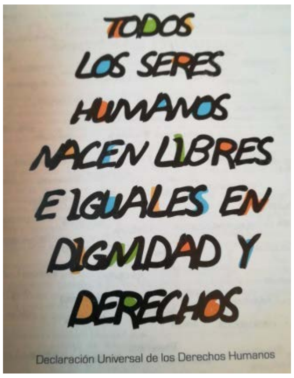
UNIDAD INTEGRADA BILINGÜE MATEMÁTICAS

LIBERTAD, IGUALDAD Y FRATERNIDAD – 20th Century

STATISTICS


TODOS LOS SERES HUMANOS NACEN LIBRES E IGUALES EN DIGNIDAD Y DERECHOS
Declaración Universal de los Derechos Humanos

What are we going to study?

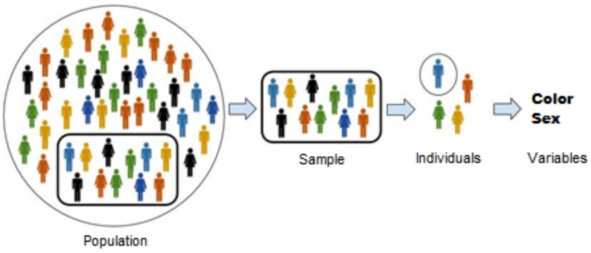
Vamos a trabajar datos de **números de muertos por edades** en una muestra de 300 individuos que provocó **la gripe que asoló España en el año 1918**. Con ellos vamos a aprender a calcular y a interpretar algunos conceptos estadísticos.

We are going to do some statistical tables and graphs with these data. They contain information that has been obtained during a statistical process.

1.- POPULATION AND SAMPLE

Para realizar un estudio estadístico es necesario conocer los siguientes conceptos:

- **Population** (población): the set of all of the elements that we are studying.
- **Sample** (muestra): a smaller group within the population. By studying this group we can infer characteristics for the entire population.
- **Individual** (individuo): one of the elements that make up the population or simple.



Un hecho importante del siglo XX que implicó miles de muertes (incluso más que la I Guerra Mundial), fue la pandemia que asoló el mundo en los años 1918 hasta 1920. Dicha pandemia fue llamada "gripe española" por la atención que se le dio en los medios de comunicación de nuestro país, ya que no estábamos sometidos a la fuerte censura de los países que participaban en la I Guerra Mundial.

Vamos a estudiar una variable numérica definida como el número de muertes causada por esta gripe en España durante dichos años, diferenciando por el grupo de edad. A continuación aparecen los 300 datos de nuestra muestra:

31	**40**	**11**	**25**	**33**	**45**	**9 meses**		**7**	**10**	**15**	**84**	**92**
17	**10**	**74**	**36**	**25**	**18**	**53**	**56**	**45**	**23**	**23**	**33**	**36**
49	**6**	**32**	**80**	**63**	**36**	**3**	**4**	**25**	**22**	**30**	**35**	**34**
50	**41**	**57**	**81**	**24**	**35**	**52**	**2 meses**		**30**	**25**	**20**	**6**
81	**20**	**33**	**18**	**63**	**21**	**3**	**4**	**11**	**3**	**3**	**30**	**25**
29	**38**	**41**	**35**	**57**	**1**	**28**	**33**	**90**	**28**	**32**	**59**	**1**
34	**31**	**36**	**88**	**60**	**20**	**1**	**4**	**28**	**32**	**95**	**4**	**24**
39	**51**	**11**	**16**	**49**	**26**	**64**	**57**	**45**	**2**	**3**	**15**	**25**
1	**5 meses**		**22**	**6**	**11**	**33**	**30**	**50**	**42**	**62**	**4**	**4**
38	**13**	**27**	**9**	**27**	**21**	**43**	**8**	**32**	**22**	**39**	**66**	**28**
3	**22**	**30**	**58**	**32**	**10**	**24**	**25**	**94**	**2 meses**		**1**	**18**
2	**24**	**43**	**89**	**2**	**22**	**39**	**51**	**55**	**14**	**2**	**19**	**8**
27	**48**	**56**	**30**	**2**	**18**	**33**	**31**	**60**	**11 meses**		**16**	**35**
46	**1**	**2**	**5 meses**		**39**	**29**	**16**	**6**	**6**	**15**	**51**	**17**
31	**19**	**1**	**39**	**4**	**4**	**4**	**45**	**23**	**5**	**19**	**41**	**79**
31	**20**	**6**	**13**	**27**	**42**	**36**	**52**	**27**	**1**	**21**	**39**	**74**
48	**19**	**8**	**26**	**24**	**15**	**16**	**37**	**40**	**40**	**27**	**7 meses**	
15	**12**	**26**	**40**	**37**	**17**	**34**	**23**	**31**	**2**	**43**	**20**	**34**
54	**13**	**96**	**19**	**41**	**38**	**74**	**35**	**8**	**28**	**23**	**35**	**19**
21	**71**	**3**	**12**	**32**	**27**	**5 meses**		**1**	**25**	**41**	**53**	**66**

16	**6**	**31**	**47**	**31**	**22**	**25**	**8**	**1**	**41**	**74**	**9**	**2**
27	**44**	**44**	**10**	**22**	**26**	**6**	**33**	**26**	**39**	**25**	**60**	**27**
28	**11**	**11**	**16**	**7**	**6 meses**		**26**	**4**	**17**	**23**	**21**	**27**
32	**32**	**23**	**46**	**22**	**31**	**31**	**28**	**40**	**63**			

2.- STATISTICAL VARIABLES

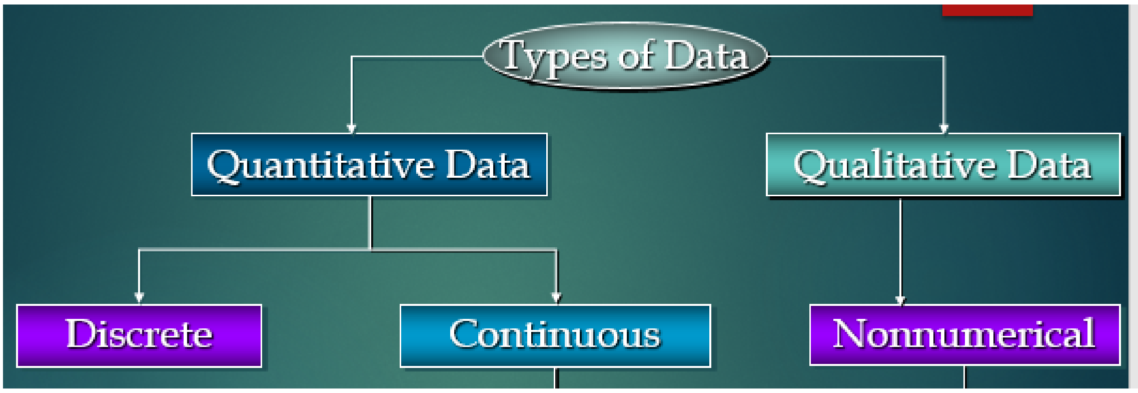
Una variable estadística (**statistical variable**) es cualquier cualidad que estudiamos en los individuos de una muestra o una población. The types of statistical variables are:

- Cualitativa (**qualitative variable**): it cannot be described with a number. It needs a quality.

- Cuantitativa (**quantitative variable**): its value is expressed with numbers.

  a) Discreta (**discrete quantitative variable**): it only supports isolated values (valores finitos).

  b) Continua (**continuous quantitative variable**): it can include all the values of an interval.



Activity 1.- Identify your statistical variable and classify it justifying your choice.

3.- PREPARING FREQUENCY TABLES

Once we have the data, we need to tabulate them. We have to make a table to organise the information. We call this a **frequency table.**

La frecuencia absoluta (**absolute frequency**) de un dato estadístico es el número de veces que se repite. Se representa por fi. La suma de las frecuencias absolutas es el número da datos.

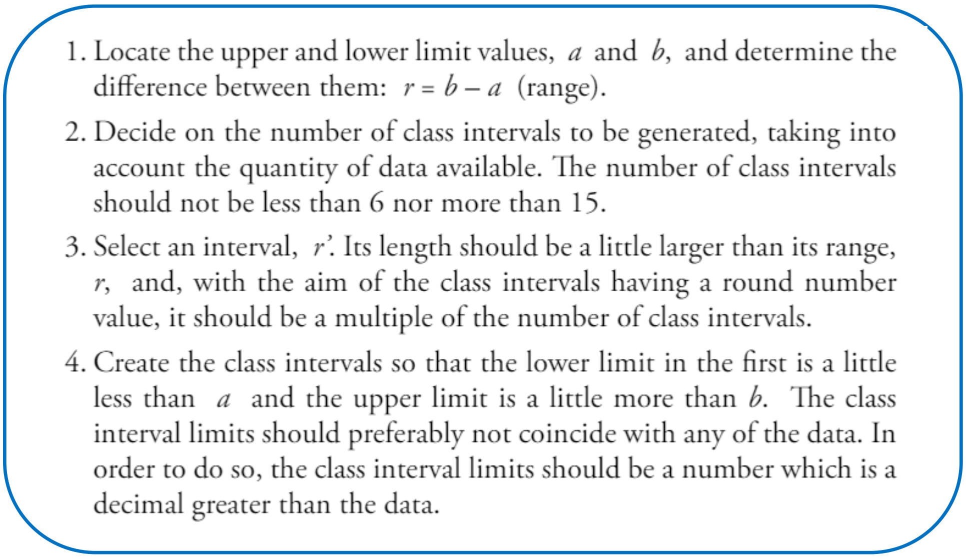
$$f_1 + f_2 + \cdots + f_n = N$$

La frecuencia relativa (**relative frequency**) de un dato estadístico es el cociente entre la frecuencia absoluta y el número total de datos. Se representa por h_i. La suma de las frecuencias relativas es igual a uno.

$$h_1 + h_2 + \cdots + h_n = 1$$

A) Creating a frequency table with data grouped into intervals

In order to prepare a frequency table with grouped data, you will need to take the following steps:



1. Locate the upper and lower limit values, a and b, and determine the difference between them: $r = b - a$ (range).
2. Decide on the number of class intervals to be generated, taking into account the quantity of data available. The number of class intervals should not be less than 6 nor more than 15.
3. Select an interval, r'. Its length should be a little larger than its range, r, and, with the aim of the class intervals having a round number value, it should be a multiple of the number of class intervals.
4. Create the class intervals so that the lower limit in the first is a little less than a and the upper limit is a little more than b. The class interval limits should preferably not coincide with any of the data. In order to do so, the class interval limits should be a number which is a decimal greater than the data.

Activity 2.- Make the frequency table with our data.

<u>CLASES</u>: Grupos de edad.

<1 **[1,4]** **[5,9]** **[10,14]**

[15,19] **[20,24]** **[25,29]** **[30,34]**

[35,39] **[40,44]** **[45,49]** **[50,54]**

[55,59] **>60**

Interval = *you can group the data into intervals following the previous steps*
Class mark x_i(*marca de clase*)
= *to find the class markfor each interval, we take the mid value of the two extremes*
f_i = *absolute frequency*
h_i = *realtive frequency*
N = *número total de datos*

Interval	***Class mark*** x_i	f_i	$h_i = \frac{f_i}{N}$
		N =	**1**

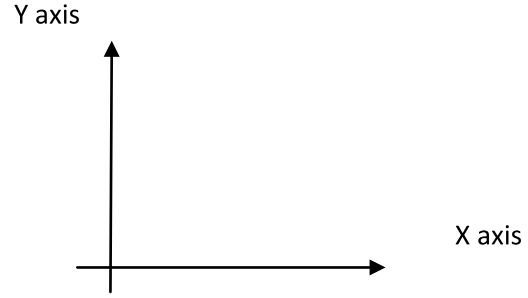
4.- STATISTICAL CHARTS

The **statistical charts** (gráficos estadísticos) allow us to easily understand what they want to tell us, with just a quick look.

a) BAR CHARTS (Diagrama de barras)

Bar chart are used to show how discrete quantitative and qualitative variables are distributed. This is why the bars are narrow and are located on specific values of the variable. La altura de cada barra indica la frecuencia de cada dato.

Activity 3.- Represent our data (discrete quantitative variable) in a bar chart using the frequency table (activity 2).

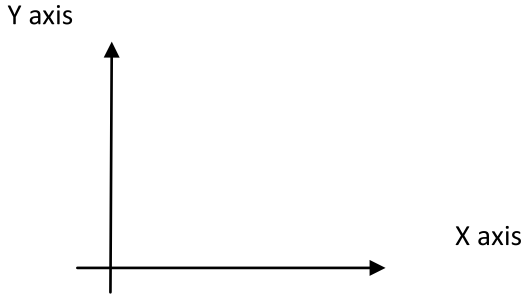


- X axis: data $\boldsymbol{x_i}$
- Y axis: absolute frequency $\boldsymbol{f_i}$

b) FREQUENCY HISTOGRAMS (Histogramas)

Histograms are used for distributions with quantitative variables (continuous or discrete). This is why rectangles whose bases are the lenght of the intervals are used.

Activity 4.- Represent our data (quantitative variable) in a frequency histogram using the frequency table (activity 2).



- X axis: intervals
- Y axis: frequency $\boldsymbol{f_i}$

c) FREQUENCY POLYGONS (Polígono de frecuencias)

Frequency polygons are used in the same cases as histograms. To create one, we draw a line along the top of the middle points of the columns (of the histogram) fronthe axis to the end of the chart.

Activity 5.- Represent the frequency polygon using the frequency histogram (activity 4).

d) PIE CHARTS (Diagrama de sectores)

In a pie chart, the angle of each sector is proportional to the corresponding frequency. It can be used for all kinds of variable, but it is very frequently used for qualitative variables.

Un diagrama de sectores (pie chart) está compuesto por un círculo dividido en sectores que representa cada uno de los valores de la variable.

La amplitud (amplitude) de cada sector, es proporcional a la frecuencia del dato que representa.

$$\text{Ángulo del sector circular } = \frac{\mathrm{f_i}}{\mathrm{N}} \cdot 360^{\text{º}} = \mathrm{h_i} \cdot 360^{\text{º}}$$

How to make a pie chart yourself?

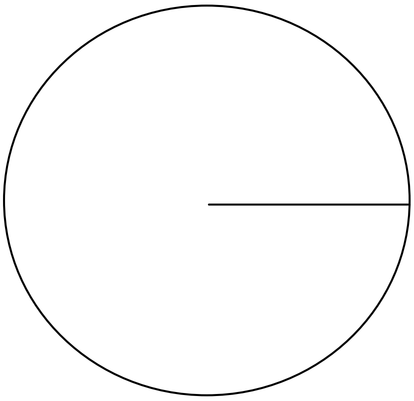
1. First, calculate the amplitude of each sector. So you have the degrees that each pie slice (sector) contains.

$h_i = \frac{f_i}{N}$	Amplitude $\mathbf{h_i \cdot 360^{\text{º}}}$

	360º

2. Then, we draw a circle and we divide it into sectors with the amplitude of the step 1. (you can use a protractor to measure the degrees of each sector).
3. Finish up by coloring each sector and giving it a label. And don´t forget a title for your pie chart.

Activity 6.- Represent our data (discrete quantitative variable) in a pie chart using the frequency table (activity 2).



□ ______________________

□ ______________________

□ ______________________

□ ______________________

□ ______________________

□ ______________________

□ ______________________

□ ______________________

□ ______________________

□ ______________________

□ ______________________

□ ______________________

□ ______________________

□ ______________________

5.- STATISTICAL PARAMETERS

We are going to study three types of statistical parameters:

A) CENTRALISATION PARAMETERS (parámetros de centralización)

- Mean (media)
- Median (mediana)
- Mode (moda)

B) DISPERSION PARAMETERS (parámetros de dispersión)

- Range (rango)
- Mean deviation (desviación media)
- Variance (varianza)
- Standard deviation (desviación típica)

C) POSITION PARAMETERS (parámetros de posición)

- Median (mediana)
- Quartiles (cuartiles)

The centralisation parameters show us which value (centre) the data is distributed around.

The dispersión parameters show us the distance of the data from the distribution values. To measure the dispersión of a distribution, the key is to measure the degree of separation from the mean (media).

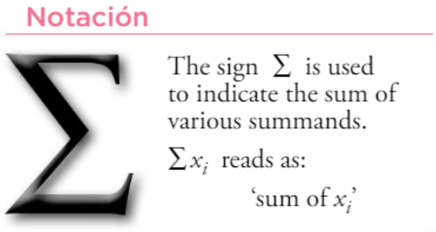
The position parameters show us a place (a position) in relation to the other values in the distribution.

- **MEAN (media)** $\bar{\mathrm{x}}$

The mean, or average, is written as $\bar{\mathrm{x}}$. Se obtiene al dividir la suma de los productos de cada dato por su frecuencia absoluta entre el número total de datos.

$$\bar{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \cdots + x_n \cdot f_n}{N}$$

$$\text{Abbreviated} \qquad \bar{x} = \frac{\sum x_i \cdot f_i}{N} = \frac{\sum x_i \cdot f_i}{\sum f_i}$$



Notación

$\sum$

The sign $\sum$ is used to indicate the sum of various summands.

$\sum x_i$ reads as:

'sum of x_i'

Activity 7.- Calculate the mean of our statistical distribution.

x_i	f_i	$x_i \cdot f_i$
	N = $\sum f_i$	$\sum x_i \cdot f_i$

♦ **<u>MEDIAN (mediana)</u> Me**

If we organise the data from smallest to largest, the median Me, is the value located in the middle. In others words, it has the same number of individuals above and below it. If the number of data is even (par), the median is the average of the two central values.

Activity 8.- Calculate the median of our statistical distribution.

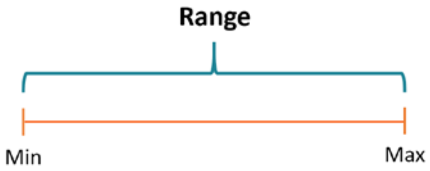
♦ **<u>MODE (moda)</u> Mo**

The mode Mo is the value that occurs most often. Es el dato (o datos) con mayor frecuencia absoluta.

Activity 9.- Calculate the mode of our statistical distribution.

♦ **<u>RANGE (rango)</u> R**

The range R is the difference between the largest and smallest number. In other words, it is the length of the section where the data is located.



Activity 10.- Calculate the range of our statistical distribution.

♦ **<u>MEAN DEVIATION (desviación media)</u> MD**

The mean deviation MD is the average distance of the data from the mean.

$$\mathrm{MD} = \frac{f_1 \cdot |x_1 - \bar{x}| + f_2 \cdot |x_2 - \bar{x}| + \cdots + f_n \cdot |x_n - \bar{x}|}{N}$$

$$\text{Abbreviated} \quad \mathrm{MD} = \frac{\sum f_i \cdot |x_i - \bar{x}|}{N}$$

Activity 11.- Calculate the mean deviation of our statistical distribution.

♦ VARIANCE (varianza)

This is the average of the squares of the distance of the data from the mean:

$$\text{Variance} = \frac{f_1 \cdot (x_1 - \bar{x})^2 + f_2 \cdot (x_2 - \bar{x})^2 + \cdots + f_n \cdot (x_n - \bar{x})^2}{N} = \frac{\sum f_i \cdot (x_i - \bar{x})^2}{N}$$

This formula is the equivalent to the following:

$$\text{Variance} = \frac{f_1 \cdot {x_1}^2 + f_2 \cdot {x_2}^2 + \cdots + f_n \cdot {x_n}^2}{N} - \bar{x}^2 = \frac{\sum f_i \cdot {x_i}^2}{\sum f_i} - \bar{x}^2$$

Activity 12.- Calculate the variance with the second formula of our statistical distribution.

x_i	f_i	$x_i \cdot f_i$	$f_i \cdot {x_i}^2$

	N = $\sum f_i$	$\sum x_i \cdot f_i$	$\sum f_i \cdot {x_i}^2$

♦ STANDARD DEVIATION (desviación típica) σ

This is the positive square root of the variance:

$$\sigma = +\sqrt{\text{Variance}} = +\sqrt{\frac{\sum f_i \cdot {x_i}^2}{\sum f_i} - \bar{x}^2}$$

With the values for $\bar{x}$ and σ, we have a pretty good idea of what the distribution is like. The mean tells us where its centre is. The standard deviation gives us an idea of how far away the mean the data is. In other words, how disperse it is.

Activity 13.- Calculate the standard deviation of our statistical distribution.