2	0,050
15-19	16	0,400
20-24	18	0,450
25-29	3	0,075
30-34	1	0,025
Total	**40**	**1**

Tablas de resumen de datos
Tabla de frecuencia absoluta

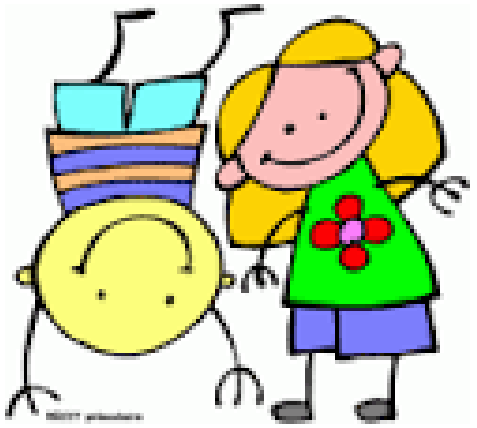
Género\Hobby	Bailar	Deporte	Televisión	**Total**
Hombre	2	10	8	**20**
Mujer	16	6	8	**30**
Total	**18**	**16**	**16**	**50**

Tabla de frecuencia relativa

Genero\ hobby	Bailar	Deporte	TV	**Total**
Hombre	0,04	0,2	0,16	**0,4**
Mujer	0,32	0,12	0,16	**0,6**
Total	**0,36**	**0,32**	**0,32**	**1**

Tablas de frecuencia relativa por fila

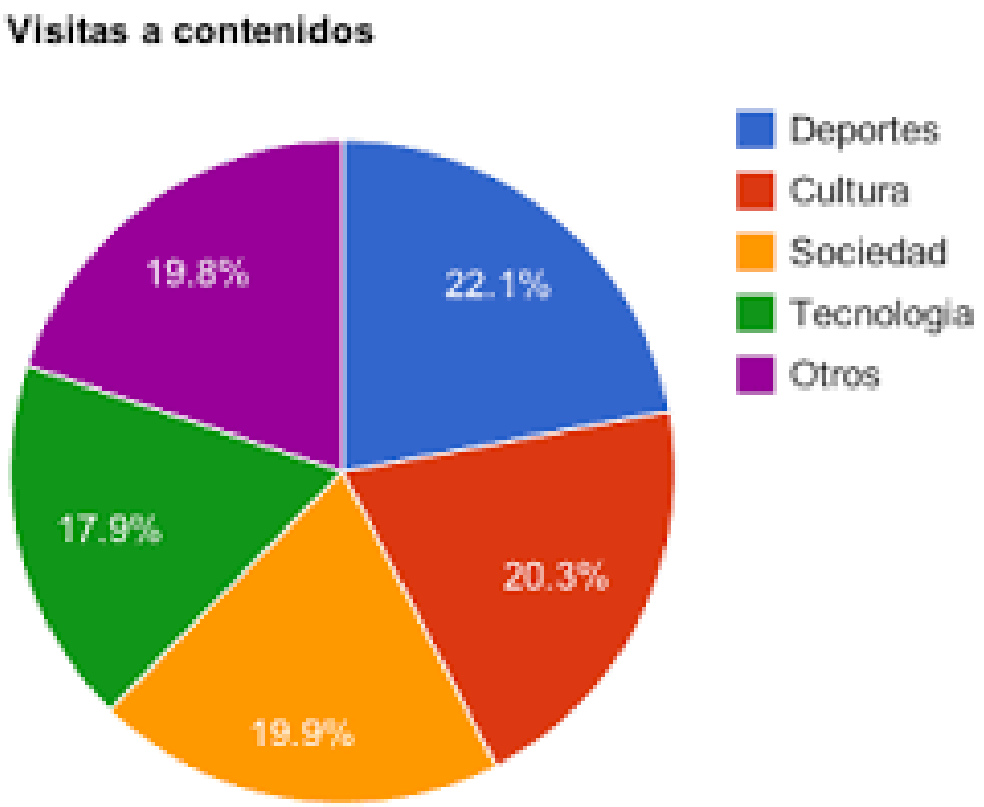
Genero\ hobby	Bailar	Deporte	Tv	**Total**
Hombre	0,10	0,50	0,40	**1**
Mujer	0,53	0,20	0,27	**1**



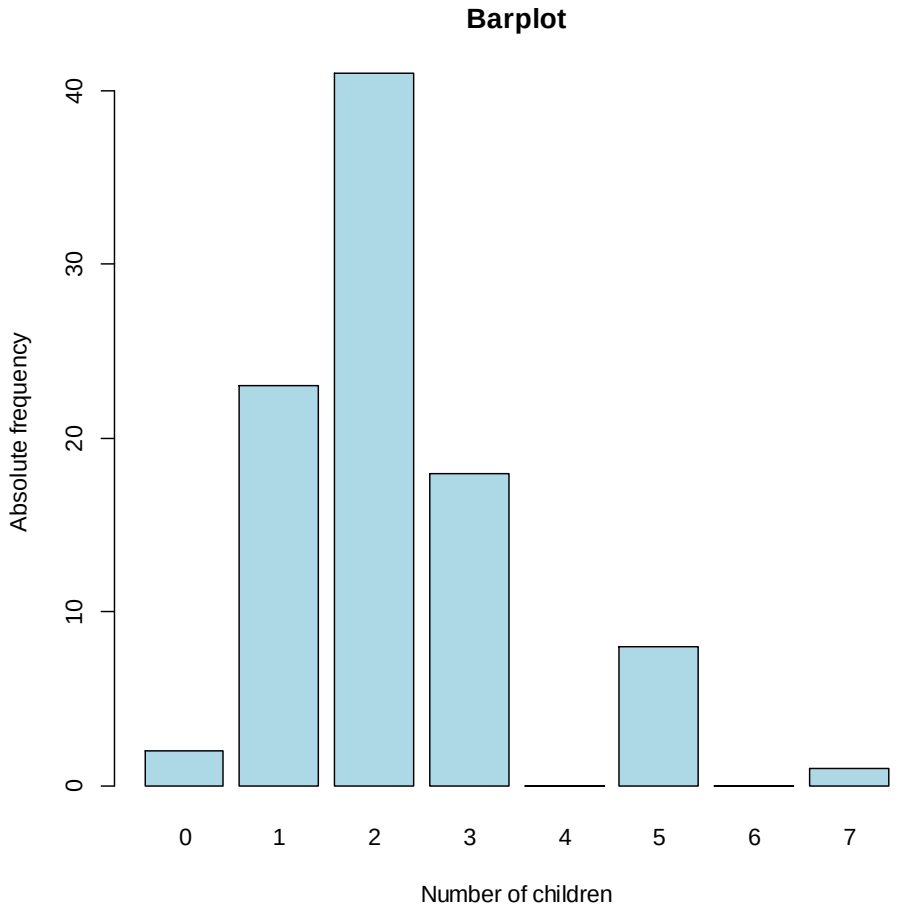
Tablas de frecuencia relativa por columna

Genero\ Hobby	Bailar	Deporte	TV
Hombre	0,11	0,63	0,50
Mujer	0,89	0,38	0,50
Total	**1**	**1**	**1**

GRAFICOS DE TORTAS



GRAFICOS DE BARRAS



HISTOGRAMA

Es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. Sirven para obtener una "primera vista" general de la distribución de la muestra, respecto a una característica.

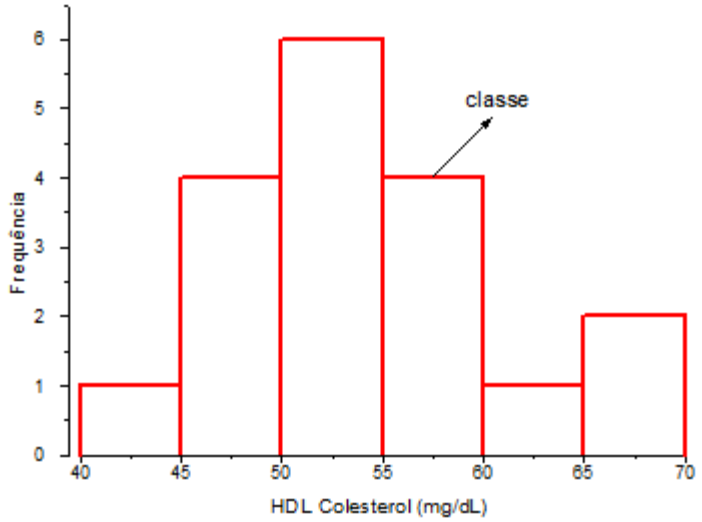
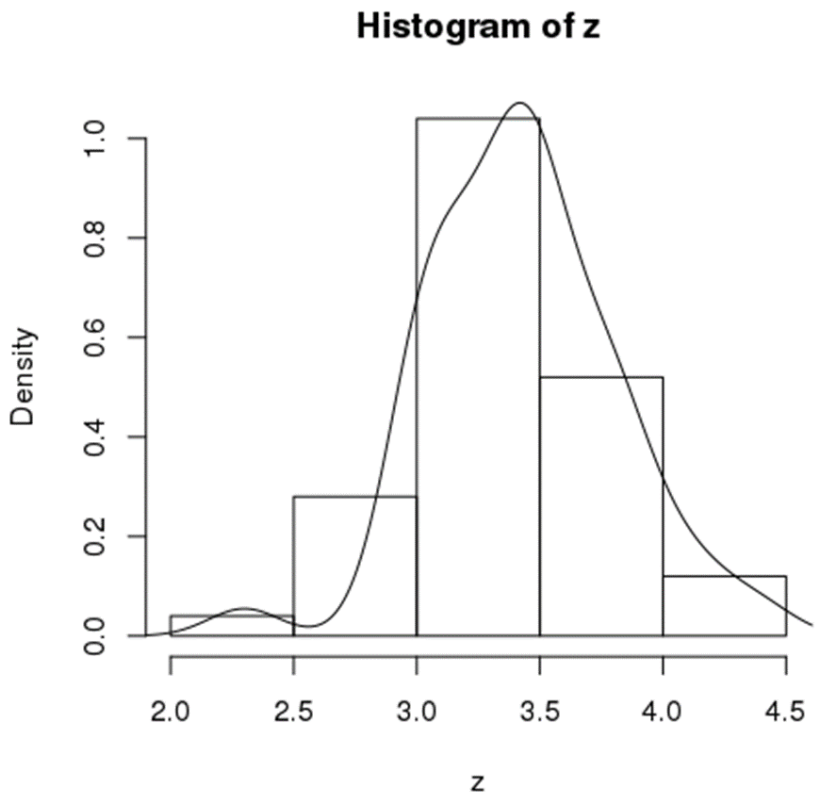
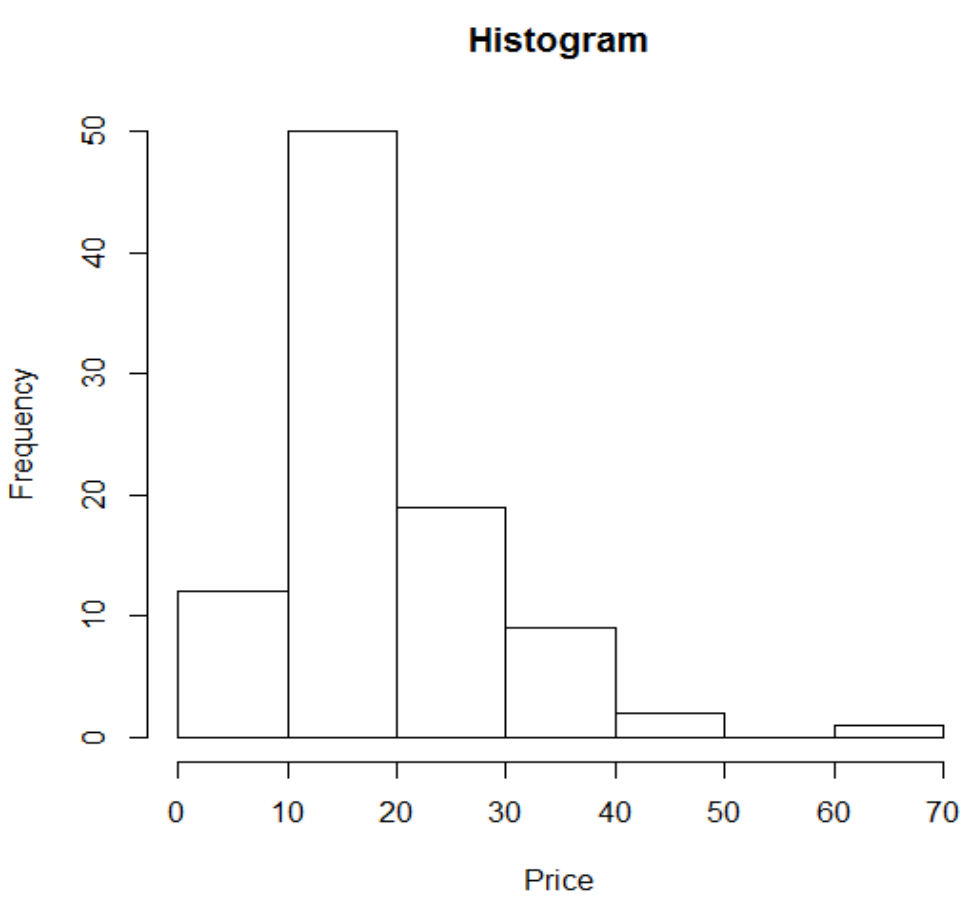


Grafico de densidad

Es un gráfico idéntico al histograma pero aplicado a distribuciones teóricas. El concepto de frecuencia relativa se cambia por el de probabilidad, pero también se representa por superficies y la suma de todas esas superficies (de todas las barras) será 1, como en el histograma.



HISTOGRAMAS Y GRAFICOS DE DENSIDAD



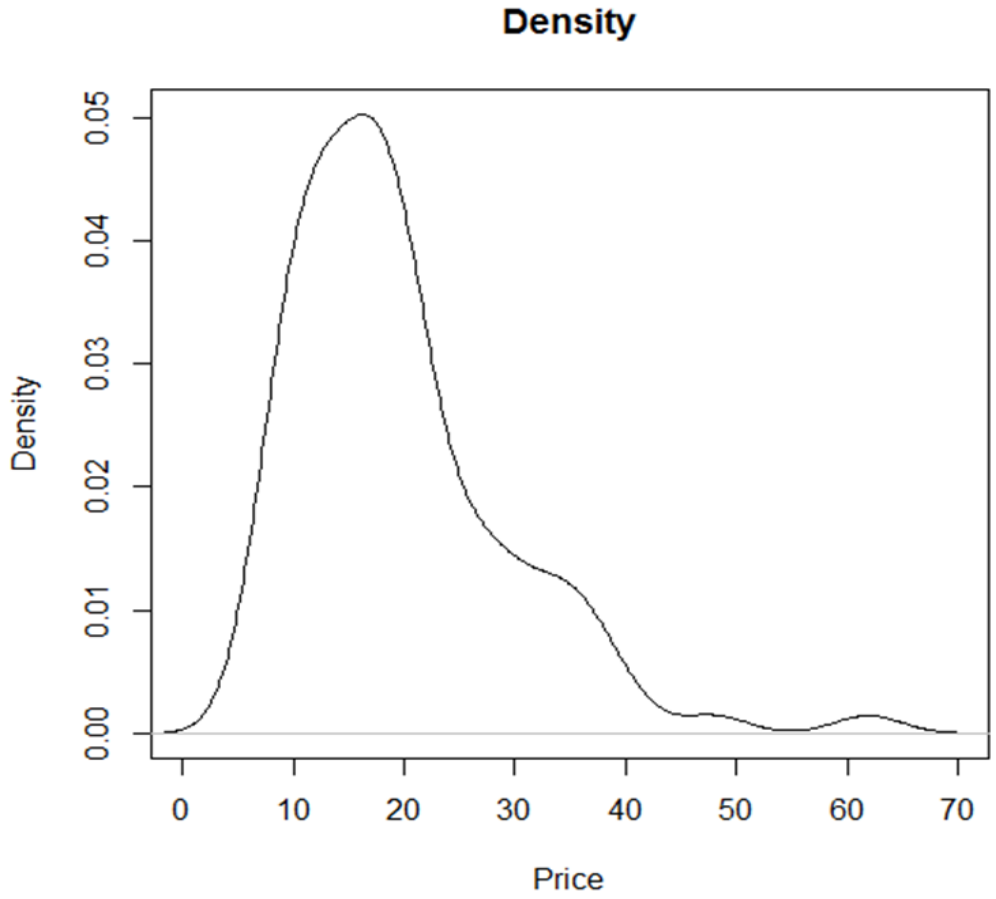
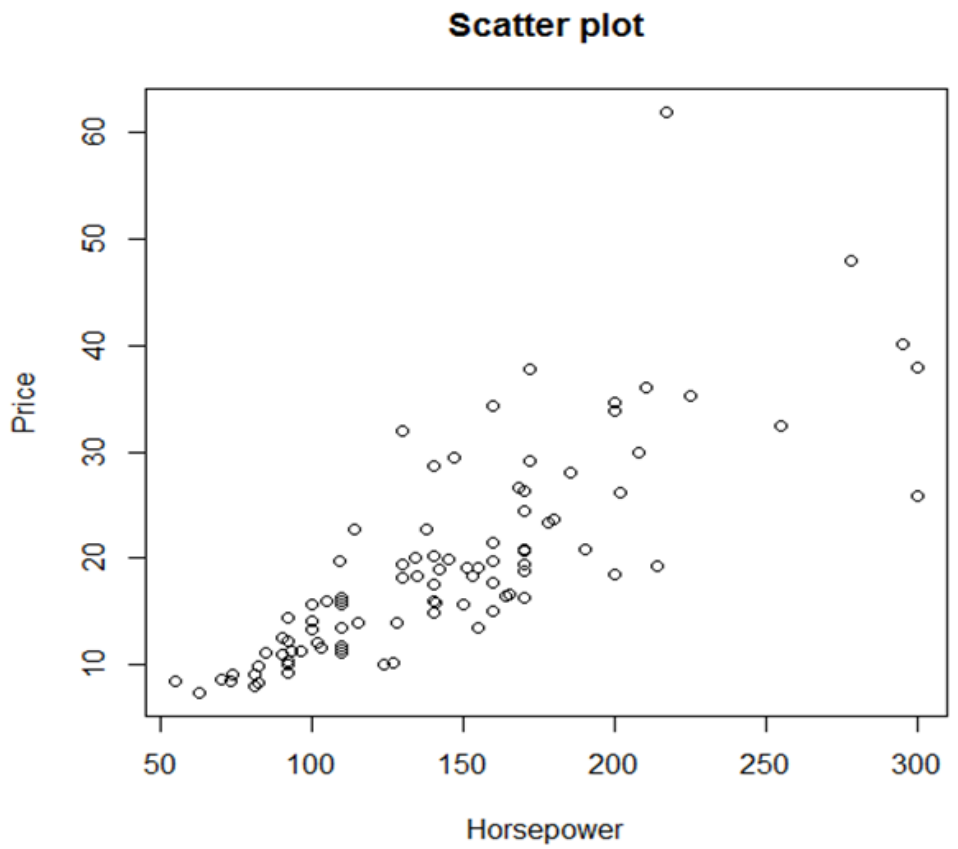


GRAFICO DE DISPERSIÓN



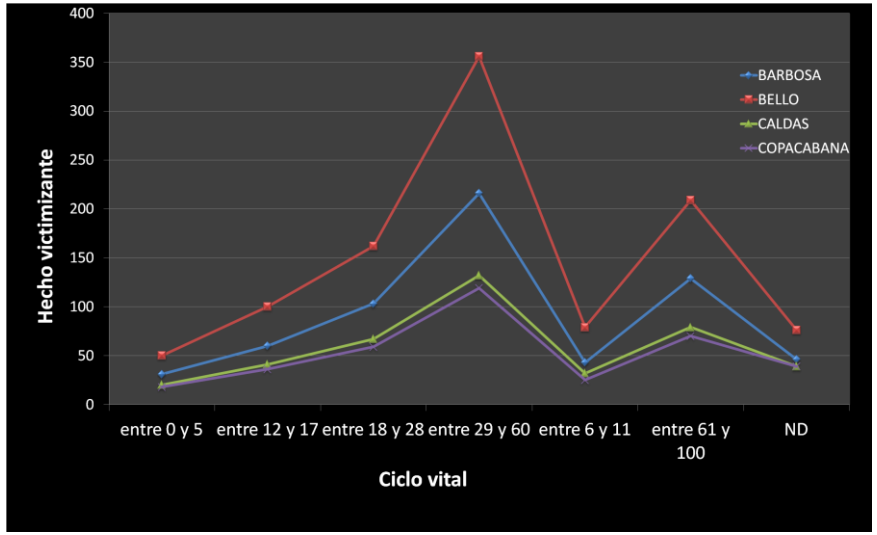
GRAFICOS DE SERIES DE TIEMPO

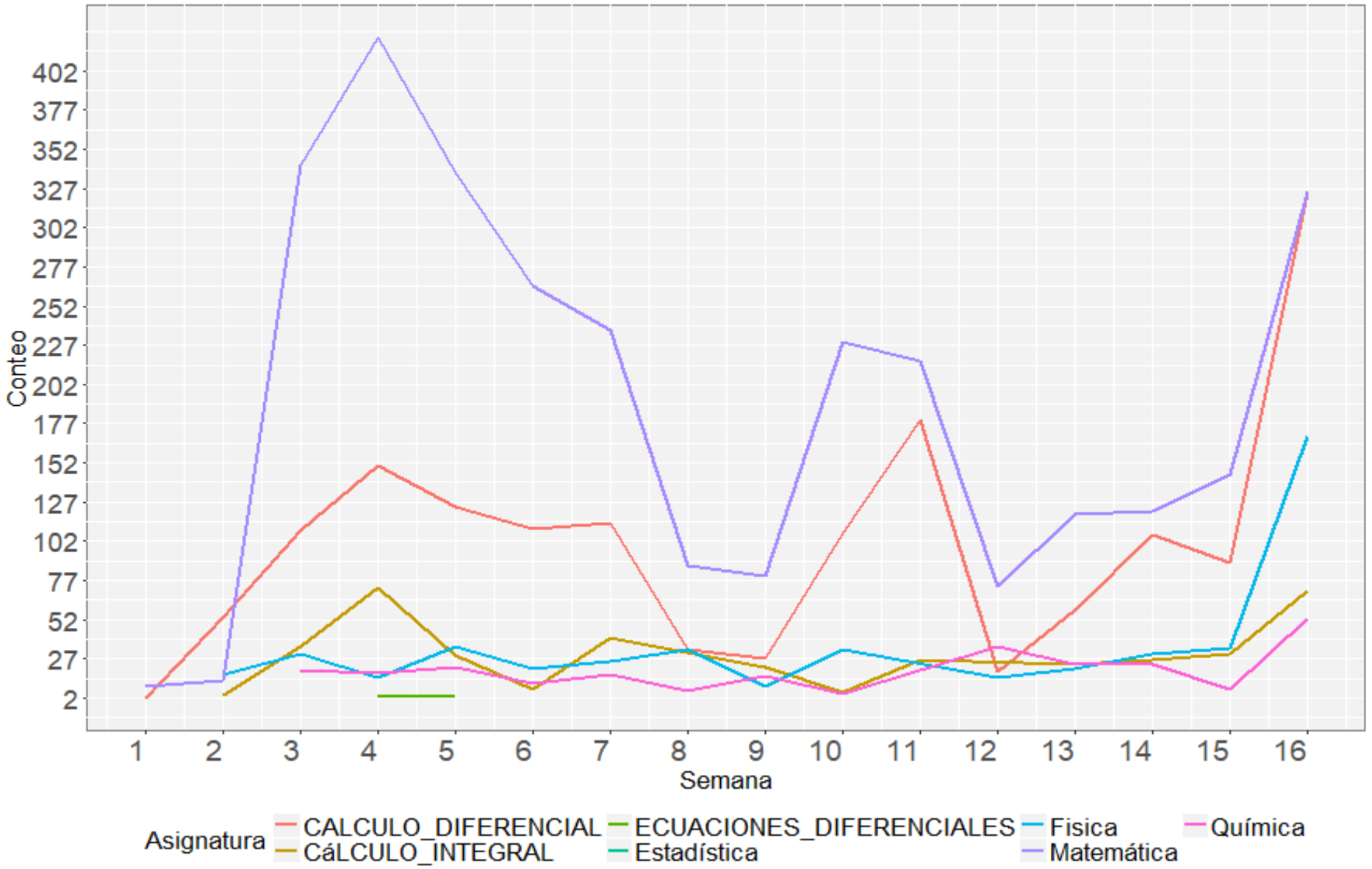
Únicamente para variables numéricas

GRAFICOS DE PERFILES

Gráfico de líneas en el que cada punto indica la media de una variable dependiente.

Sirve para detectar la variabilidad existente dentro de cada sujeto o nivel del factor y entre sujetos




Conteo
402
377
352
327
302
277
252
227
202
177
152
127
102
77
52
27
2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Semana
Asignatura
CALCULO_DIFERENCIAL
CáLCULO_INTEGRAL
ECUACIONES_DIFERENCIALES
Estadística
Fisica
Matemática
Química

EJERCICIO

Durante el mes de julio, en una ciudad se han registrado las siguientes temperaturas máximas:

32, 31, 28, 29, 33, 32, 31, 30, 31, 31, 27, 28, 29, 30, 32, 31, 31, 30, 30, 29, 29, 30, 30, 31, 30, 31, 34, 33, 33, 29, 29

Construya una tabla de frecuencias absoluta y relativa

Construya el histograma

MEDIDAS DE TENDENCIA CENTRAL

Es un valor que resume y representa la información contenida en un conjunto de datos.

Las tres medidas más usadas son la media la mediana y la moda

La media o promedio (μ ó $\bar{x}$)

Medida de tendencia central

Media Poblacional

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

Media muestal

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{\mathrm{n}}$$

Mediana

Es un valor que se encuentra en la mitad de los datos, cuando estos están ordenados

a. si los datos son pares, la mediana es la suma de los dos valores centrales dividida por dos

$$m_e = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}\right)+1}}{2}$$

b. si los datos son impares la mediana es el valor que se encuentra en la posición $\frac{(n+1)}{2}$

$$m_e = x_{\frac{(n+1)}{2}}$$

Ejemplo

Las edades de una muestra aleatoria de 10 estudiantes del programa diurno y nocturno del pregrado en administración de empresas son:

Diurno	**24**	**30**	**28**	**23**	**25**	**22**	**26**	**27**	**25**	**28**
Nocturno	26	33	29	28	27	29	33	34	28	27

Calcule la media y la mediana para cada uno de los grupos

Moda

Es el valor que más se repite

si no hay datos que se repiten se dice que no hay moda.

Si dos datos se repiten con la misma frecuencia se dice que los datos son bimodales.

Ejemplo

Durante el mes de julio, en una ciudad se han registrado las siguientes temperaturas máximas:

32, 31, 28, 29, 33, 32, 31, 30, 31, 31, 27, 28, 29, 30, 32, 31, 31, 30, 30, 29, 29, 30, 30, 31, 30, 31, 34, 33, 33, 29, 29

Calcule la media, la mediana y la moda

MEDIA PONDERADA

La media ponderada es la media de un conjunto de datos cuyas entradas tienen diferentes pesos, está dada por

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

Donde w es el peso de cada entrada x.

EJEMPLO

Las siguientes son las notas obtenidas por los estudiantes de matemática básica, con sus respectivos porcentajes. ¿Cual es la nota promedio final?

EVENTOS EVALUATIVOS	**1**	**2**	**3**	**4**
NOTA	4	3	2	1
VALOR (%)	35	25	20	20

MEDIDAS DE VARIACIÓN

La media es un buen indicador de tendencia central, pero no da una evidencia real acerca de los datos.

						Mean	Median
Data set A:	5	6	7	8	9	7	7
Data set B:	1	2	7	12	13	7	7

RANGO

El rango de un conjunto de datos es la diferencia entre el valor máximo y el valor mínimo

$$Rango = Valor\ mayor - Valor\ menor$$

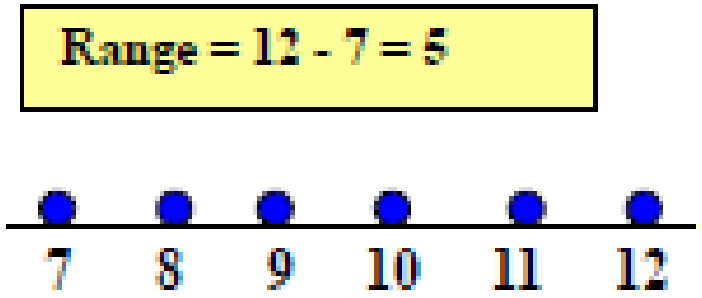
Ejemplo: los siguientes son los datos de los precios de un inventario. Encuentre el rango

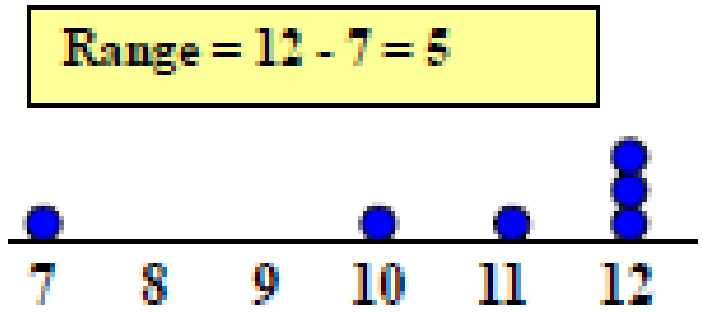
Stock	56	56	57	58	61	63	63	67	67	67

The range is 67 – 56 = 11.

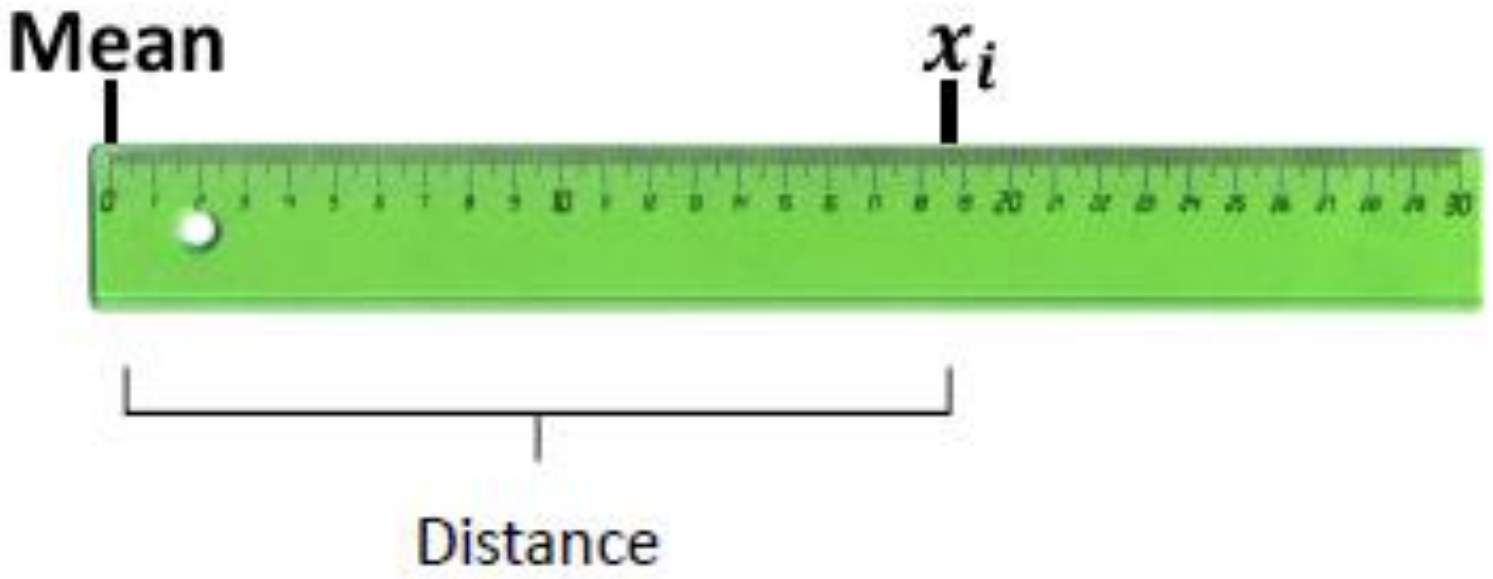
RANGO

Ignora la manera en que los datos están distribuidos





Varianza poblacional y desviación estándar



Varianza y desviación estándar poblacional

La varianza poblacional de un conjunto de datos está dada por:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

La desviación estándar poblacional de un conjunto de datos está dada por:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

En el caso muestral

La varianza muestral está dada por:

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

y la desviación estándar muestral está dada por:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Ejemplo

Un grupo de matemáticas obtuvo las siguientes notas en un examen

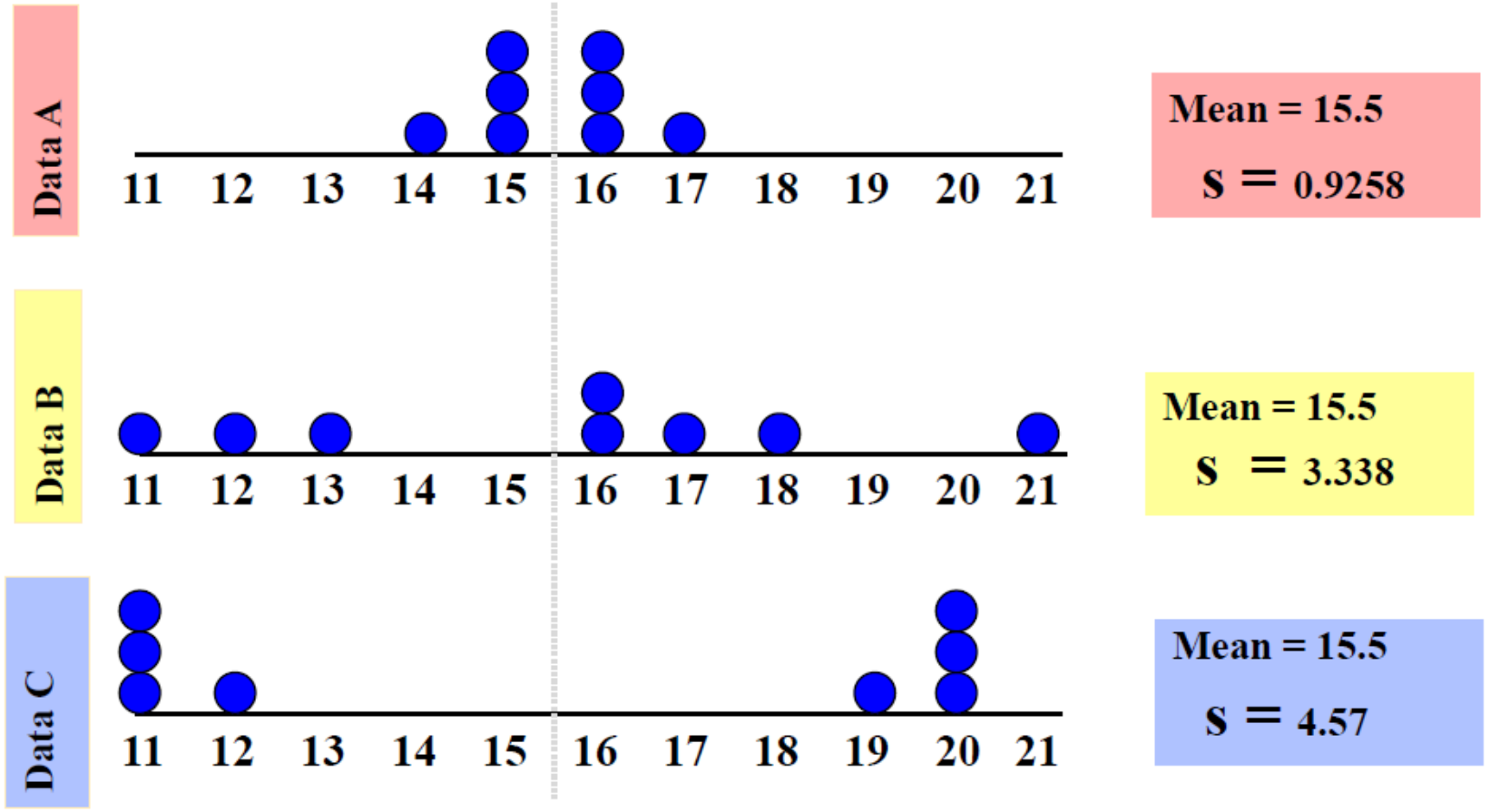
92 95 83 76 54

Encuentre la varianza y la desviación estándar poblacional

PASOS

1. Encuentre la media
2. Encuentre la desviación de cada uno de los valores con respecto a la media
3. Eleve cada valor obtenido al cuadrado
4. Halle la suma de cada uno de los cuadrados
5. Divida la suma de los cuadrados por el número de ítems
6. Encuentre la raíz cuadrada de la varianza

Comparemos…


Data A
11 12 13 14 15 16 17 18 19 20 21
Mean = 15.5
S = 0.9258
Data B
11 12 13 14 15 16 17 18 19 20 21
Mean = 15.5
S = 3.338
Data C
11 12 13 14 15 16 17 18 19 20 21
Mean = 15.5
S = 4.57

¿Cómo obtener la media y la varianza en la calculadora?

https://www.youtube.com/watch?v=qguhqq0xvM0&t=34s

Diferencia en medidas entre la población y la muestra

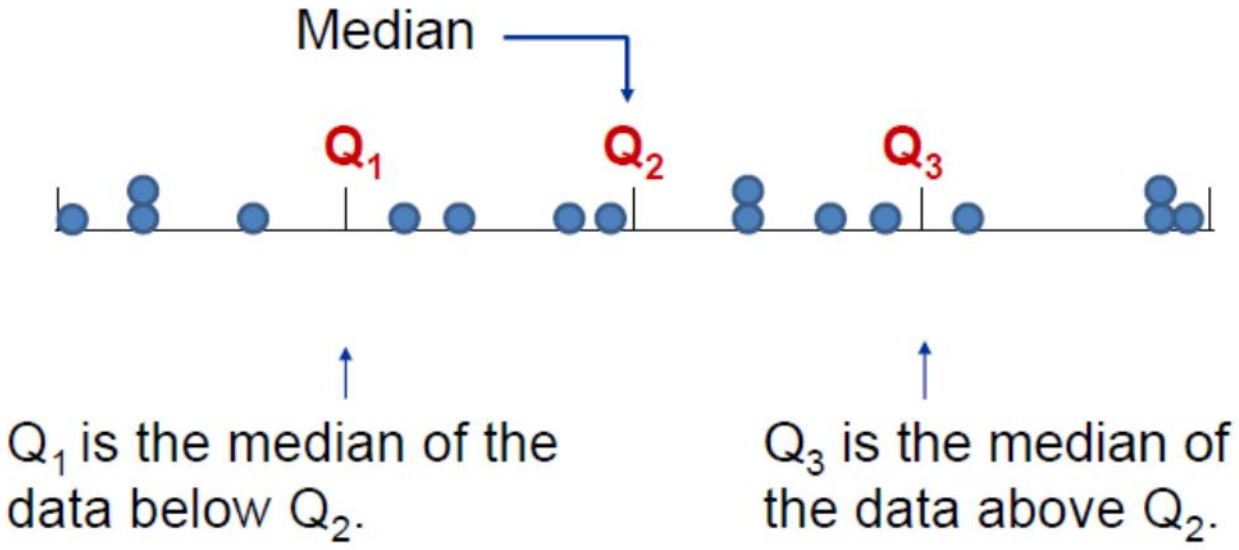
Medida	Población	Muestra
Tamaño	N	n
Media	μ	$\bar{x}$
Varianza	σ^2	S^2
Desviación estándar	σ	S

MEDIDAS DE POSICIÓN

Las medidas de posición más utilizadas son los cuartiles, deciles y percentiles.

Cuartiles

son tres valores que distribuyen la serie de datos ordenada, en cuatro tramos iguales, en los que cada uno de ellos se concentra el 25% de los resultados.



Cómo estimar los cuartiles

$$Q_k = x_{\left(k\frac{n+1}{4}\right)}$$

with k = 1, 2, 3

Los puntajes en una prueba se listan a continuación

28 43 48 51 43 30 55 44 48 33 45 37 37 42 38

Order the data.

Lower half | Upper half

28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

$Q_1 = X_{(4)}$ $Q_2 = X_{(8)}$ $Q_3 = X_{(12)}$

DECILES

Son 9 valores que distribuyen la serie de datos ordenada, en diez tramos iguales, en los que cada uno de ellos concentra el 10% de los resultados.

$$D_k = x_{\left(k\,\frac{n+1}{10}\right)}$$

with k = 1, 2, ..., 9

PERCENTILES

Divide un conjunto de datos ordenados en 100 partes iguales, es decir hay 99 percentiles

$$P_k = x_{\left(k\,\frac{n+1}{100}\right)}$$

with k = 1, 2, ..., 99

BOXPLOT

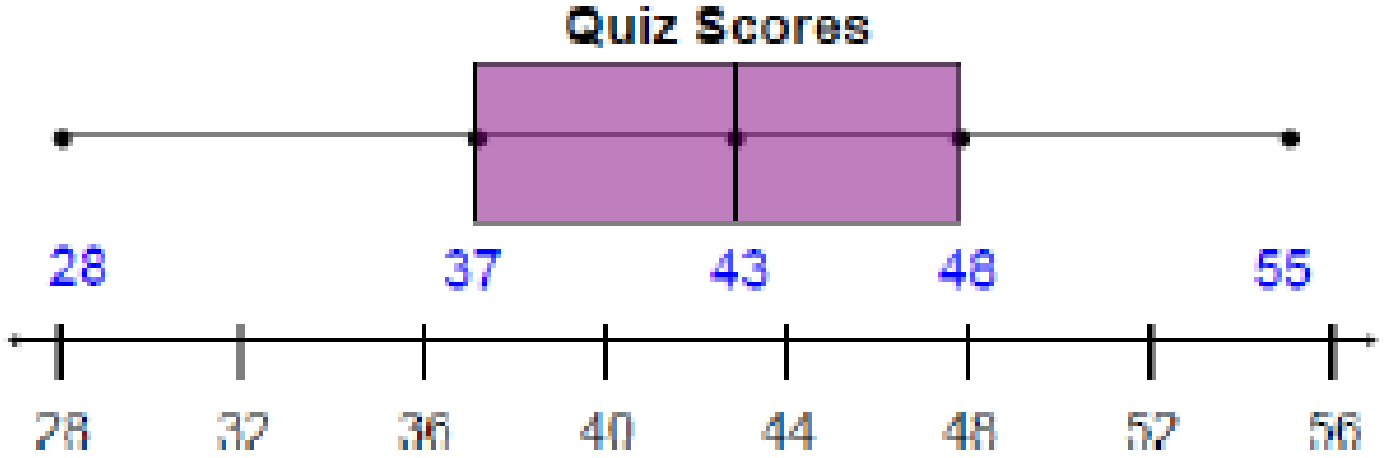
El boxplot es una herramienta de análisis que resalta las principales características de un conjunto de datos.

Los 5 números usados son:

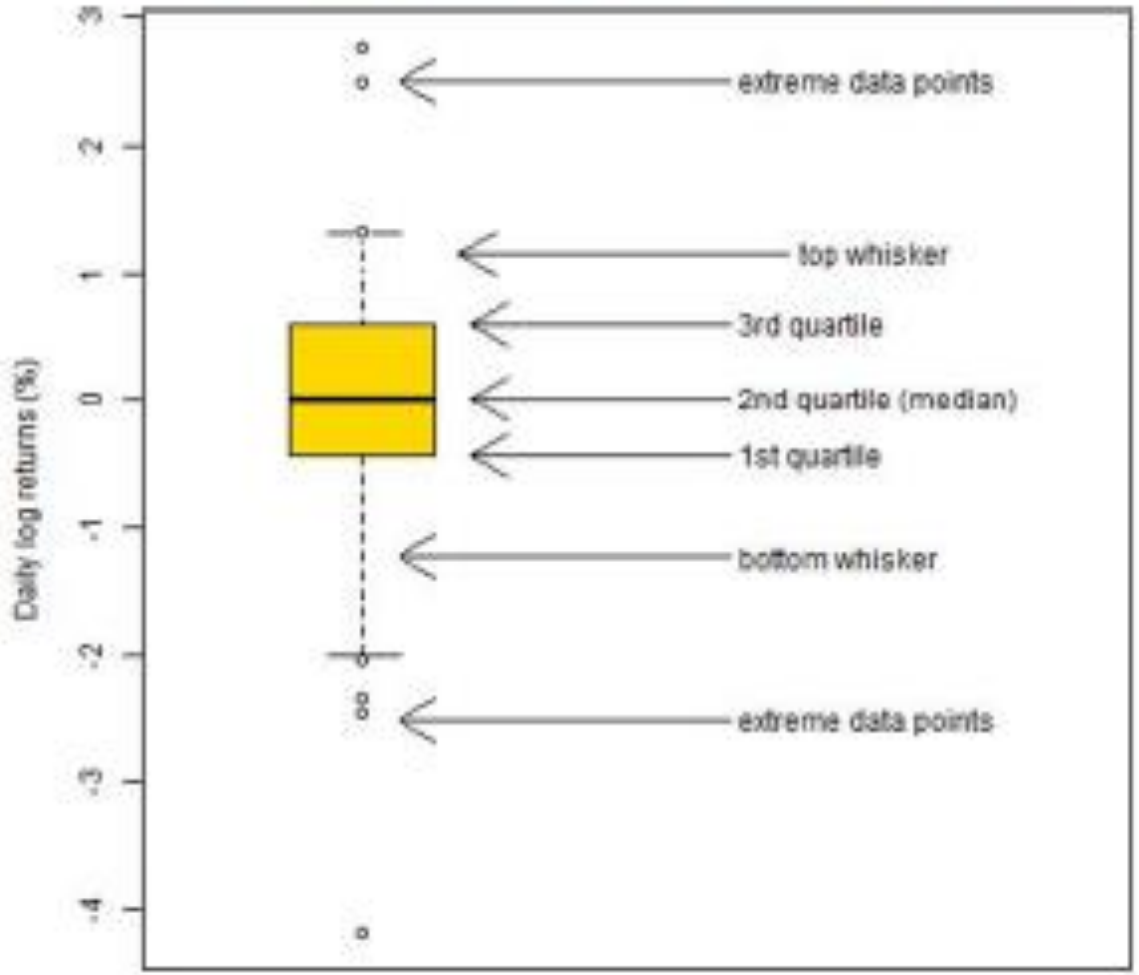
- Valor mínimo
- Q1
- Q2 (Mediana)
- Q3
- Valor máximo

Five-number summary

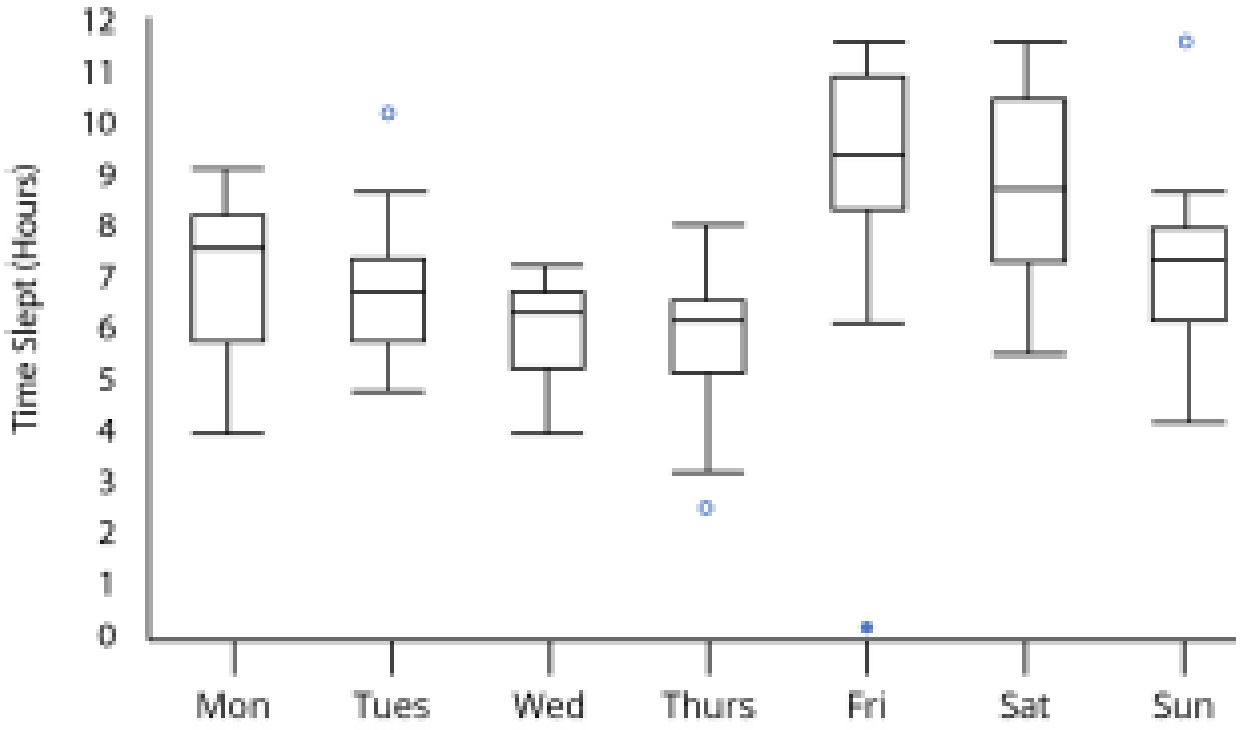
- The minimum entry 28
- Q_1 37
- Q_2 (median) 43
- Q_3 48
- The maximum entry 55



Partes de un boxplot



Ejemplo: Horas de sueño por días de la semana para un grupo de 20 estudiantes



Coeficiente de variación

Es una medida de la dispersión relativa de un conjunto de datos, la cual relaciona la desviación típica de una muestra y su media.

Se expresa en términos porcentuales.

$$CV = \frac{s}{\bar{X}}$$

No depende de las unidades de medición, por lo que sirve para comparar la variabilidad de dos conjuntos de datos, siempre que sus medias sean positivas.

EJEMPLO

El promedio de exportación semanal de flores de la corporación A fue de 4420 kilos con una desviación estándar de 615 k, en tanto que la corporación B fue de 4320K con una desviación estándar 620. En qué corporación hubo mayor variabilidad?.

Corporación	Media	Desviación estándar	Coeficiente de variación
A	4420	615	13.91
B	4320	620	14.34

MEDIDAS DE FORMA

Permiten identificar si una distribución de frecuencia presenta uniformidad.

Son necesarias para determinar el comportamiento de los datos.

Se clasifican en:

Medidas de asimetría y medidas de curtosis

Medidas de asimetría

Permite establecer el grado de simetría que tiene una distribución.

Coeficiente de asimetría de Pearson: relaciona la media y la moda de un conjunto de datos

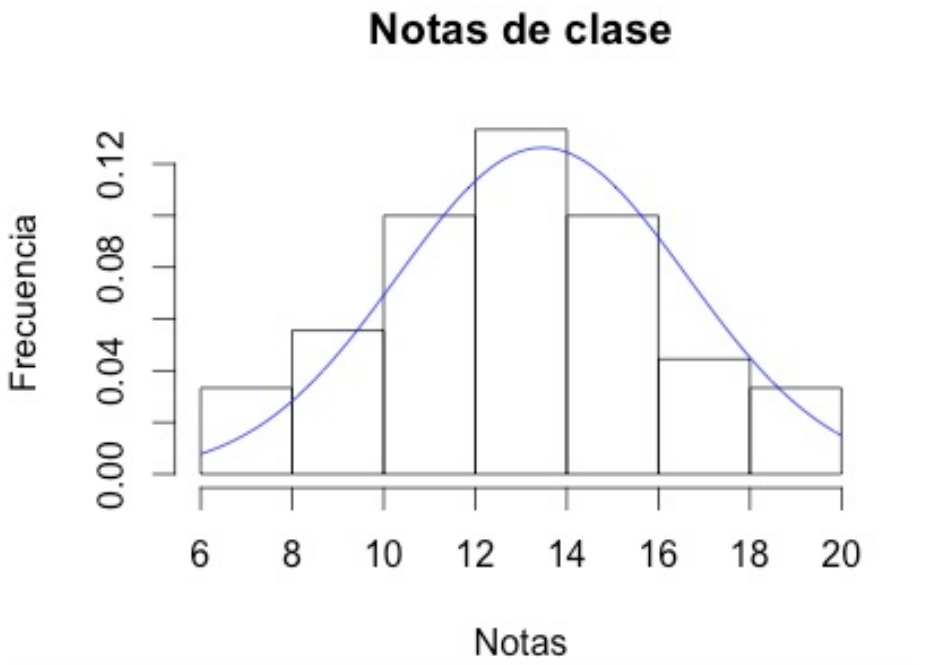
$$A_s = \frac{\bar{X} - M_o}{S}$$

Distribución simétrica

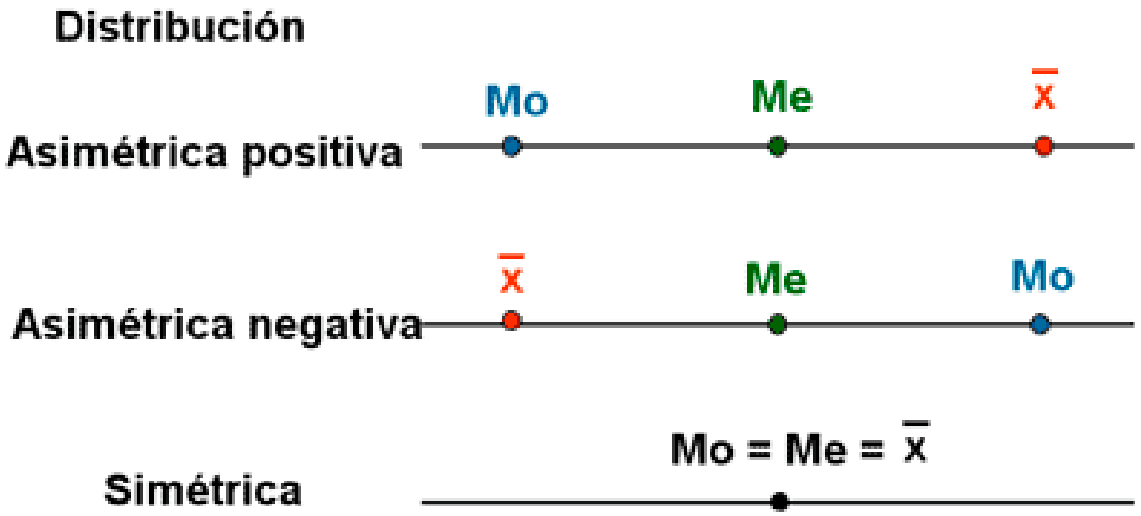
Una distribución es simétrica cuando la media la mediana y la moda coinciden es decir

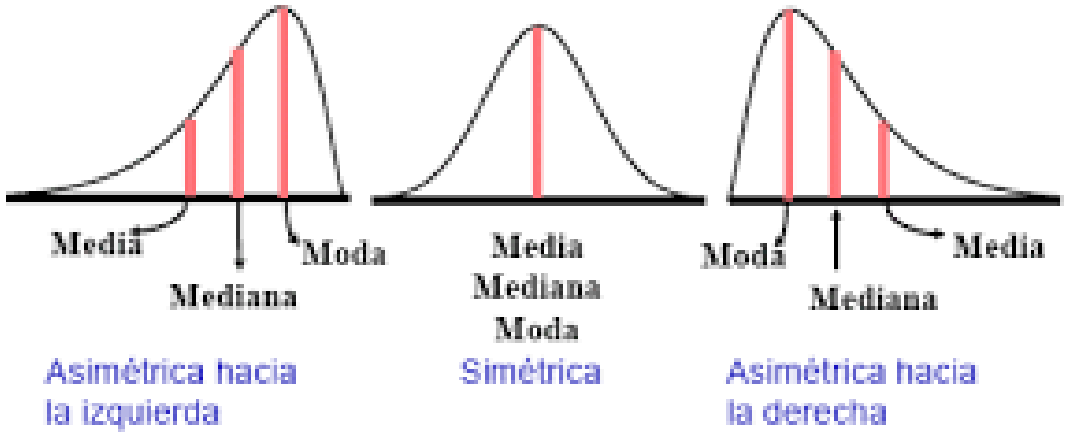
$$\bar{X} = M_o = M_e$$

Luego el coeficiente de pearson es 0



Distribución asimétrica

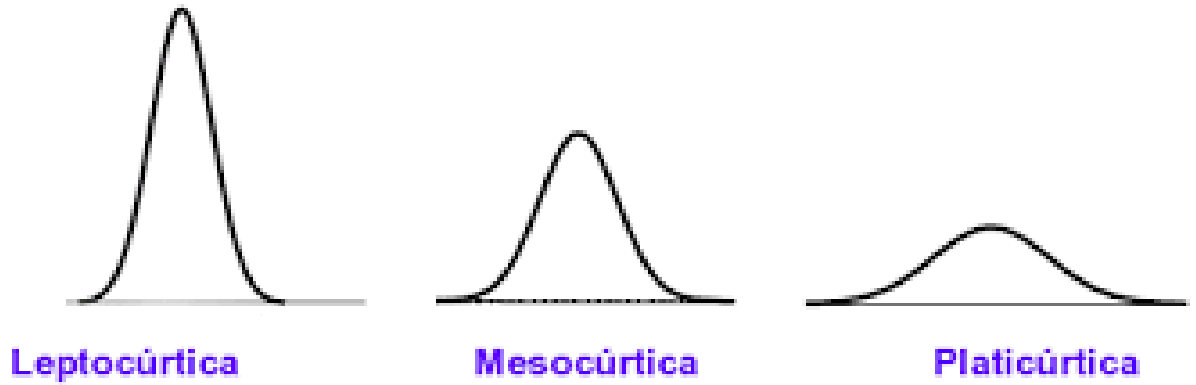




MEDIDAS DE CURTOSIS

Curtosis: Medida que sirve para analizar el grado de concentración que presentan los valores de una variable analizada alrededor de la zona central de la distribución de frecuencias

Se definen tres tipos de distribuciones según su grado de curtosis.



Tablas de frecuencia para datos agrupados

Cuando los valores de la variable (continuas o discretas) son muchos, conviene agrupar los datos en intervalos o clases para así realizar un mejor análisis e interpretación de ellos.

Proceso para la construcción de la tabla de frecuencia para datos agrupados

1. Se elige ó se calcula el número de intervalos de clase representado por la variable m se puede usar la fórmula:

$$m = 1 + 3.3 \log(n)$$

$$m = \sqrt{n}$$

2. Calcule el rango R

3. Calcule la amplitud de los intervalos, el cual se denota por la letra a:

$$a = \frac{R}{m}$$

4. Calcule los intervalos de clase: son dos columnas que delimitan el límite superior e inferior del intervalo (LI y LS)

El intervalo es cerrado a la izquierda y abierto a la derecha [x1,x2).

El primer y último intervalo es cerrado a ambos lados, para no dejar información fuera del rango

(Ll_1): Límite inferior del intervalo o clase

(LS_i): Límite superior del intervalo o clase

i	Intervalos de clase	
1	X_{min}	$LS_1 = LI_1 + a$
2	$LI_2 = LS_1$	$LS_2 = LI_2 + a$
⋮	⋮	⋮
n	$LI_n = LS_{n-1}$	

i = 1, 2, 3, ..., n

$(LS_i) = (Ll_i) + a$

5. se busca la marca de clase que se denota por X_i, dada por:

$$X_i = \frac{LI_i + LS_i}{2}$$

6. La segunda columna de la tabla, para datos agrupados es la frecuencia absoluta (f_i):

Es el número de observaciones que caen en un intervalo sin incluir el limite superior, es decir número de datos mayores o iguales a LI_i pero menores que LS_i.

Es decir el intervalo es cerrado a la izquierda y abierto a la derecha [)

El último intervalo es cerrado a ambos lados, para no dejar información fuera del rango

7. Se busca las otras columnas que corresponden a:

f_i: Frecuencia absoluta F_i: Frecuencia absoluta acumulada

$$F_1 = f_1$$

$$F_2 = f_1 + f_2$$

.

.

.

$$F_i = \sum_{i=1}^{n} F_i$$

h_i *:* Frecuencia relativa, se calcula dividiendo cada frecuencia absoluta f_i por el numero total de observaciones n, así:

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

H_i: Frecuencia relativa acumulada

$$H_1 = h_1$$

$$H_2 = h_1 + h_2$$

$$H_3 = h_1 + h_2 + h_3$$

$$H_i = \sum_{i=1}^{n} h_i$$

$h_i * (100)\%$

Intervalos Ll_i - LS_i	f_i	X_i	F_i	h_i	H_i	h_i*%
Ll_1 - LS_1	f_1	X_1	F_1	$h_1 = f_1 / n$	h_1	h_1 (100)
Ll_2 - LS_2	f_2	X_2	F_2	$h_2 = f_2 / n$	$h_2 + h_2$	h_2 (100)
⋮	⋮	⋮	⋮	⋮	⋮	
Ll_n - LS_n	f_n	X_n	F_n	$h_n = f_n / n$	$h_1 + h_2 + \ldots + h_n$	h_n (100)

Ejemplo

Calculo de la media para datos agrupados

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

Donde:

X_i es la marca de clase

f_i es la frecuencia relativa

n tamaño de la muestra

Como calcular la mediana para datos agrupados

1. calcular n/2
2. Encontrar la F_i en la tabla que contenga a n/2
3. Calcular la mediana con la fórmula

$$m_e = L_i + \frac{a\left(\frac{n}{2} - F_{i-1}\right)}{(F_i - F_{i-1})}$$

Donde:

L_i Límite inferior del intervalo i que corresponde a F_i

F_i Frecuencia absoluta acumulada del que corresponde al intervalo de la mediana

F_{i-1} Frecuencia absoluta acumulada que corresponde al intervalo anterior de la mediana

a amplitud del intervalo

Moda para datos agrupados

Es el valor que representa la **mayor frecuencia absoluta**. En tablas de frecuencias con datos agrupados, hablaremos de intervalo modal. La moda se representa por **Mo**.

$$m_o = L_i + \frac{a \quad (f_i - F_{i-1})}{(f_i - f_{i-1}) + (f_i - f_{i+1})}$$

Donde:

L_i Límite inferior del intervalo modal

f_i Frecuencia absoluta del intervalo modal

f_{i-1} Frecuencia absoluta del intervalo anterior al modal

f_{i+1} Frecuencia absoluta del intervalo posterior al modal