Estadística
Algunos Conceptos

Introducción

¿Qué es la estadística?

La estadística, en general, es la ciencia que trata de la recopilación, organización presentación, análisis e interpretación de datos numéricos con e fin de realizar una toma de decisión más efectiva.

¿En qué áreas se aplica la estadística?

Actualmente se aplica en todas las áreas del saber, por ejemplo en *Sociología*, *Educación*, *Psicología*, *Administración*, *Economía*, *Medicina*, *Ciencias Políticas*, entre otras.

Ejemplos de su aplicación son:

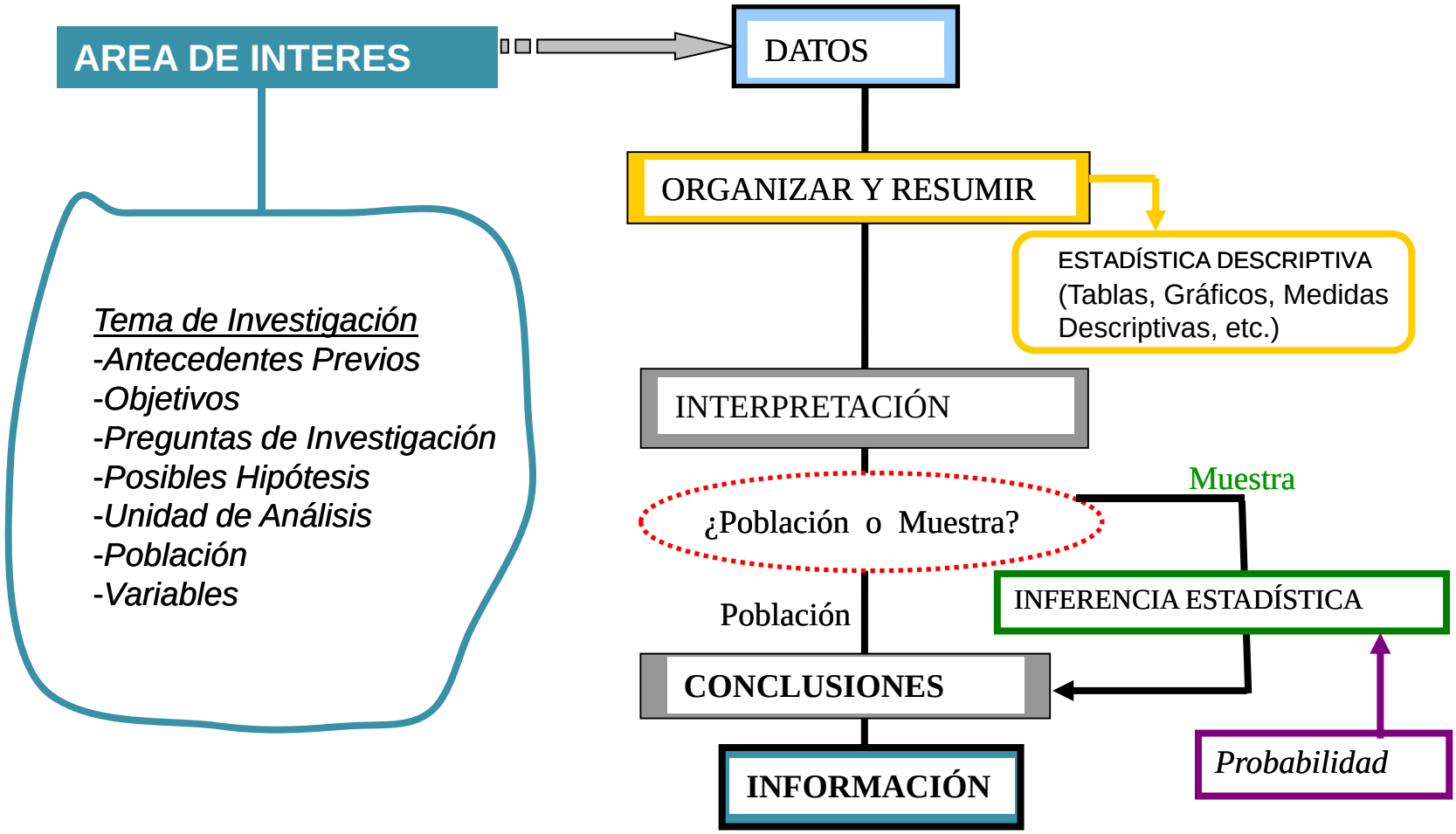
1) En Administración de Empresas: la estadística se utiliza para evaluar un producto antes de comercializarlo.
2) En Economía: para medir la evolución de los precios mediante números índice o para estudiar los hábitos de los consumidores a través de encuestas de presupuestos familiares.

Etapas de un estudio estadístico

Un análisis estadístico se lleva a cabo siguiendo las etapas habituales en el llamado método científico cuyas etapas son:

1. Planteamiento del problema: consiste en definir el objetivo de la investigación y precisar el universo o población.
2. Recogida de la información: consiste en recolectar los datos necesarios relacionados al problema de investigación.
3. Análisis descriptivo: consiste en resumir los datos disponibles para extraer la información relevante en el estudio.
4. Inferencia estadística: consiste en suponer un modelo para toda la población partiendo de los datos analizados para obtener conclusiones generales.
5. Diagnóstico: consiste en verificar la validez de los supuestos del modelo que nos han permitido interpretar los datos y llegar a conclusiones sobre la población.

Esquema de las etapas de un estudio estadístico



Ejemplos de algunos problemas a estudiar

1. Se quiere estudiar si en cierto colectivo existe discriminación salarial debida al sexo de la persona empleada.
2. Se quiere determinar el perfil de los trabajadores en términos de condiciones económicas y sociales en diferentes comunidades.
3. Se quiere estudiar el consumo de las personas de una zona determinada en cuanto a vestuario, alimentación, ocio y vivienda.
4. Se quiere determinar las tallas estándar en vestuario para mujeres españolas.
5. Se quiere determinar el tiempo que dedican al trabajo y a la familia los trabajadores de distintas empresas del país.
6. Se quiere determinar el perfil sociodemográfico de los estudiantes de una Universidad.
7. Se quiere estudiar el gasto en teléfono móvil mensual de los estudiantes de una Universidad, y si éste tiene alguna relación con su edad u otras características.

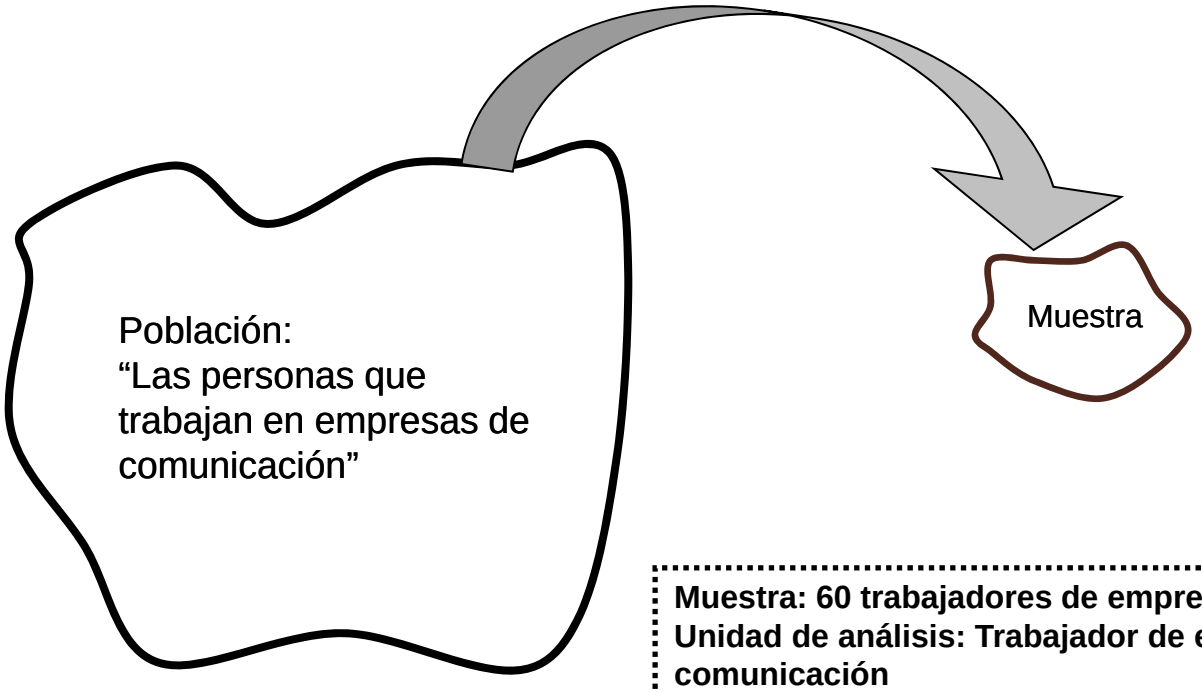
Algunos conceptos

- VARIABLE: es lo que se va a medir y representa una característica de la UNIDAD DE ANÁLISIS.

- ¿QUIÉNES VAN A SER MEDIDOS?: Los sujetos u objetos o Unidades de Análisis de una Población o una Muestra

- POBLACIÓN : Es el total de unidades de análisis que son tema de estudio.

- MUESTRA: Es un conjunto de unidades de análisis provenientes de una población.



Muestra: 60 trabajadores de empresas de comunicación
Unidad de análisis: Trabajador de empresa de comunicación
Variables: sexo, edad, salario, Nº de horas de trabajo, etc.

EJEMPLO

Problema de Investigación: Se quiere establecer el perfil de las industrias de conserva en función de algunas características.
Unidad de Análisis: Industria de Conserva
Población: Industrias de Conservas del país

Variables

- Tipo de Industria: se clasifica en industria tipo A, B, C o D. (*cualitativa nominal*)
- Nº de Empleados: se refiere al número de empleados en las líneas de producción. (*cuantitativa discreta*)
- Superficie: se refiere a los *metros cuadrados* (*unidad de medida*) disponibles para las áreas de producción. (*cuantitativa continua*)
- Calificación: calificación realizada por una institución pública sobre cumplimiento de ciertos estándares (Muy Bien, Bien, Regular, Mal). (*cualitativa ordinal*)

Datos

Industria nº	Tipo	Nº Empleados	Superficie	Calificación
1	A	100	1000,6	Muy Bien
2	B	150	1200,4	Bien
. . .	. . .	. . .	. . .	. . .
299	D	250	800,3	Mal
300	C	300	4000,2	Regular

EJEMPLO

TABLAS DE FRECUENCIA

Problema de Investigación: Se quiere establecer el perfil de las industrias de conserva en función de algunas características.
Unidad de Análisis: Industria de Conserva
Población: Industrias de Conservas del país

(1)

Tipo de Industria	*Frecuencia Absoluta* (F_i)	*Frecuencia Relativa* (f_i)	*Porcentaje* (%)
A			
B			
C			
D			
Total	300	1	100

(2)

Calificación	*Frec. Absoluta* (F_i)	*Frec.Relativa* (f_i) o %	*Frec. Absol. Acum.* (FAA_i)	*Frec. Relat. Acum.* (fra_i) o %
Muy Bien				
Bien				
Regular				
Mal			300	1 (o 100)
Total	300	1 (o 100)		

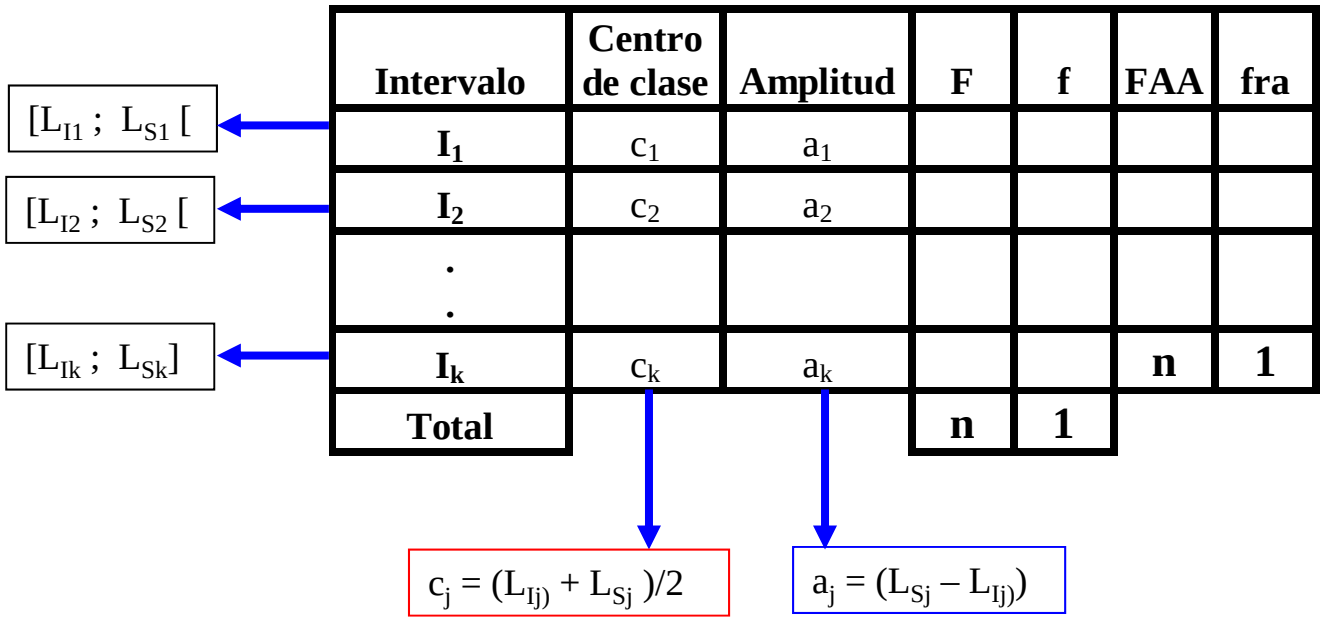
(3)

Numero de Empleados	*Frec. Absoluta* (F_i)	*Frec.Relativa* (f_i) o %	*Frec. Absol. Acum.* (FAA_i)	*Frec. Relat. Acum.* (fra_i) o %
<100				
[100-150[				
. .				
[950-1000]			300	1 (o 100%)
Total	300	1 (o 100%)		

(4)

Superficie (mt^2)	*Frec. Absoluta* (F_i)	*Frec.Relativa* (f_i) o %	*Frec. Absol. Acum.* (FAA_i)	*Frec. Relat. Acum.* (fra_i) o %
<200				
[200-400[				
. .				
[50000-5200]			300	1 (o 100%)
Total	300	1 (o 100%)		

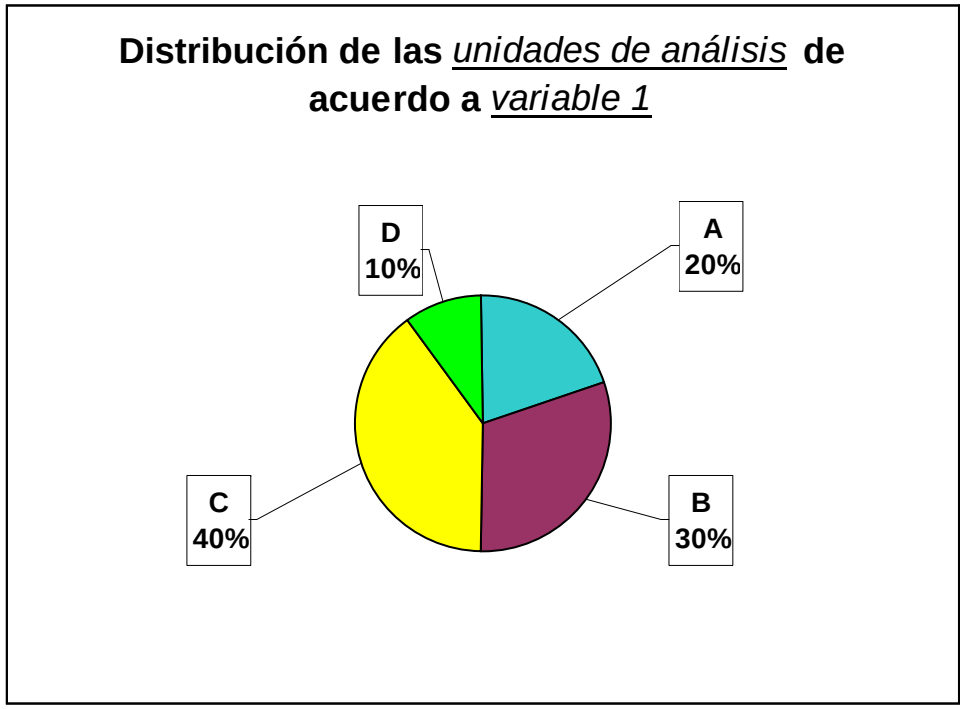
Elementos de una tabla de frecuencia cuando la variable es continua (x)

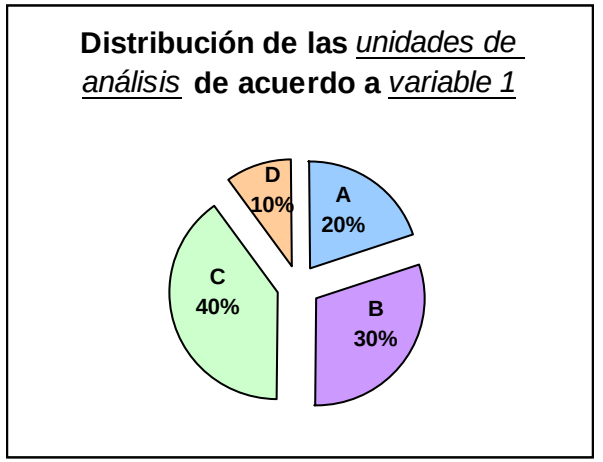


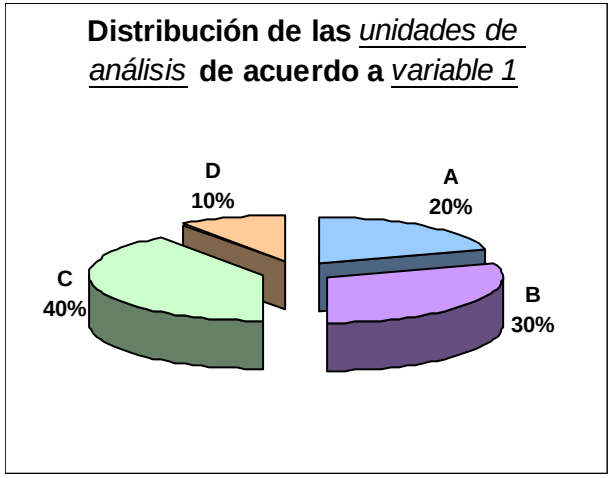
Intervalo	Centro de clase	Amplitud	F	f	FAA	fra
I_1	c_1	a_1				
I_2	c_2	a_2				
. .						
I_k	c_k	a_k			n	1
Total			n	1		

Tipos de gráficos

- 1. Gráfico de Sectores Circulares (de *Torta*)

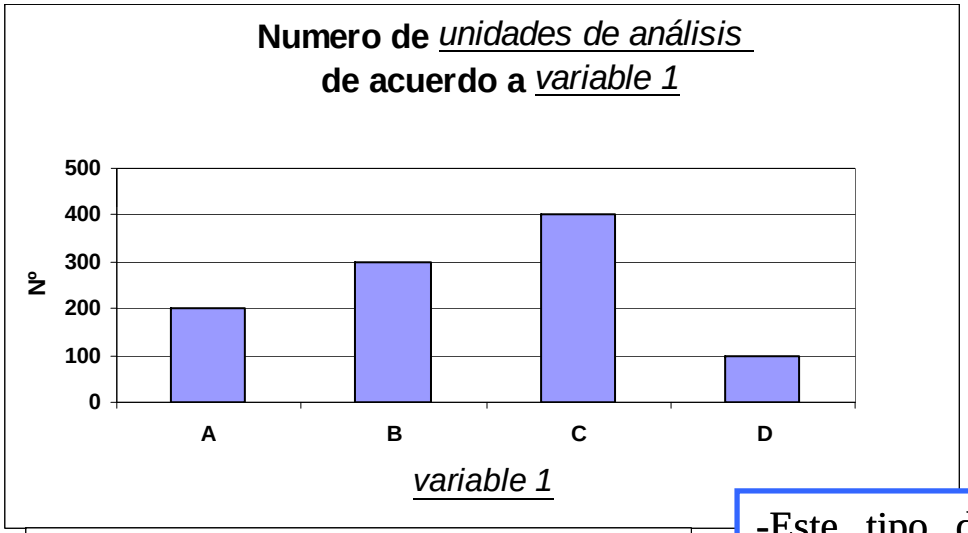


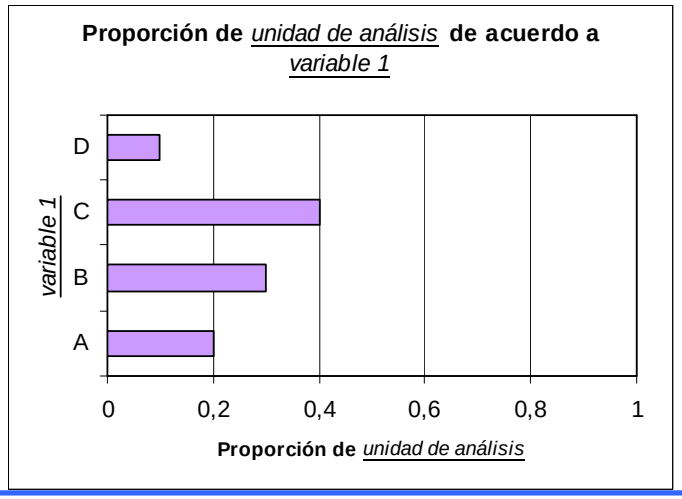


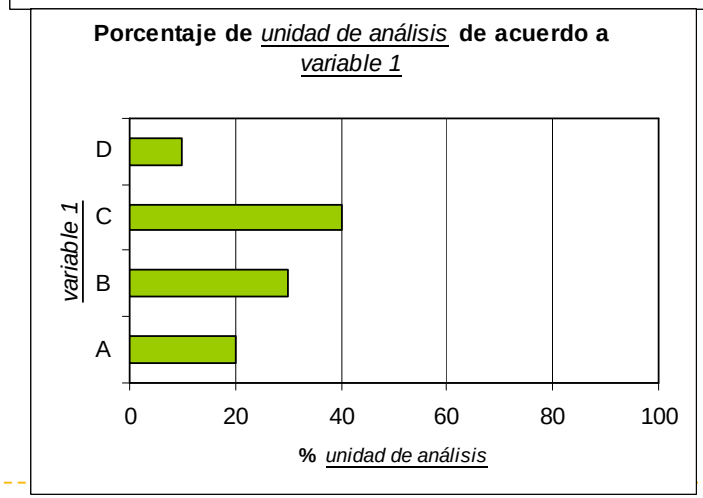


Tipos de gráficos

- 2. Gráfico de Barras





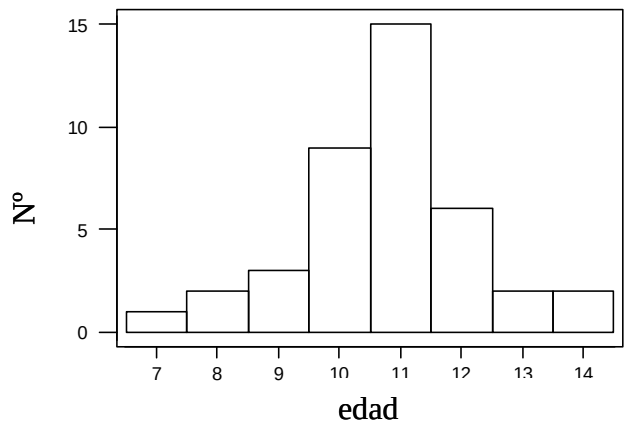


-Este tipo de gráfico se utiliza generalmente para *representar la frecuencia* de las categorías de una variable cualitativa.
-Cuando una variable es cuantitativa se puede utilizar este tipo de gráfico sólo si la variable se ha transformada en categorías.
-Hay distintas versiones de estos gráficos (por ejemplo en Excel), y en algunos casos son muy útiles para describir el comportamiento de una variable en distintos grupos.

Tipos de gráficos

- 3. Histograma

Distribución de los hijos de trabajadores de la empresa de acuerdo a edad

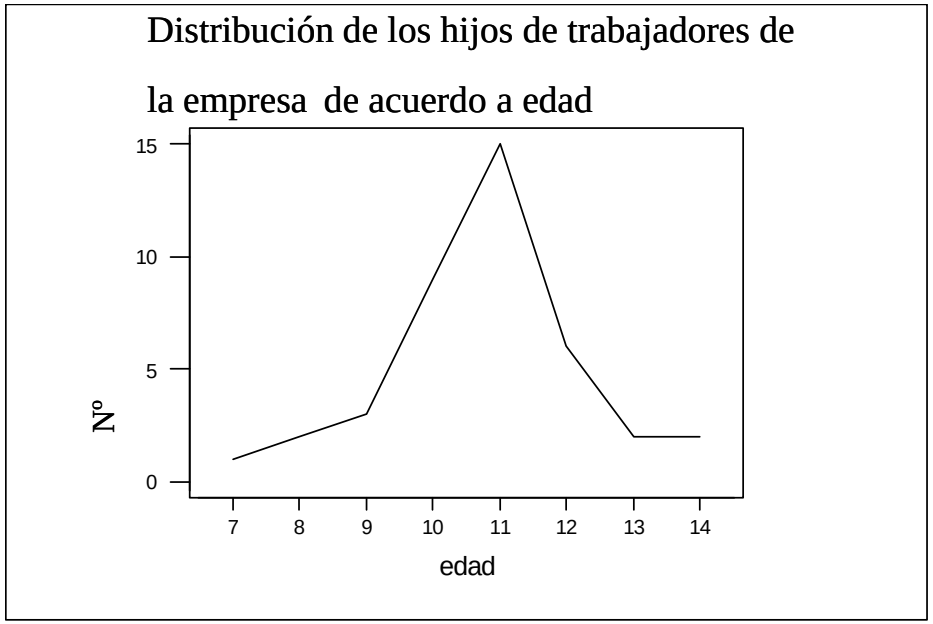


Ejemplo
En el gráfico se puede observar el número de hijos , de menor edad (7-8 años), las de mayor edad (13-14 años); y además que la mayoría de hijos de los trabajadores están entre los 10 y 12 años.

Histograma
- Permite la representación de la frecuencia de una variable Cuantitativa.
- El **eje** x se refiere a la variable.
- El **eje** y se refiere a la frecuencia (Nº , %).
- Cada **barra** representa la frecuencia de la variable en la población en estudio (o la muestra).
-El histograma se puede construir desde los datos de la tabla de frecuencia de la variable en estudio.

Tipos de gráficos

- 4. Polígono de Frecuencia

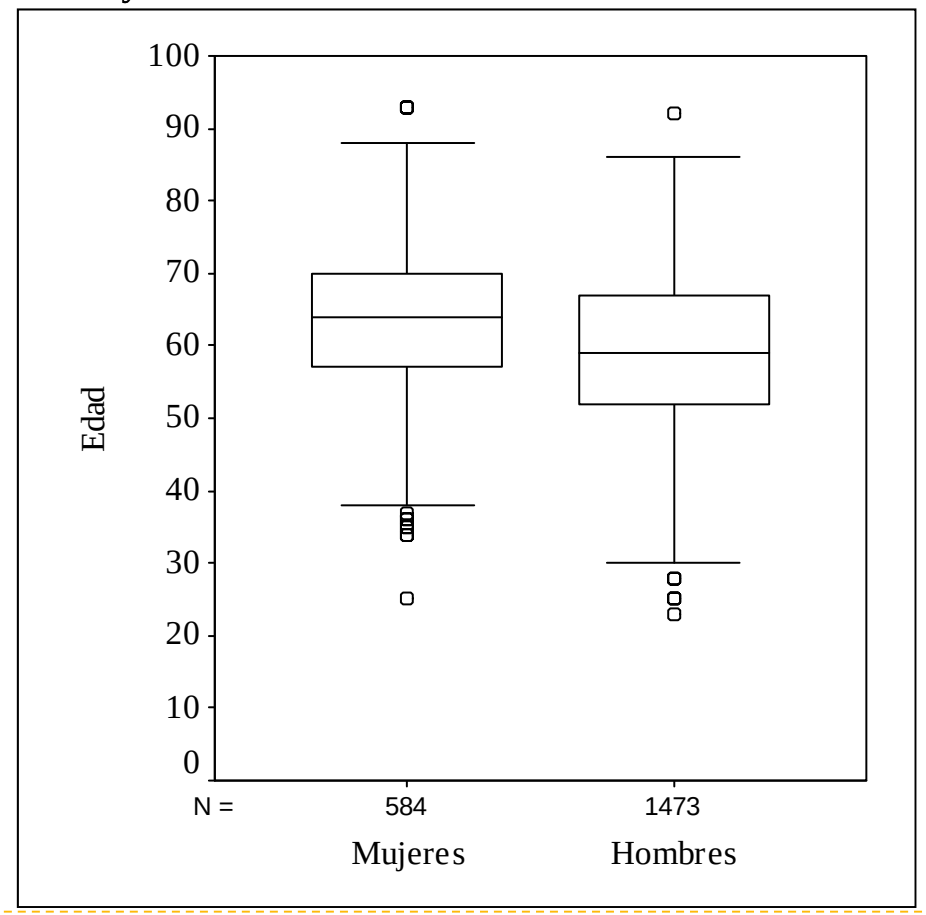


-Esta representación se basa en el Histograma.
-Sólo es útil para variables cuantitativas.
-El eje x se refiere a la variable.
- El eje y se refiere a la frecuencia (Nº , %).
-Los puntos que permiten la unión de las líneas representa el *centro de clase (o marca de clase).*

Tipos de gráficos

- 5. Diagrama de Caja

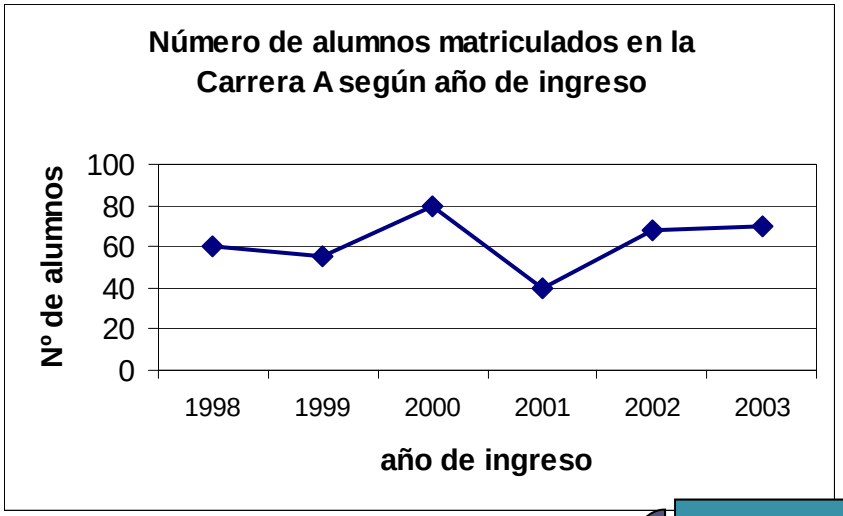
Edad de las personas que se realizaron angioplastía entre 1980 y 2000

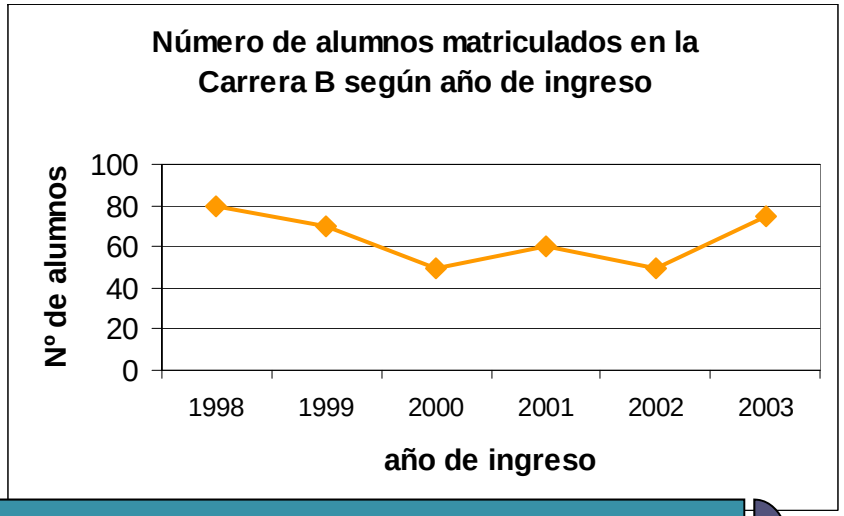


- Permite identificar gráficamente la mediana, los cuartiles 1 y 3 (percentiles 25 y 75), mínimo y máximo de una variable.
- Sólo es útil para variables **cuantitativas**.
-El **eje x** permite identificar la poblacion en estudio.
- El **eje y** representa los valores de la variable en estudio.

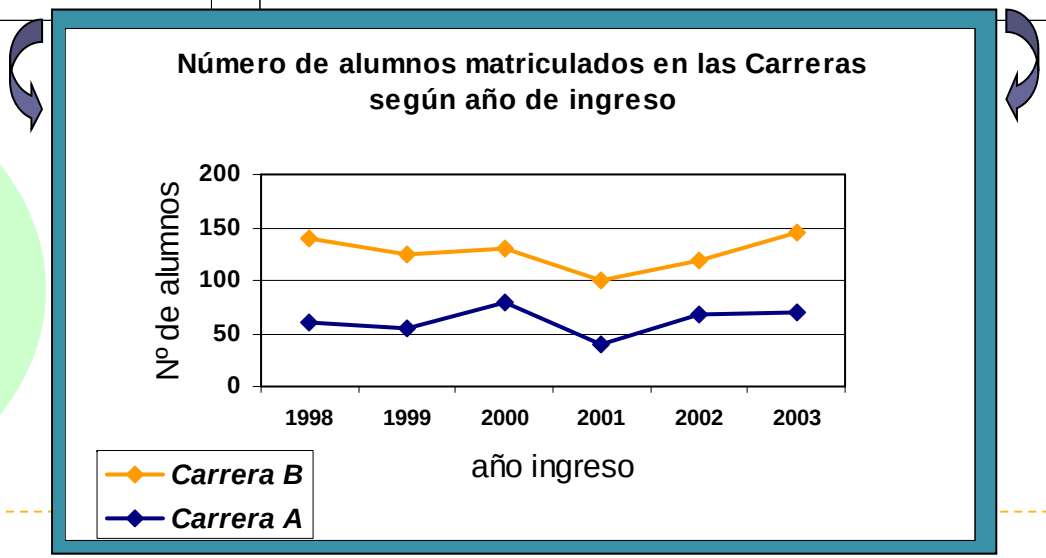
Tipos de gráficos

- 6. Otros





	Nº de alumnos	
año de ingreso	*Carrera A*	*Carrera B*
1998	60	80
1999	55	70
2000	80	50
2001	40	60
2002	68	50
2003	70	75



Notación

- *Variables Cuantitativas*

x = variable $\quad x_i$ = valor de la variable en el individuo i $\quad i = 1,\ldots,n$

y = variable $\quad y_i$ = valor de la variable en el individuo i

a, b, c : constantes

$$\sum_{i=1}^{n} c = c + \cdots + c = nc$$

$$\sum_{i=1}^{n} cx_i = cx_1 + \cdots + cx_n = c\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + \cdots + x_n^2$$

$$\sum_{i=1}^{n} (ax_i + b) = (ax_1 + b) + \cdots + (ax_n + b) = a\sum_{i=1}^{n} x_i + b$$

$$\left(\sum_{i=1}^{n} x_i\right)^2 = (x_1 + \cdots + x_n)^2$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1) + \cdots + (x_n + y_n)$$

$$\sum_{i=1}^{n} (x_i y_i) = (x_1 y_1) + \cdots + (x_n y_n)$$

Medidas de posición o tendencia central

Medidas de tendencia central

- -Media Aritmética (Promedio)
- -Mediana
- -Moda

Media Aritmética o Promedio

- La media aritmética o simplemente media, que denotaremos por X , es el número obtenido al dividir la suma de todos los valores de la variable entre el numero total de observaciones, y se define por la siguiente expresión

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

Media Aritmética o Promedio

- Si los datos están agrupados en intervalos, la expresión de la media aritmética, es la misma, pero utilizando la marca de clase (Xi).

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

Mediana

- Dada una distribución de frecuencias con los valores ordenados de menor a mayor, llamamos mediana y la representamos por Me, al valor de la variable, que deja a su izquierda el mismo número de frecuencias que a su derecha.
- Calculo de la mediana:
  - Variara según el tipo de dato

$$\boxed{M_E = x_{(k)}} \quad \text{Si } n \text{ es impar}$$

$$\boxed{M_E = \frac{x_{(k)} + x_{(k+1)}}{2}} \quad \text{Si } n \text{ es par}$$

$$x_{(k)} = \text{dato del centro}$$

Moda

- La moda es el valor de la variable que más veces se repite, y en consecuencia, en una distribución de frecuencias, es el valor de la variable que viene afectada por la máxima frecuencia de la distribución.
- En distribuciones no agrupadas en intervalos se observa la columna de las frecuencias absolutas, y el valor de la distribución al que corresponde la mayor frecuencia será la moda. A veces aparecen distribuciones de variables con más de una moda (bimodales, trimodales, etc), e incluso una distribución de frecuencias que presente una moda absoluta y una relativa.

Datos:	**Moda**
Cualitativos y Cuantitativos	M_o ="el dato que más se repite"

Moda

- En el caso de estar la variable agrupada en intervalos de distinta amplitud, se define el intervalo modal, y se denota por (L_{i-1} , L_i], como aquel que posee mayor densidad de frecuencia (h_i); la densidad de frecuencia se define como

$$h_i = \frac{n_i}{a_i}$$

- Una vez identificado el intervalo modal procederemos al cálculo de la moda, a través de la fórmula:

$$Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} c_i$$

- En el caso de tener todos los intervalos la misma amplitud, el intervalo modal será el que posea una mayor frecuencia absoluta (n_i) y una vez identificado este, empleando la fórmula:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} c_i$$

Medidas de posición no central (Cuantiles)

- Los cuantiles son aquellos valores de la variable, que ordenados de menor a mayor, dividen a la distribución en partes, de tal manera que cada una de ellas contiene el mismo número de frecuencias.
- Los cuantiles más conocidos son:

a) Cuartiles (Qi): Son valores de la variable que dividen a la distribución en 4 partes, cada una de las cuales engloba el 25 % de las mismas. Se denotan de la siguiente forma: Q1 es el primer cuartil que deja a su izquierda el 25 % de los datos; Q2 es el segundo cuartil que deja a su izquierda el 50% de los datos, y Q3 es el tercer cuartil que deja a su izquierda el 75% de los datos. (Q2 = Me)

b) Deciles (Di): Son los valores de la variable que dividen a la distribución en las partes iguales, cada una de las cuales engloba el 10 % de los datos. En total habrá 9 deciles. (Q2 = D5 = Me)

c) Centiles o Percentiles (Pi): Son los valores que dividen a la distribución en 100 partes iguales, cada una de las cuales engloba el 1 % de las observaciones. En total habrá 99 percentiles. (Q2 = D5 = Me = P50)

Percentiles, Deciles o Cuartiles

- Percentil, Decil o Cuartil: corresponde al valor que toma la variable (cuantitativa), cuando los *n* datos están ordenados de Menor a Mayor

Percentiles, Deciles o Cuartiles
- -Percentil (ejemplo: 25, 50, 75)
- -Decil (ejemplo: 4, 5, 8)
- -Cuartil (ejemplo: 1, 2, 3)

Percentiles, Deciles o Cuartiles

El Percentil va de 1 a 100

El percentil 25 (25/100): es el valor de la variable que reúne al menos el 25% de los datos

Ejemplo: Si N=80, el 25% de 80 es 20; por lo tanto, se busca el dato que este en la posición 20.

Si N=85, el 25% de 85 es 21,25; por lo tanto se busca el dato que este en la posición 22.

El Decil va de 1 a 10

El Decil 4 (4/10): es el valor de la variable que reúne al menos el 40% de los datos

Ejemplo: Si N=80, el 40% de 80 es 32; por lo tanto, se busca el dato que este en la posición 32.

Si N=85, el 40% de 85 es 34; por lo tanto se busca el dato que este en la posición 34.

El Cuartil va de 1 a 4

El Cuartil 3 (3/4): es el valor de la variable que reúne al menos el 75% de los datos

Ejemplo: Si N=80, el 75% de 80 es 60; por lo tanto, se busca el dato que este en la posición 60.

Si N=85, el 75% de 85 es 63,75; por lo tanto se busca el dato que este en la posición 64.

Medidas de dispersión

Medidas de Dispersión

- **-Rango**
- **-Varianza**
- **-Desviación Estándar**

Recorrido o Rango

- Se define como la diferencia entre el máximo y el mínimo valor de la variable:

$$R = \max(x_i) - \min(x_i)$$

Varianza

- La varianza mide la mayor o menor dispersión de los valores de la variable respecto a la media aritmética. Cuanto mayor sea la varianza mayor dispersión existirá y por tanto menor representatividad tendrá la media aritmética.
- La varianza se expresa en las mismas unidades que la variable analizada, pero elevadas al cuadrado.

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

Desviación Típica o Estandar

- Se define como la raíz cuadrada con signo positivo de la varianza.

$$s = \sqrt{s^2}$$

Comparación entre Variables

- Se refiere al comportamiento de las variables cuantitativas en un grupo.
  - *Por ejemplo: Si se tiene un conjunto de personas a las que se les mide Estatura, Peso, Edad: Entre estas variables ¿cuál presenta mayor variación?*

Coeficiente de Variación

$$cv = \frac{s}{\overline{x}}$$

Medidas de forma

- Además de la posición y la dispersión de los datos, otras medidas de interés en una distribución de frecuencias son:
  - Asimetría
  - Kurtosis o Apuntamiento

Coeficiente de Asimetría de Fisher

$$CA = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n \cdot s^3}$$

- Según sea el valor de *CA*, diremos que la distribución es asimétrica a derechas o positiva, a izquierdas o negativa, o simétrica, o sea:
  - Si *CA*=0 si la distribución es simétrica alrededor de la media.
  - Si *CA*<0 si la distribución es asimétrica a la izquierda.
  - Si *CA*>0 si la distribución es asimétrica a la derecha.

Coeficiente de Apuntamiento o Kurtosis

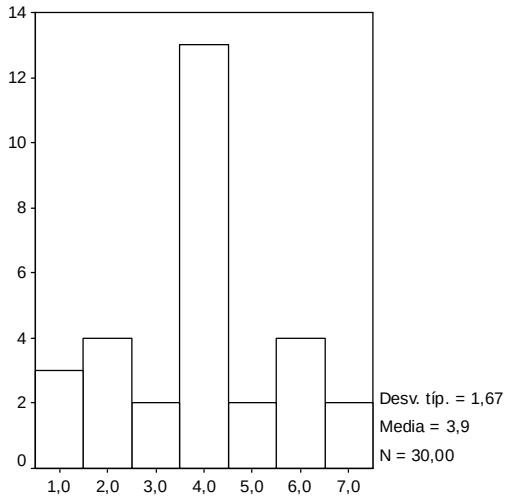
- Se refiere al grado de apuntamiento que tiene una distribución; para determinarlo, emplearemos el coeficiente de curtosis de Fisher.

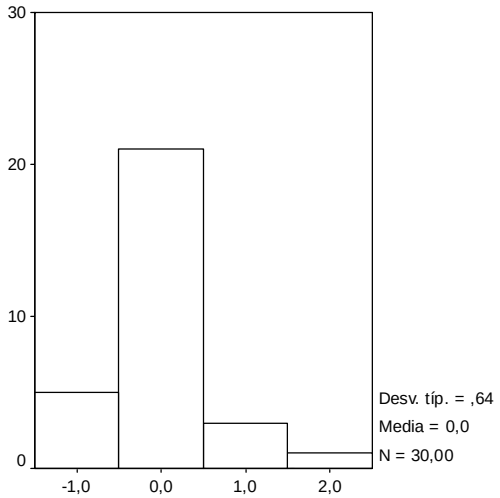
$$CAp = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n \cdot s^4}$$

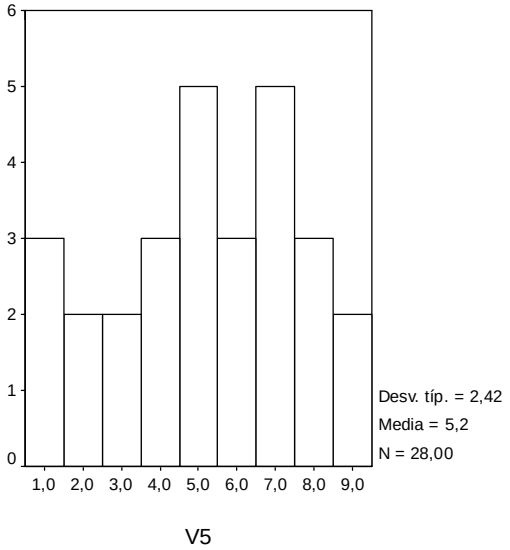
- Si *CAp*=0 la distribución se dice normal (similar a la distribución normal de Gauss) y recibe el nombre de *mesocúrtica*.
- Si *CAp*>0, la distribución es más puntiaguda que la anterior y se llama *leptocúrtica*, (mayor concentración de los datos en torno a la media).
- Si *CAp*<0 la distribución es más plana y se llama *platicúrtica*.

Otras medidas o Coeficientes

- Ejemplos Histogramas con distinta asimetría y apuntamiento





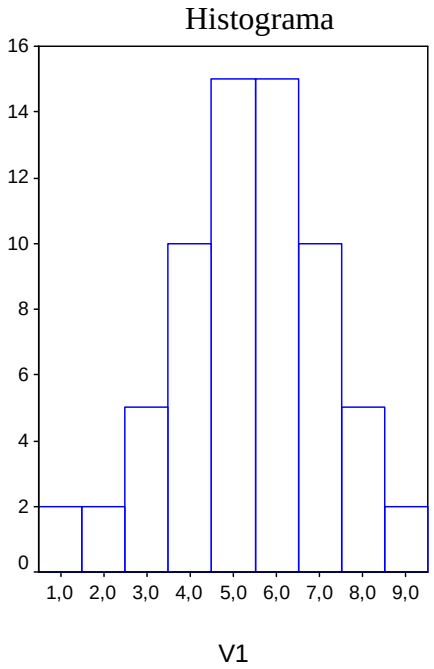


Otras medidas o Coeficientes

- Ejemplos

Datos

1	4	4
1	4	4
1	4	5
2	4	5
2	4	6
2	4	6
2	4	6
3	4	6
3	4	7
4	4	7



Medidas descriptivas

Media	3,9
Mediana	4
Moda	4
Desviación estándar	1,67
Varianza de la muestra	2,78
kurtosis	-0,43
Coeficiente de asimetría	-0,02
Rango	6
Mínimo	1
Máximo	7
Cuenta	30

Media, Desviación típica, Coeficientes de Asimetría y Apuntamiento para datos Agrupados (tabla de frecuencias)

Tabla de frecuencia (para variable cuantitativa)

Intervalo	Centro de clase	Amplitud	F	f	FAA	fra
I_1	c_1	a_1	n_1	f_1		
I_2	c_2	a_2	n_2	f_2		
⋮	⋮	⋮	⋮	⋮		
I_k	c_k	a_k	n_k	f_k	**n**	**1**
Total			**n**	**1**		

Sea c_j la marca de clase (o centro de clase) y f_j la frecuencia relativa de la clase *j*, donde *j*=1, 2,…, k.

1) La *Media para datos agrupados* es igual a la suma de los productos de las marcas de clase por sus frecuencias relativas, de la forma:

$$Media_c = \bar{x}_c = \sum_{j=1}^{k} c_j f_j$$

2) La *Desviación típica* para datos agrupados esta dada por:

$$s_c = \sqrt{\sum_{j=1}^{k} (c_j - \bar{x}_c)^2 f_j}$$

3) El *Coeficiente de Asimetría* para datos agrupados esta dado por:

$$CA_c = \frac{\sum_{j=1}^{k} (c_j - \bar{x}_c)^3 f_j}{s_c^3}$$

4) El *Coeficiente de apuntamiento* para datos agrupados esta dada por:

$$CAp_c = \frac{\sum_{j=1}^{k} (c_j - \bar{x}_c)^4 f_j}{s_c^4}$$