UNIT 3. UNIVARIATE FREQUENCY DISTRIBUTIONS AND GRAPHIC PRESENTATION

FREQUENCY TABLES

It often happens that, once the data under study have been obtained, they are sufficiently numerous that a simple visual inspection of them provides practically no information. It is therefore necessary to summarize all the data and group them together if necessary.

Tabulation is the process of sorting and grouping a set of data. As a result, the frequency table of the observations are obtained. A frequency table allows grouping data into mutually exclusive classes showing the number of observations in each class.

The following notation is used:

- Total number of observations: N
- Values of the variable: x_i
- Number of different values of the variable: k
- Range: $\{x_1, \dots, x_k\}$
- Assumption (if possible): $x_1 < x_2 < \dots < x_k$

Frequencies

Let X be a variable or characteristic measured in a population of size N that takes the values $\{x_i; \quad i = 1 \dots k\}$. The following frequencies can be defined:

- **Absolute frequency** (n_i): number of times that value x_i is observed.
- **Relative frequency** (f_i): fraction of the total number of observations equal to $\boldsymbol{x_i}$ (usually in %):

$$f_i = \frac{n_i}{N}$$

It is always fulfilled that:

$$N = \sum_{i=1}^{k} n_i$$

and

$$\sum_{i=1}^{k} f_i = 1$$

Absolute and relative frequencies can always be obtained, regardless of the type of variable we are dealing with. For those data that present some order, i.e., all except those of nominal type, we can obtain the **cumulative frequencies**:

- **Cumulative absolute frequency** (N_i): number of times that a value less than or equal x_i is observed. It can be expressed as:

$$N_i = \sum_{j=1}^{i} n_j$$
$$\text{or} \quad N_i = N_{i-1} + n_i \quad i > 1$$

The last value of the cumulative absolute frequencies will be:

$$N_k = \sum_{j=1}^{k} n_j = N$$

- **Cumulative relative frequency** (F_i): fraction of the total number of observations less than or equal to xi (usually in %)

$$F_i = \sum_{j=1}^{i} f_j \quad \text{or} \quad F_i = \frac{N_i}{N}$$
$$\text{or} \quad F_i = F_{i-1} + f_i \quad i > 1$$

The last value of the cumulative relative frequencies will be:

$$F_k = \sum_{j=1}^{k} f_j = 1$$

Frequency distribution

The frequency distribution represents the values of a variable and their associated frequencies:

$\{x_i; n_i \quad i = 1, \dots, k\} \rightarrow$ Absolute ordinary frequency distribution

$\{x_i; f_i \quad i = 1, \dots, k\} \rightarrow$ Relative ordinary frequency distribution

$\{x_i; N_i \quad i = 1, \dots, k\} \rightarrow$ Absolute cumulative frequency distribution

$\{x_i; F_i \quad i = 1, \dots, k\}$ → Relative cumulative frequency distribution

Obtaining this distribution will depend on the type of data we are working with. We will distinguish the frequency distribution for the following types of variables:

- Nominal qualitative variables.
- Ordinal qualitative variables.
- Discrete (or non-grouped) quantitative variables.
- Continuous (or grouped) quantitative variables.

The calculation of the frequency distribution of ordinal qualitative variables and discrete (or non-grouped) quantitative variables is similar.

FREQUENCY DISTRIBUTION FOR NOMINAL QUALITATIVE VARIABLES

For this type of distributions cumulative frequencies are meaningless. We will provide a frequency table with the values of the variable (x_i) and the associated ordinary frequencies (n_i and f_i):

X	n_i	f_i
x_1	n_1	$f_1 = n_1/N$
x_2	n_2	$f_2 = n_2/N$
⋮	⋮	⋮
x_k	n_k	$f_k = n_k/N$
$\sum$	N	1

Example: Activity sector of companies in the province of Zaragoza (December 2020). Source: Ministry of Employment and Social Security

Sector	Companies	Relative Freq.
x_i	n_i	$f_i = n_i/N$
Agriculture	1,811	6.5%
Industry	2,984	10.7%
Construction	2,424	8.7%
Services	20,582	74.0%
N=	27,801	100.0%

FREQUENCY DISTRIBUTION FOR ORDINAL QUALITATIVE VARIABLES AND DISCRETE (OR NON-GROUPED) QUANTITATIVE VARIABLES

We will assume that the values are ranked: $x_1 < x_2 < \ldots < x_k$. The frequency table will include now the cumulative frequencies:

X	n_i	f_i	N_i	F_i
x_1	n_1	$f_1 = n_1/N$	$N_1 = n_1$	$F_1 = f_1$ ó $F_1 = N_1/N$
x_2	n_2	$f_2 = n_2/N$	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$ ó $F_2 = N_2/N$
⋮	⋮	⋮	⋮	⋮
x_k	n_k	$f_k = n_k/N$	$N_k = n_1 + \ldots + n_k = N$	$F_k = f_1 + \ldots + f_k = 1$ ó $F_k = N_k/N = 1$
$\sum$	N	1		

Example for an Ordinal Qualitative Variable: Type of medical care of the cases confirmed by COVID in Zaragoza on 16/05/2020.

This is an ordinal variable (order according to severity of the type of medical care). The frequencies distribution, following the general notation, are expressed as:

Type of medical care	*Cases*			
x_i	n_i	f_i	N_i	F_i
No hospitalisation	3,195	60.43%	3,195	60.43%
Normal hospitalisation	1,901	35.96%	5,096	96.39%
ICU hospitalisation	191	3.61%	5,287	100.00%
	5,287	100.00%		

Example for a Discrete (or non-grouped) Quantitative Variables: Number of children per family. Source: Survey of 1,200 families.

This is a discrete quantitative variable with few different values. The frequencies distribution, following the general notation, are expressed as:

Number of children	*Absolute frequencies (number of families)*	*Relative frequencies*	*Cumulative absolute frequencies*	*Cumulative relative frequencies*
x_i	n_i	f_i	N_i	F_i
0	213	17.75%	213	17.75%
1	255	21.25%	468	39.00%
2	375	31.25%	843	70.25%
3	194	16.17%	1037	86.42%
4	83	6.92%	1120	93.33%
5	54	4.50%	1174	97.83%
6	20	1.67%	1194	99.50%
7	6	0.50%	1200	100.00%
	1200	100.00%		

FREQUENCY DISTRIBUTION FOR CONTINUOUS (OR GROUPED) QUANTITATIVE VARIABLES

If the number of different values is too high, it is usual to group data into mutually exclusive classes showing the number of observation in each class. To do this, it is necessary to calculate both the classes and a point value representing each of the classes.

The classes are represented by the expression $(L_{i-1}, L_i]$ where L_{i-1} is the lower limit of the class, and Li is the upper limit of the class. Then, a representative value for each of them is determined, called the *class midpoint* (x_i), defined as the halfway between the limits of the class:

$$x_i = \frac{L_{i-1} + L_i}{2}$$

A variable with grouped data will receive the same treatment as one with non-grouped data.

If the classes are not all of the same width, a new characteristic of the population must be calculated, its *frequency density*, which can be different for each class, and is defined as:

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

where a_i represents the class width and is calculated as $a_i = L_i - L_{i-1}$.

It is also possible to calculate the density using the relative frequency:

$$d_i = \frac{f_i}{a_i}$$

The frequency table will include all these elements:

Class	**Class midpoint**	**Width**	**Frequency density**	**Frequencies**			
$(L_{i-1}, L_i]$	x_i	a_i	d_i	n_i	f_i	N_i	F_i
$(L_0, L_1]$	x_1	a_1	d_1	n_1	f_1	N_1	F_1
$(L_1, L_2]$	x_2	a_2	d_2	n_2	f_2	N_2	F_2
…	…	…	…	…	…	…	…
$(L_{k-1}, L_k]$	x_k	a_k	d_k	n_k	f_k	N_k	F_k
Total				N	1		

Example: Number of workers in Spanish companies (dated 31/12/20). Source: Ministry of Employment and Social Security

Number of workers	Number of companies
From 1 to 2	697,528
From 3 to 5	290,574
From 6 to 9	131,113
From 10 to 49	147,851
From 50 to 249	23,834
From 250 to 499	2,550
More than 499	2,206

- This is a discrete quantitative variable
- The number of values that the variable may present is very high.
- This makes it necessary to group the values of the variable by intervals or ranges (classes).

Example: Monthly salary of the workers of a company

Monthly salary (€)	Number of workers
500-1000	50
1000-1500	150
1500-2000	200
2000-2500	90
2500-3000	10

- This is a continuous quantitative variable
- Then, it is necessary to group the values of the variable by intervals or ranges (classes).

L_{i-1}	L_i	n_i	x_i	f_i	N_i	F_i	a_i	d_i
500	**1000**	**50**	750	10%	50	10%	500	0.1000
1000	**1500**	**150**	1250	30%	200	40%	500	0.3000
1500	**2000**	**200**	1750	40%	400	80%	500	0.4000
2000	**2500**	**90**	2250	18%	490	98%	500	0.1800
2500	**3000**	**10**	2750	2%	500	100%	500	0.0200
TOTAL		**N=500**		100%				

GRAPHIC PRESENTATIONS

The graphical representation of the data will depend, again, on the type of variable with which we are working.

For Nominal Qualitative Variables:

- Cartograms, Pictograms.
- Pie Charts.
- Bar Charts.

For Ordinal Qualitative and Discrete Quantitative Variables:

- Bar Charts.
- For Continuous (or grouped) Quantitative Variables.
- Histogram.
- Frequency Polygon.

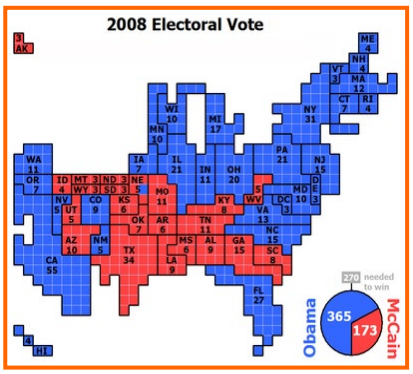
Other Graphic Presentations:

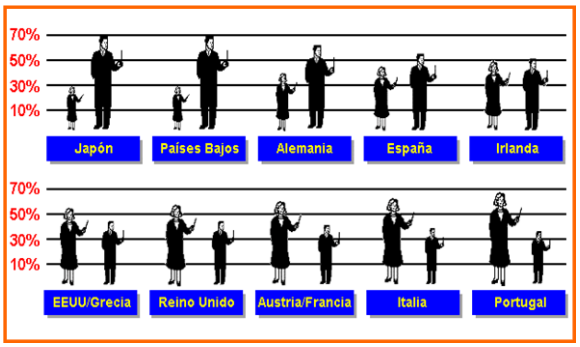
- Stem-and-Leaf Plots.

GRAPHIC PRESENTATIONS FOR QUALITATIVE VARIABLES

Cartograms and Pictograms

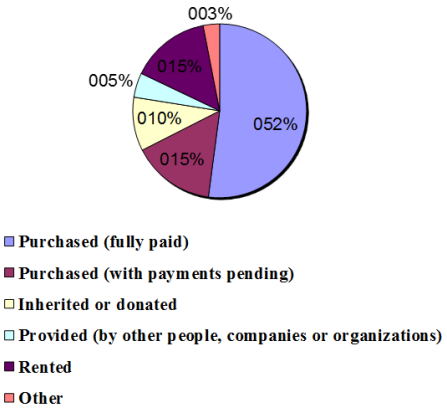
The first of these charts is a cartogram (presenting information on the elections 2008 in the USA) and the second is a pictogram (presenting data on the salaries of teachers for different countries distinguishing by gender)





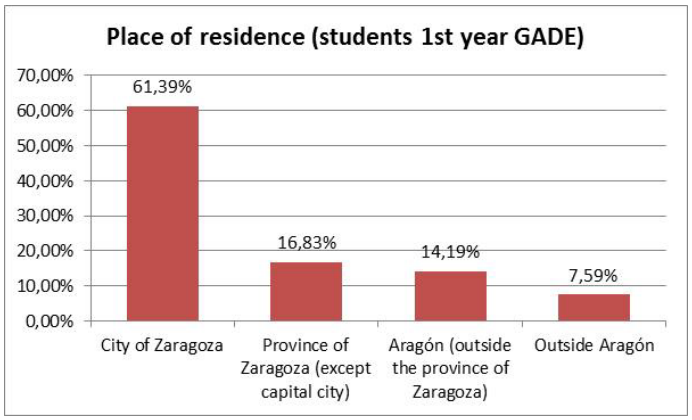
Pie charts

They are particularly suitable for nominal qualitative variables.

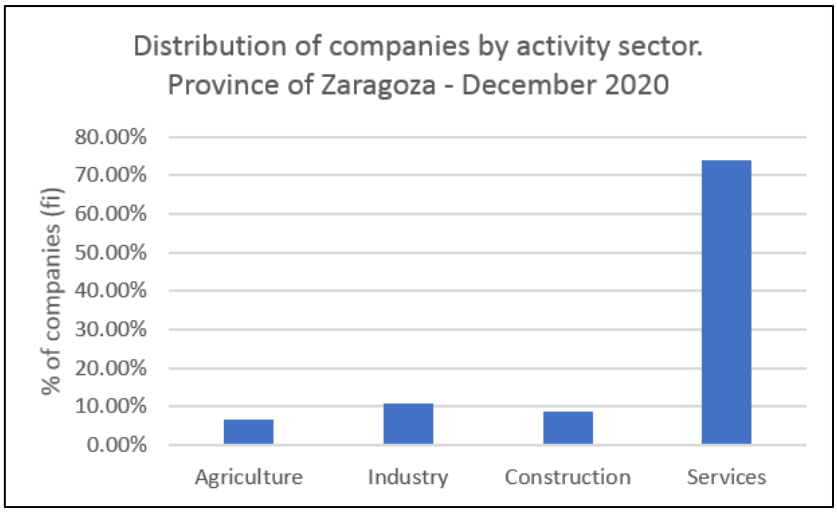


Bar charts

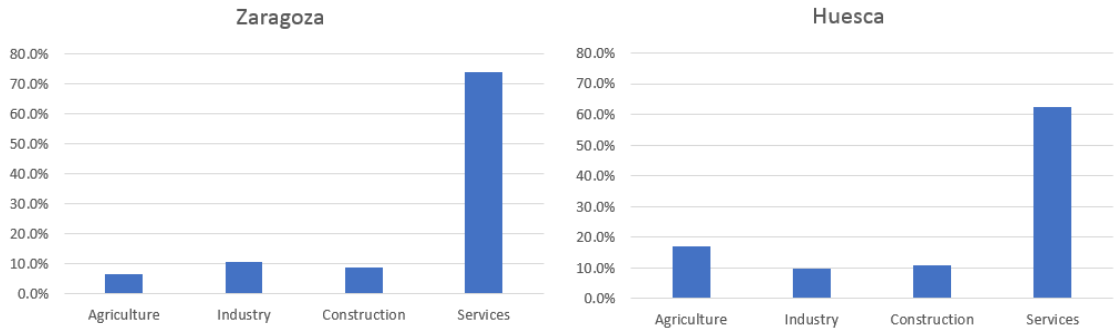
On a pair of axes, the different classes of the variable are represented on the X-axis and, for each of them, a bar with a height proportional to the frequency (absolute or relative) of that class is presented. This is the ideal representation for **ordinal** qualitative variables.



In the case of **nominal** variables, the X-axis represents each of the values of the variable, regardless of the order followed:



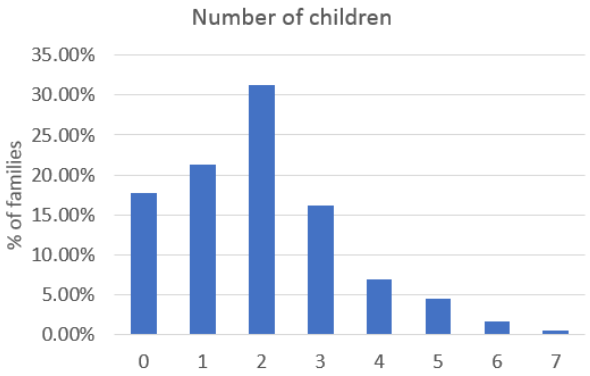
On the Y-axis we can represent absolute or relative frequencies, but if we have to compare distributions using a graphic presentation, we must use relative frequencies:



GRAPHIC PRESENTATIONS FOR QUANTITATIVE DISCRETE VARIABLES

Bar Charts

Number of children	*Absolute frequencies (number of families)*	*Relative frequencies*
x_i	n_i	f_i
0	213	17.75%
1	255	21.25%
2	375	31.25%
3	194	16.17%
4	83	6.92%
5	54	4.50%
6	20	1.67%
7	6	0.50%
	1200	100.00%



- Both absolute and relative frequencies can be used.
- *Only* use *relative frequencies* if the data are to be *compared* with those of another population.

GRAPHIC PRESENTATIONS FOR CONTINUOUS (OR GROUPED) QUANTITATIVE VARIABLES

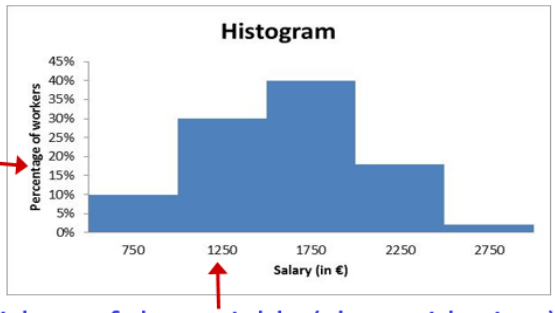
Histograms

A histogram is a graphical representation of a variable using bars, where the area of each bar is proportional to the frequency of the values represented. On the Y-axis the frequency densities (or frequencies in the case of intervals of the same width) are represented, and on the X- axis the intervals of the variables, indicating the class midpoint.

The X-axis has to reflect the continuity of the values of the variable (xi) and the widths of the bars have to coincide with the width of each interval. Be **careful** if classes **widths** are **not equal.** It is worth noting that the height should be proportional to the density.

- If the widths (a_i) are the **same** for all the intervals

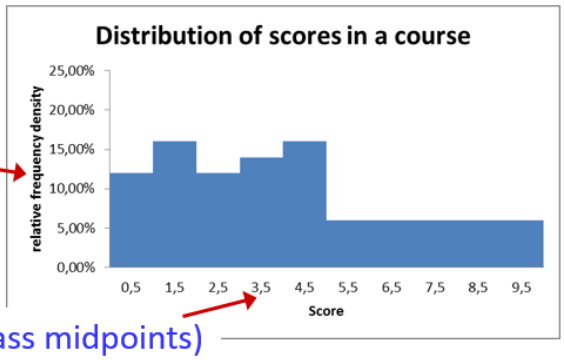
Absolute Frequencies (n_i)
or
Relative Frequencies (f_i)
(preferably)
or
Frequency Densities (d_i)



Values of the variable (class midpoints)

- If the widths (a_i) are **NOT** the same for all the intervals

Frequency Densities (d_i)
(preferably the relative ones)



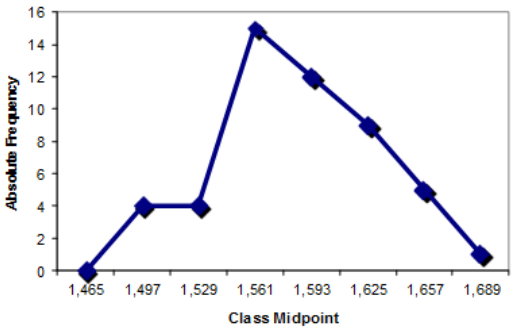
Values of the variable (class midpoints)

Frequency Polygons

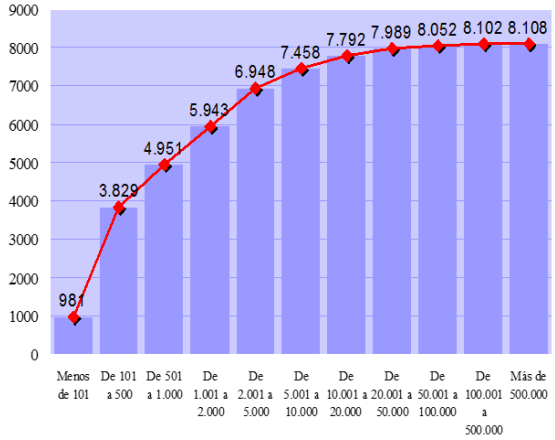
The frequency polygon is a graphical representation of the frequency distribution that is essentially equivalent to the histogram, with the following characteristics:

- The pairs **(x_i, d_i)** are plotted and joined by a **polygonal line**.
- If the **intervals** have the **same width**, it is possible to choose to represent the pairs (x_i, **n_i**) or (x_i, **f_i**).

Ordinary Frequency Polygons

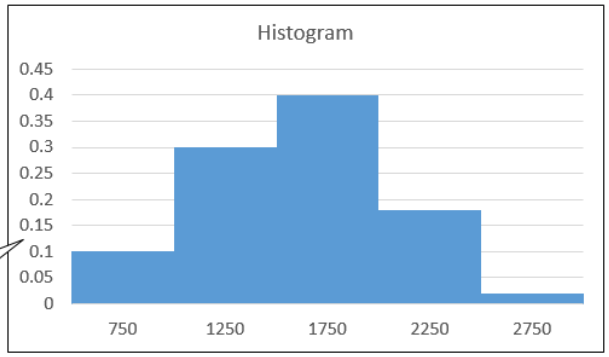


Cumulative Frequency Polygons

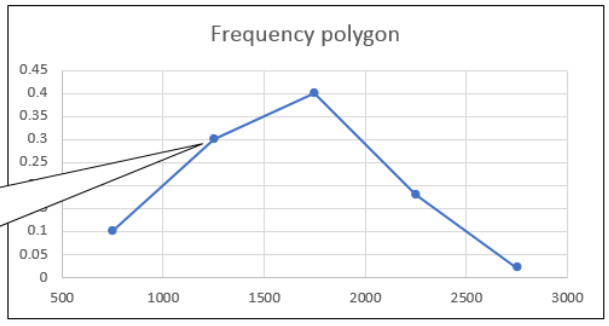


Example: Monthly salary of the workers of a company

Monthly salary (€)	Number of workers	a_i	d_i
500-1000	50	500	0.10
1000-1500	150	500	0.30
1500-2000	200	500	0.40
2000-2500	90	500	0.18
2500-3000	10	500	0.02



Frequency densities have been used, but frequencies could have been employed, since all classes have the same width.



It can be seen that the frequency polygon is obtained by joining the midpoints of the upper part of each bar of the histogram.

The same precautions given for histograms should be taken into account with frequency polygons:

- If the widths (ai) are NOT the same: use frequency densities (preferably the relative ones)
- To compare distributions: use the relative frequency densities (di=fi/ai)

When the widths are different, it is too complex to construct a histogram with Excel. We will then use a frequency polygon to present the distribution.

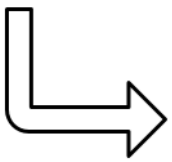
Cumulative Frequency Polygons are used to present graphically grouped quantitative variables. They are created in a similar way to histograms and frequency polygons but using cumulative frequencies.

Stem-and-Leaf Plots

They can be used for quantitative variables, either discrete or continuous.

Example: heights of a group of children.

1.56	1.59	1.63	1.62	1.65
1.61	1.59	1.51	1.62	1.62
1.53	1.49	1.57	1.54	1.53
1.59	1.58	1.57	1.47	1.64
1.55	1.59	1.53	1.56	1.53
1.47	1.57	1.60	1.54	1.56
1.50	1.62	1.59	1.62	1.54
1.68	1.52	1.62	1.59	1.49
1.65	1.53	1.59	1.56	1.54
1.58	1.52	1.63	1.56	1.62



Frequency	Stem & Leaf
4	14 . 7799
13	15 . 0122333334444
18	15 . 566666777889999999
12	16 . 012222222334
3	16 . 558

SUMMARY

Variable types	Tabulation	Graphic presentation
Binary (nominal with 2 values)	Non-Grouped Frequencies	Pie or bar charts
Nominal with more than 2 values	Non-Grouped Frequencies	Pie or bar charts
Ordinal	Non-Grouped Frequencies	Bar charts
Quantitative discrete with few values	Non-Grouped Frequencies	Bar charts
Quantitative discrete with many different values OR continuous	Grouped Frequencies	Histograms or Frequency polygons