UNIVERSIDAD CENTRAL

Ingeniería de Sistemas

Prof. Hugo Franco

Session N ° 10 | Exploratory Analysis

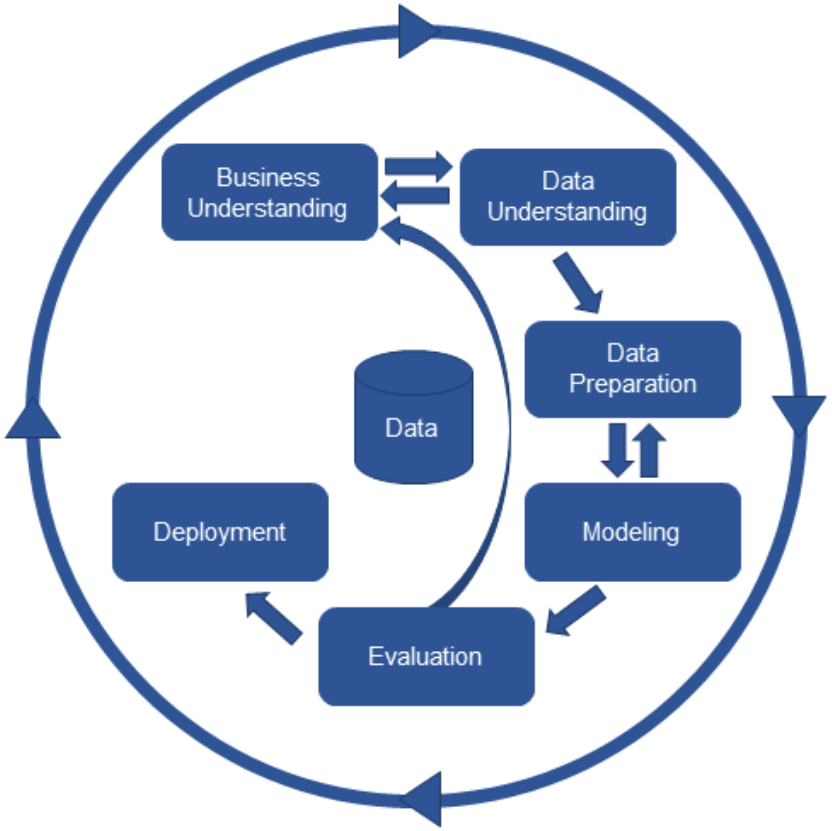
Descriptive Statistics

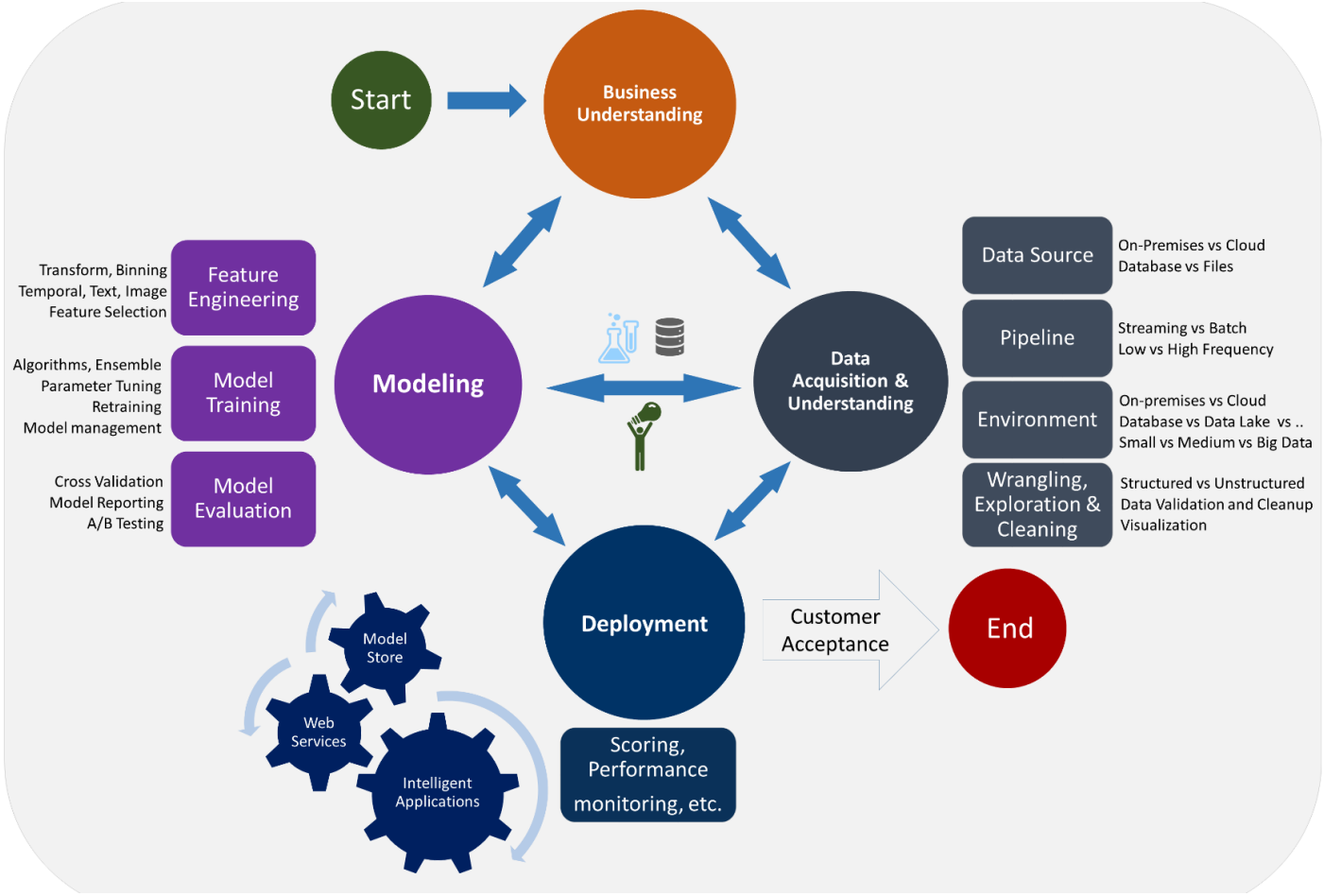
Bogotá D.C., October 4th, 2022

Review: Data Analytics Workflow

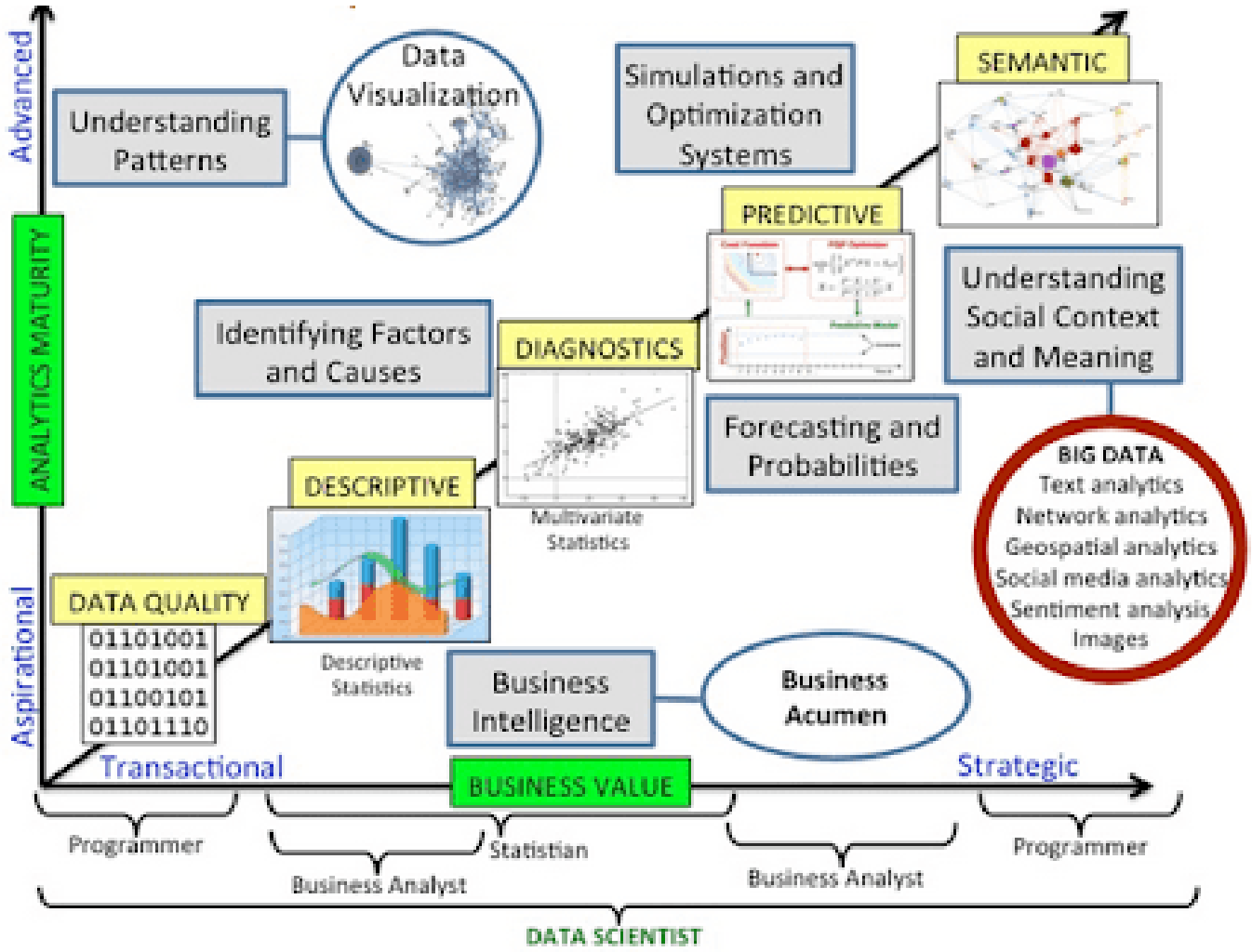
Modeling data

CRISP-DM (1999) vs Microsoft TDSP (2016)





Data analytics: professional scope



Modeling data I: Statistics

Descriptive Statistics

Statistics and Data Analytics

- Data analytics has existed from the beginning of the experimental version of science:
  - The hypothetical deductive method introduced the “falsifiability” of scientific theories: experience can support or reject according to observations (data)
- The advent of Computer Science allowed the systematic (computerized) analysis of large amounts of data for Enterprise-scale and Nation-scale applications
  - Several statistical processes benefited of computer-based implementations

What is statistics?

- **Statistics** is the science that explains and provides tools to work with data.
  - It has experienced a fast development over the last few decades.

Applications:

- **Statistics is currently applied in all areas of knowledge**, e.g., in Biology, Physics research, Environmental Sciences, Computer Science (e.g., *Machine Learning*), Ecology, Sociology, Education, Psychology, Administration, Economics, Medicine, and Political Science, among others.

Application examples

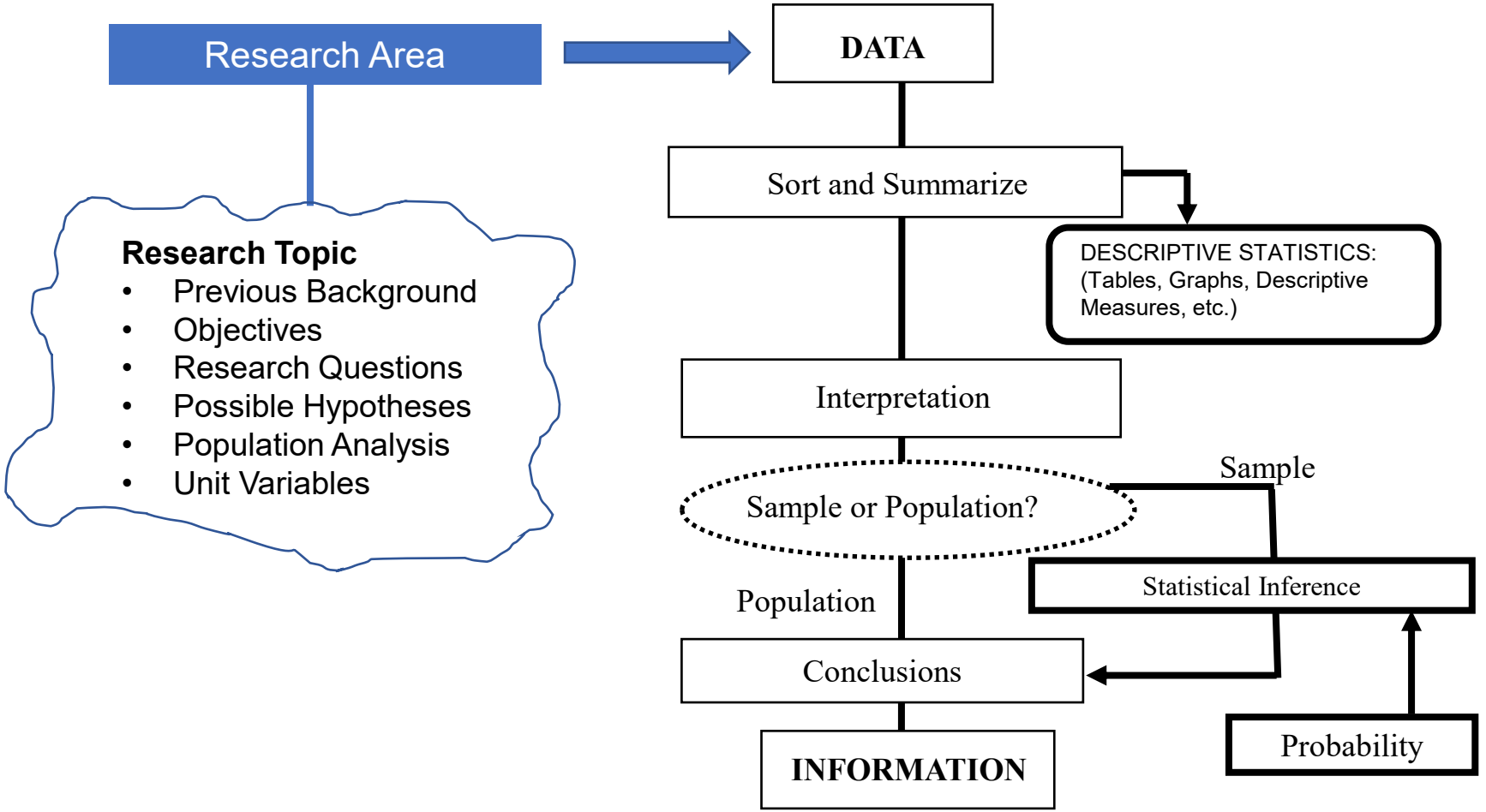
1. *In Business Administration:* statistics are used to evaluate a product before marketing it.
2. *In Economics*: to measure the evolution of prices through index numbers or to study the habits of consumers through surveys of family budgets.
3. *In Politics**: to know the preferences of the voters before a vote through polls and thus guide the strategies of the candidates.
4. *In Sociology**: to study the opinions of social groups on current issues.
5. *In Psychology*: to elaborate the scales of the tests and quantify aspects of human behavior (for example the tests that are applied to candidates for a position in a company).
6. *In Medicine*: one among many uses of statistics, is to determine the health status of the population (Epidemiology – Public Health).
7. *In Artificial Intelligence*: to evaluate the performance of a machine learning model, according to the difference between its actual behavior (after training) and the expected output (knowledge base, training / testing sets).

Phases of a statistical study

Statistical Analysis is carried out following the usual phases in the so-called scientific method whose phases are:

1. **Problem statement:** define the objective of the research and specify the universe or population.
2. **Information collection:** collect the necessary data related to the research problem.
3. **Descriptive analysis:** summarize the available data to extract the relevant information in the study.
4. **Statistical inference:** assume a model for the entire population based on the data analyzed to obtain general conclusions.
5. **Diagnosis:** evaluate the validity of the assumptions of the model that have allowed us to interpret the data and reach conclusions about the population

Statistical study workflow



Sample problems

1) To study whether in a certain group there is wage difference due to the sex of the person employed (discriminative analysis).

2) To determine the profile of workers in terms of economic and social conditions in different communities (distributions, characterization).

3) To study the consumption of people in a certain area in terms of clothing, food, leisure and housing (proportions).

4) To determine the standard sizes in clothing and shoes in several countries (central tendency measures, dispersion).

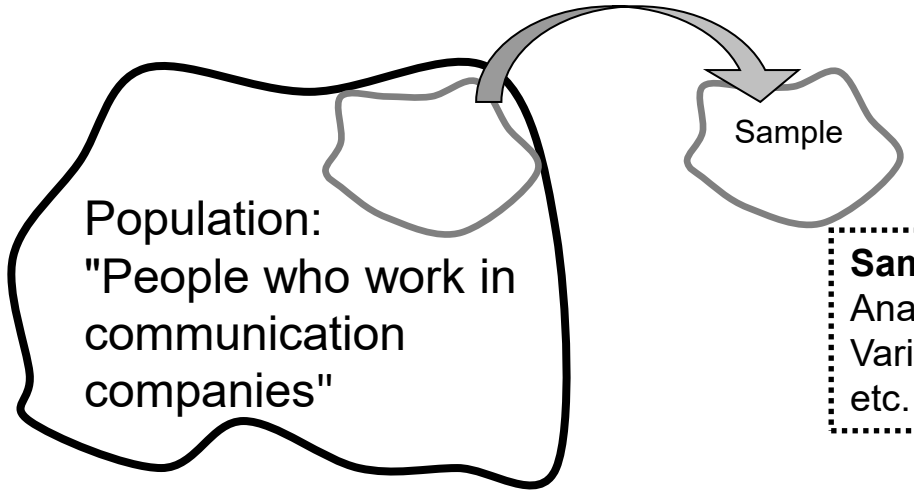
5) To determine the time spent at work and family by workers in different companies in the country (distributions).

6) To determine the sociodemographic profile of the students at a University (distributions, characterization).

7) To study the monthly mobile phone expenditure of the students at a University, and if it has any relationship with their age or other characteristics (inferential statistics).

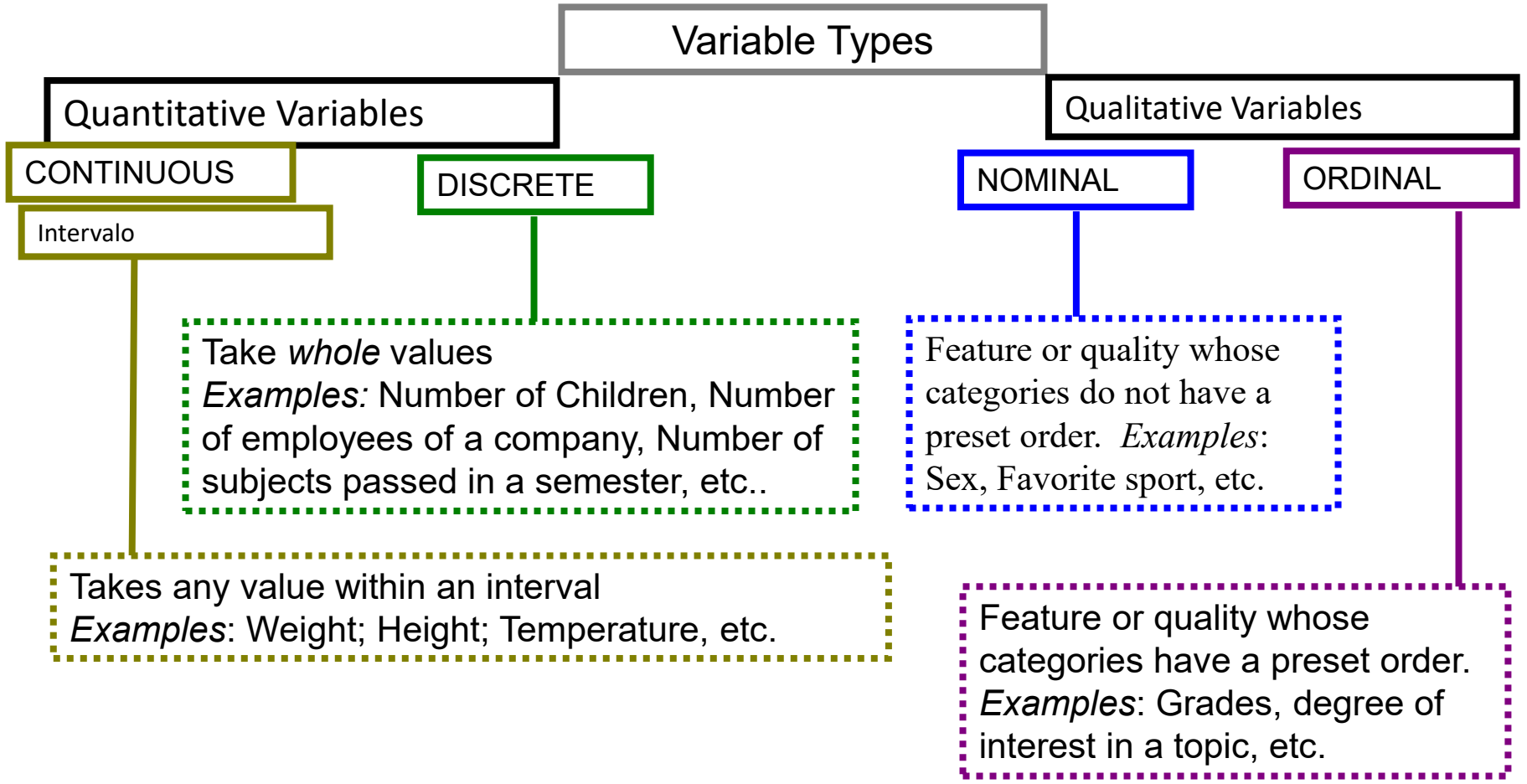
Descriptive statistics: main concepts

- **Variable**: it is what is going to be measured and represents a characteristic of the UNIT OF ANALYSIS.
  - The subjects or objects are the Units of Analysis within a Population or a Sample:
    - POPULATION: it is the total of units of analysis that are the subject of study.
    - SAMPLE: it is a set of units of analysis coming from a population**.**



Sample: 60 workers of communication companies
Analysis unit: Worker of communication company
Variables: sex, age, salary, Number of working hours, etc.

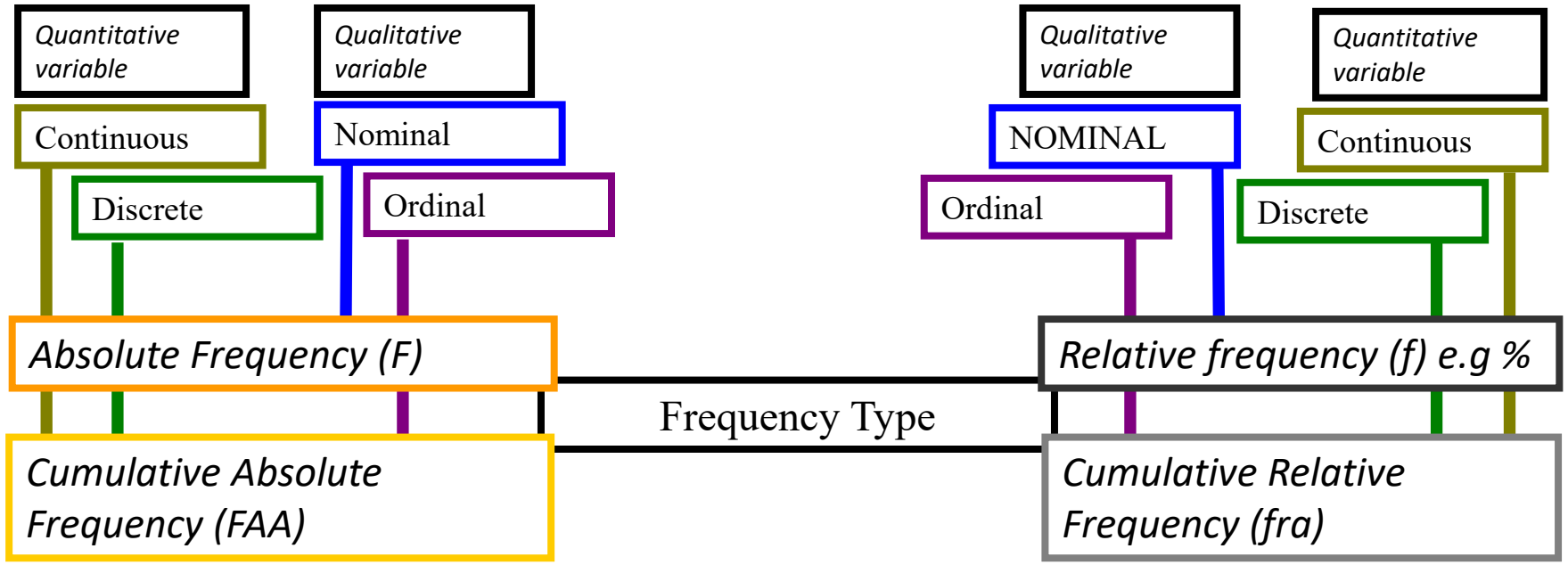
Variables as characteristics of the Analysis Unit



Unit of Measurement: Grams or Kilos for the variable Weight; Degrees C or F for the variable Temperature*

Frequency

- Given a set of units, it corresponds to the Number or Percentage of times a feature is presented.



Example

- Research Problem: to establish the profile of the car assembly companies based on available features.
  - Unit of Analysis: car assembly industry
  - Population: assembly companies in Colombia

Variables

- Type of Industry: classified into industry type A (manufacturing), B (assembly), C (importer) or D (sales); (nominal qualitative)
- Number of Employees: refers to the number of employees in the production lines. (discrete quantitative)
- Area: refers to the square meters (unit of measurement) available for production areas. (continuous quantitative)
- Rating: rating made by a public institution on compliance with certain standards (Very Good, Good, Regular, Bad). (ordinal qualitative)

Data

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

Example: Frequency Tables

- **Research Problem:** to establish the profile of the car assembly companies based on available features.
  - Unit of Analysis: car assembly industry
  - Population: assembly companies in Colombia

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

(1)

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)		

(2)

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%)		

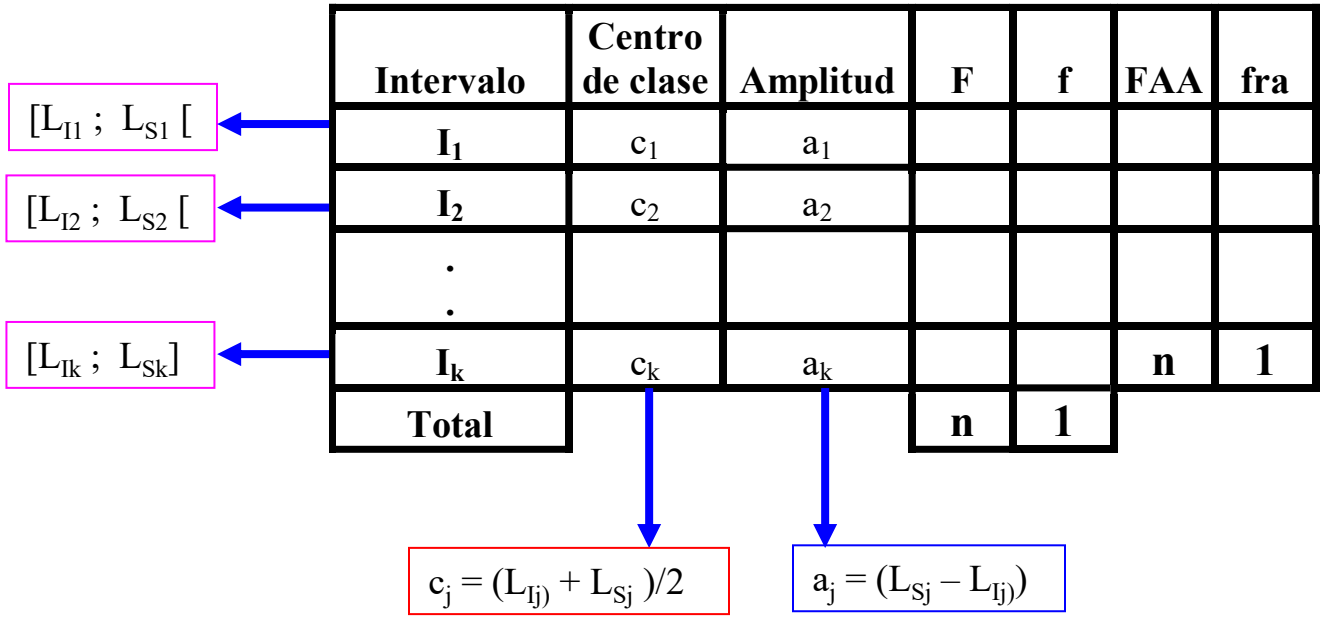
(3)

(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%)		

Example

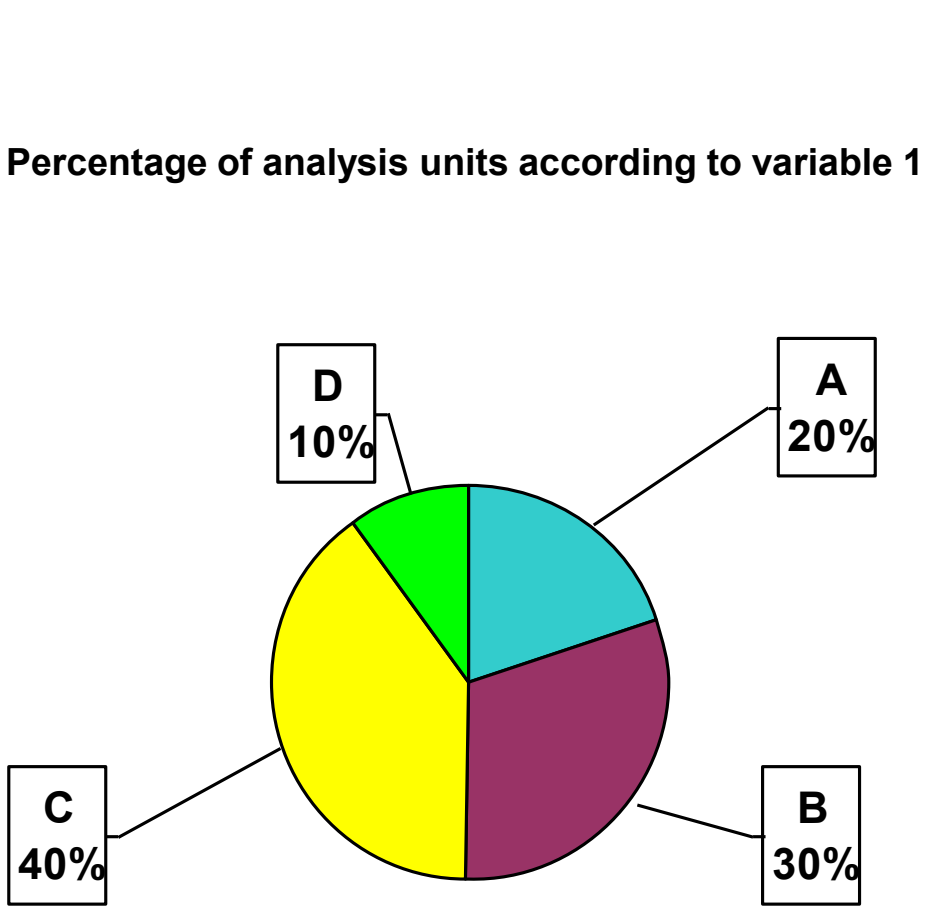
- Elements of a frequency table when the variable is continuous

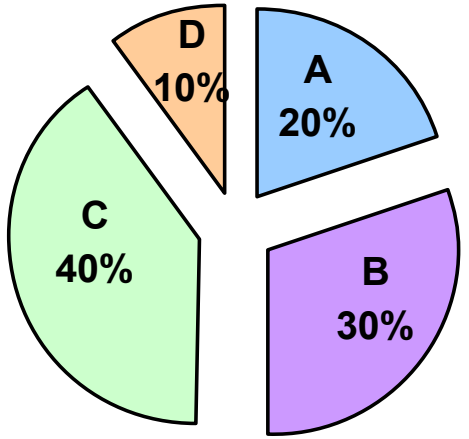


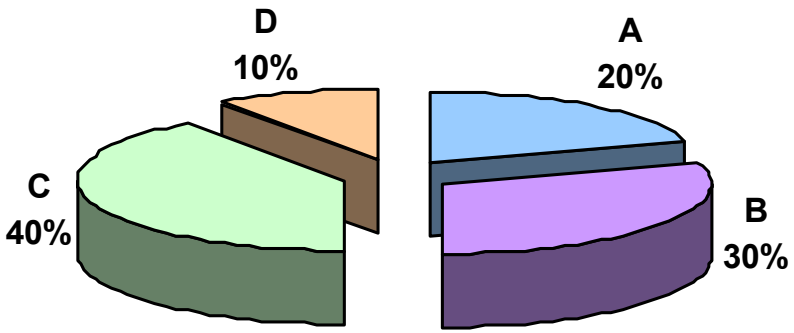
Intervalo	Centro de clase	Amplitud	F	f	FAA	fra
I_1	c_1	a_1				
I_2	c_2	a_2				
. .						
I_k	c_k	a_k			n	1
Total			n	1		

Data visualization as a descriptive tool

Plots: Pie Chart (sector)

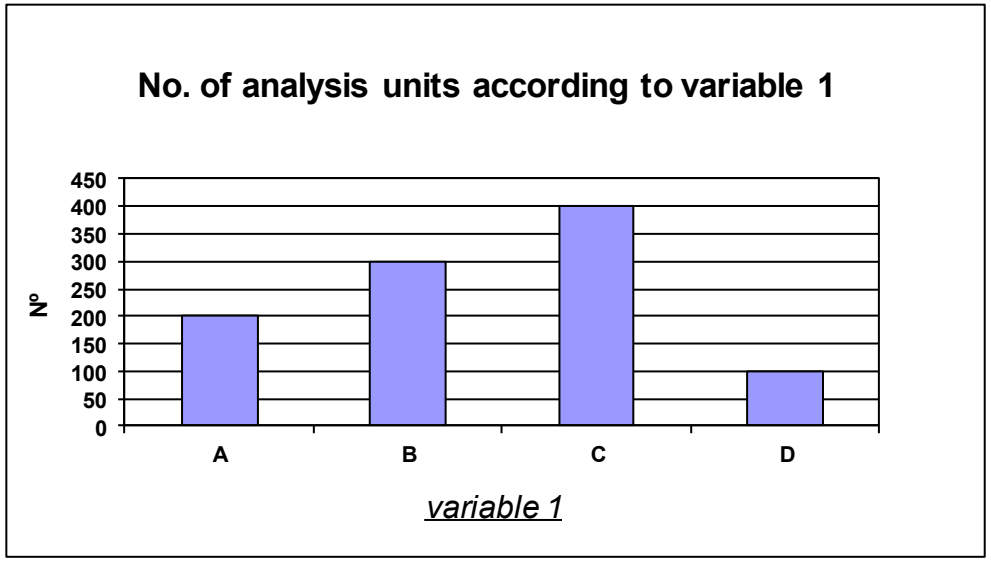


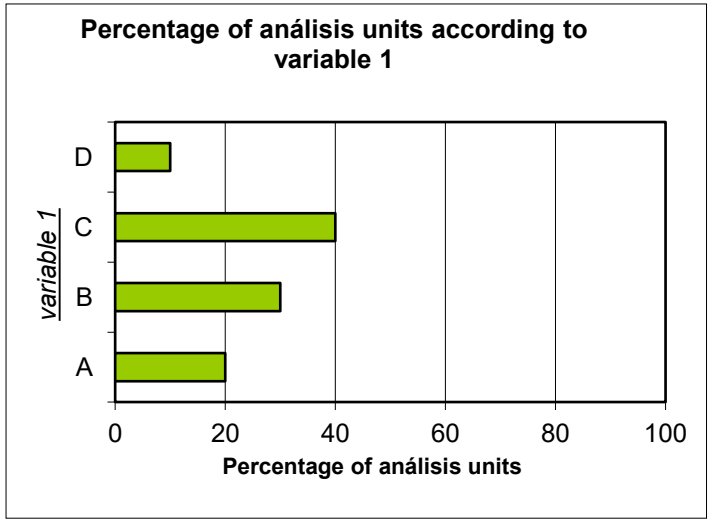


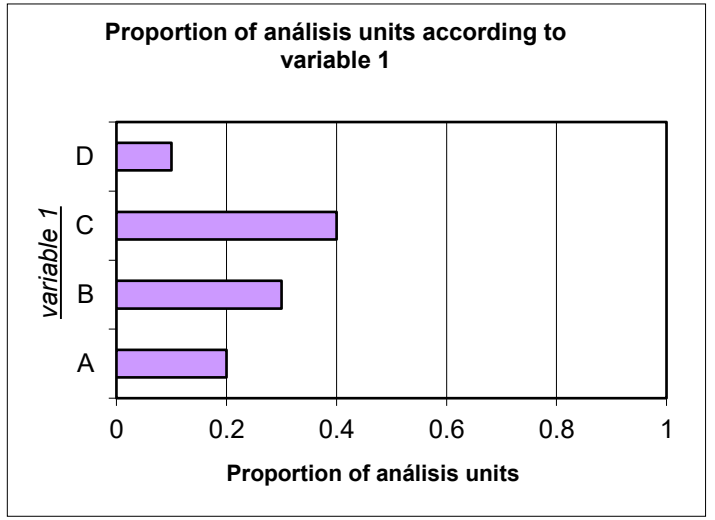


Bar plot

- This type of graph is generally used to represent the frequency of categories of a qualitative variable.
- When a variable is quantitative, this type of graph can be used only if the variable has been transformed into categories.
- There are different versions of these charts (for example in Excel), and in some cases they are very useful for describing the behavior of a variable in different groups.

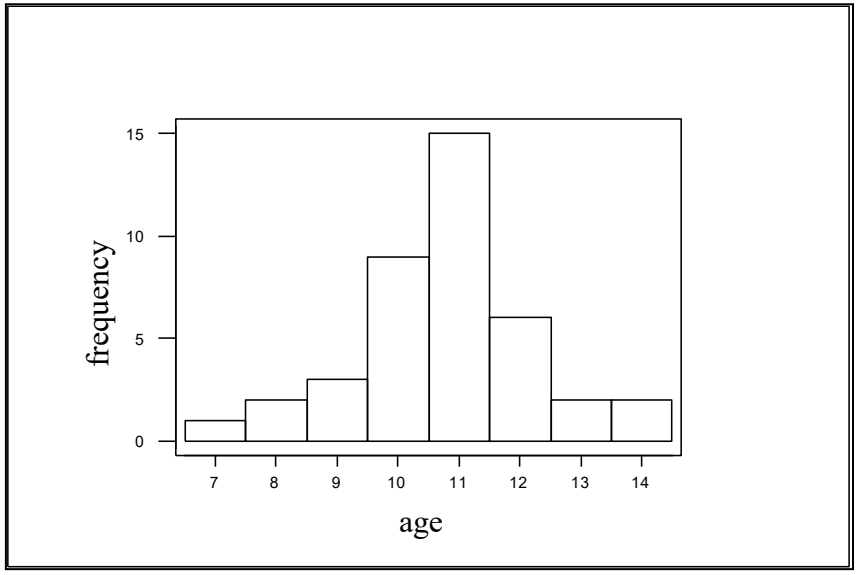






Histogram (descriptive distribution)

Histogram Distribution of the children of company workers according to their age



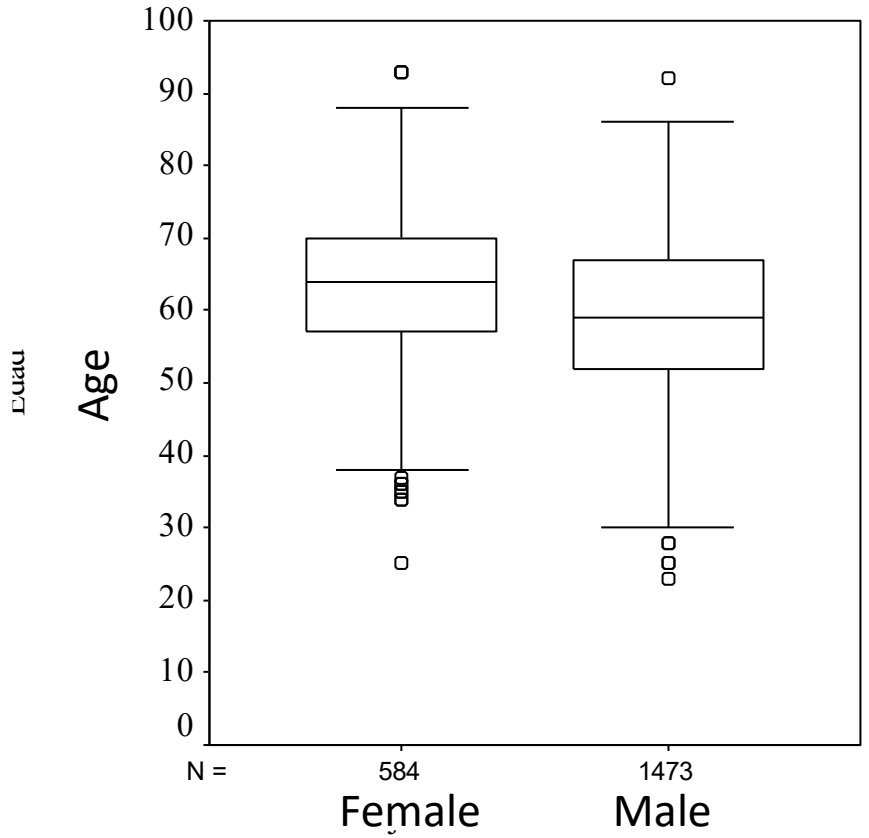
According to the plot, the number of children, lies between 7 and 14 years); and most of the workers' children are between 10 and 12 years old.

- Allows the representation of the frequency of a Quantitative variable.
  - The x-axis refers to the variable.
  - The y-axis refers to the frequency (No., %, …).
- Each bar represents the frequency of the variable in the study population (or sample).
- The histogram can be constructed from the data in the frequency table of the variable under study.

Frequency Polygon

- This representation is based on the Histogram.
- It is only useful for quantitative variables.
  - The *x*-axis refers to the variable.
  - The *y*-axis refers to the frequency (number, %).
- The points that allow the lines to join represent the class center (or class mark).

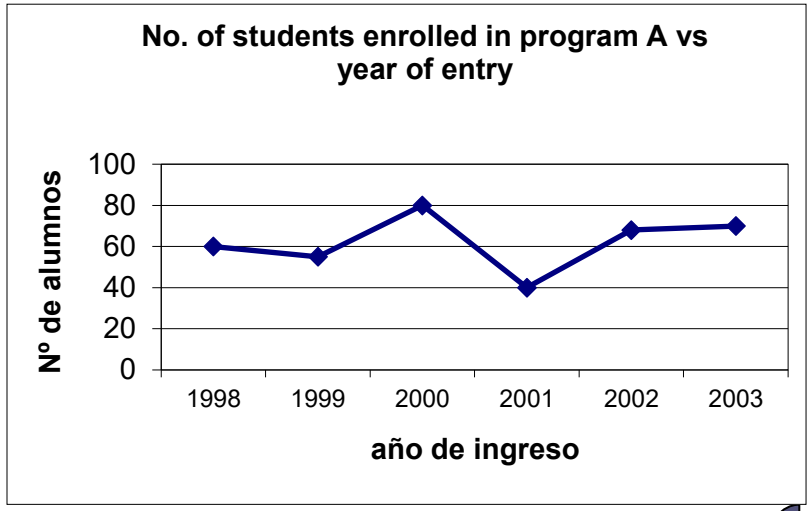
Box plot

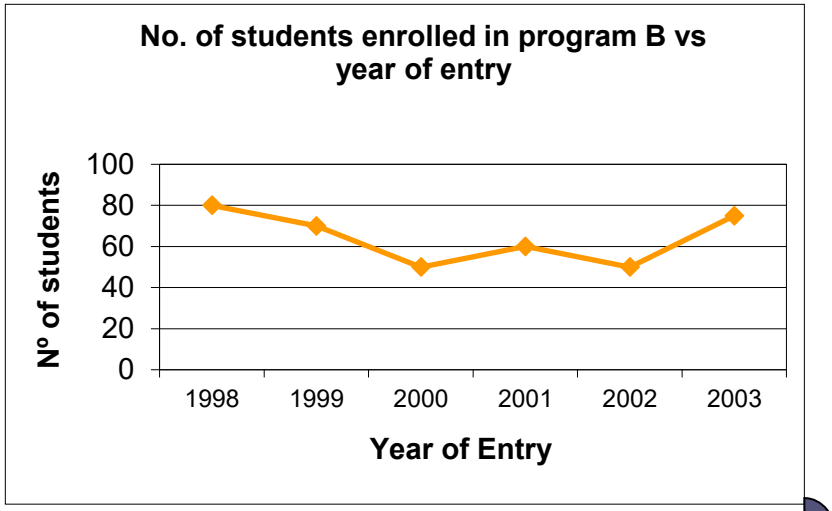


Age of angioplasty patients between 1980 and 2000

- Box plots allow graphically identifying the median (horizontal line within the box), quartiles 1 and 3 (25th and 75th percentiles).
  - The outliers are detected according to the "interquartile range": the difference between the 3rd and the 1st quartiles, i.e., the "whiskers" do not extend to the maximum and minimum values in the data.
- They are only useful for quantitative variables.
- Their *x*-axis makes it possible to identify the population (and its groups) under study.
- Their *y*-axis represents the values of the variable under study.

Dot-Line plot (evolution)





	Number of students	
Number of students	Program A	Program B
1998	60	80
1999	55	70
2000	80	50
2001	40	60
2002	68	50
2003	70	75

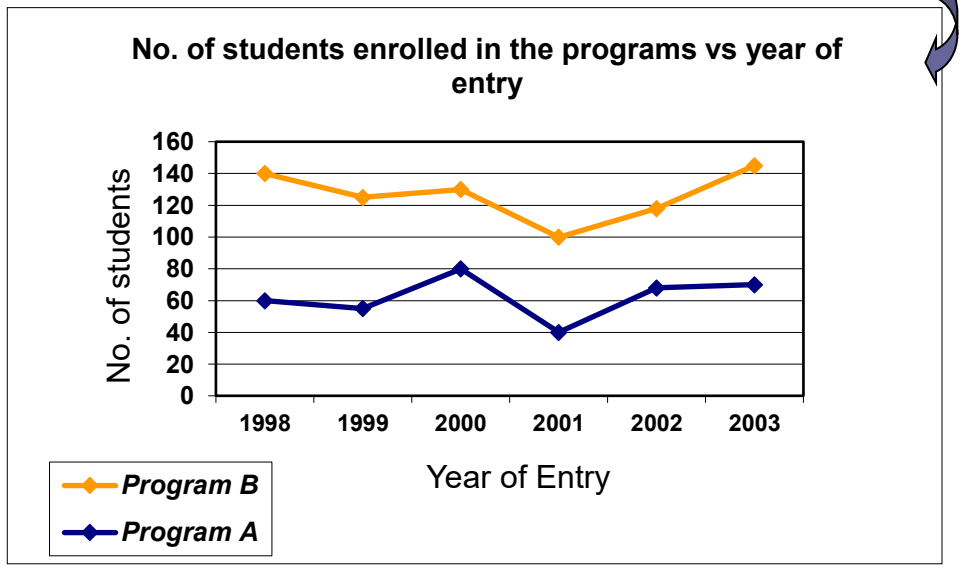


Chart usage and best practices

- The selected chart type will depend on the variable under study.
- The chart must contain a General Title and the identification of each axis:
  - Variable under study and frequency.
- Sometimes a chart is more illustrative than a frequency table.
- Like tables, graphs should be self-explanatory.

Descriptive statistics I

Data interpretation

Notation for quantitative variables

x = variable $\quad x_i$ = variable value for the subject i

y = variable $\quad y_i$ = variable value for the subject i $\quad i = 1, \ldots, n$

a, b, c: constants

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

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

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

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

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

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

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

Central trend measures

- Aritmethic mean (average)
- Median
- Mode

Quantitative Data

x
x_1
x_2
$\vdots$
x_n

Aritmethic mean of average

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

Ordered quantitative data (lower to higher values)

x
$x_{(1)}$
$x_{(2)}$
$\vdots$
$x_{(n)}$

Median

$M_E = x_{(k)}$ If n is odd

$$M_E = \frac{x_{(k)} + x_{(k+1)}}{2}$$ If n is even

$x_{(k)}$ = value in the middle

Quantitative and Qualitative Data

Mode

M_o = the most repeated data

Percentiles, Deciles and Cuartiles

This values are intended to provide a first approach to the value distribution of the variables

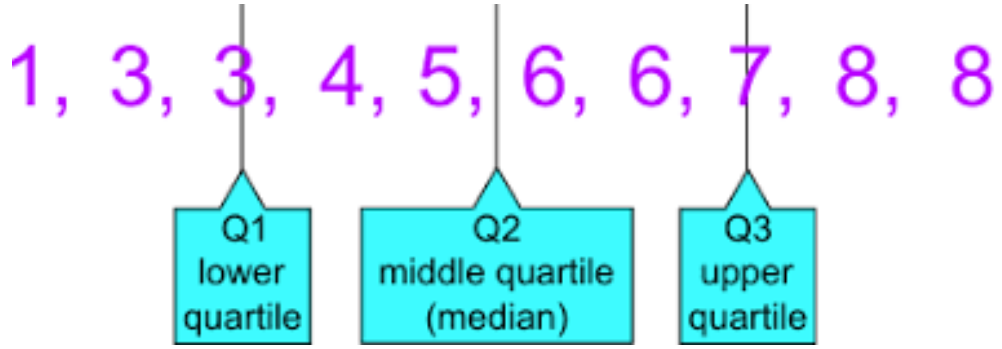
- Percentile (25, 50, 75)
- Decile (4, 5, 8)
- Cuartile (1, 2, 3)

Percentile, Decile and Cuartile (sometimes Quintile) correspond to the value of the quantitative variable, when the *n* data are ordered from smallest to largest

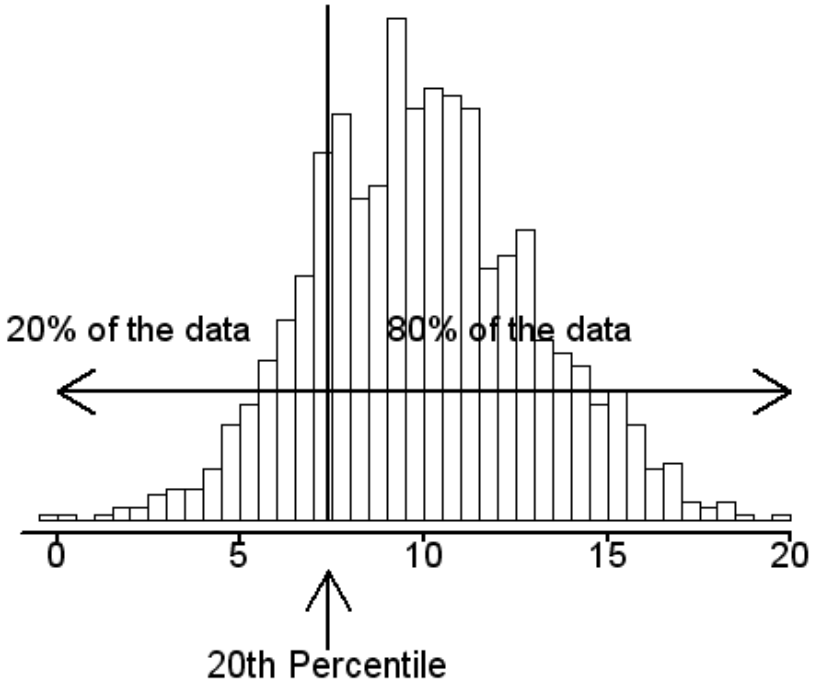
- The Percentile is in the range of 1 to 100
  - The percentile 25 (25/100) is the value of the variable that gathers the 25% of the data, at least
  - Example: If N=80, the 25% of 80 is 20; therefore, the data that is in position 20 is sought.
  - If N=85, 25% of 85 is 21.25; therefore the data that is in position 22 is sought.
- The Decile ranges from 1 to 10
  - The decile 4 (4/10): is the value of the variable that gathers at least 40% of the data
  - Example: If N=80, 40% of 80 is 32; then, the data that is in position 32 is sought.
  - If N=85, 40% of 85 is 34; therefore the data that is in position 34
- The Quartile ranges from 1 to 4
  - The Quartile 3 (3/4) is sought: it is the value of the variable that gathers at least 75% of the data
  - Example: If N = 80, 75% of 80 is 60; therefore, the data that is in position 60 is sought.
  - If N=85, 75% of 85 is 63.75; therefore the data that is in position 64 is sought.

Interpretation

Quartiles



Percentiles



Dispersion measures

- Standard Deviation
- Variance
- Range

Quantitative Data

x
x_1
x_2
$\vdots$
x_n

Range

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

Variance

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

Standard deviation

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

Variable comparison

It refers to the behavior of quantitative variables in a group.

Example: a set of people, measured by Height, Weight, Age. Among these variables, which one presents the greatest variation?

Variation coefficient

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

Higher order distribution descriptors

- Skewness
- Kurtosis

In addition to the position and dispersion of the data, another measure of interest in a frequency distribution is symmetry and pointing or kurtosis.

Skewness

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

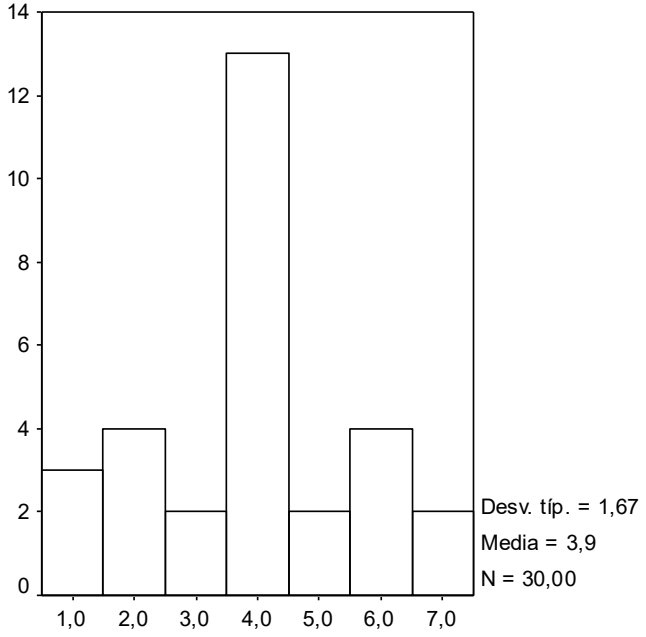
CA=0 if the distribution is symmetric around the mean
CA<0 if the distribution is asymmetric to the left
CA>0 if the distribution is asymmetric to the right

Kurtosis

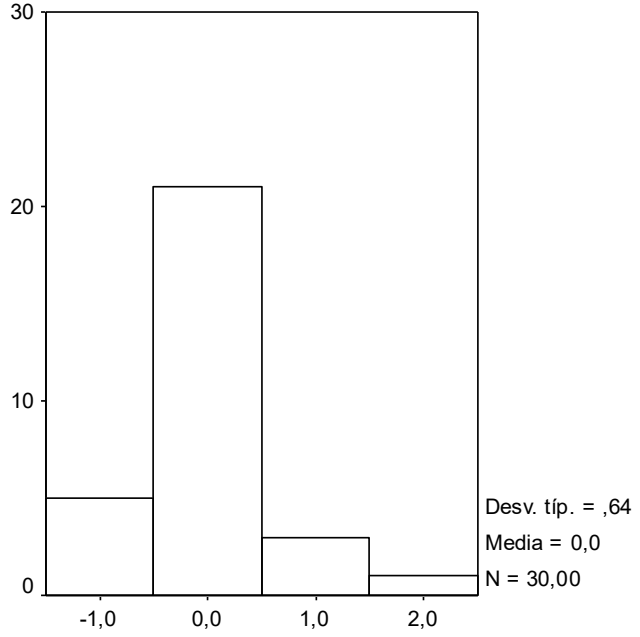
$$CAp = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n \cdot s^4}$$

- If *CAp* = 0 the distribution is said to be normal (similar to the normal Gauss distribution) and is called mesokurtic.
- If *CAp* >0, the distribution is more pointed than the previous one and is called leptokurtic, (greater concentration of the data around the mean).
- If *CAp*<0 the distribution is flatter and is called platykurtic.

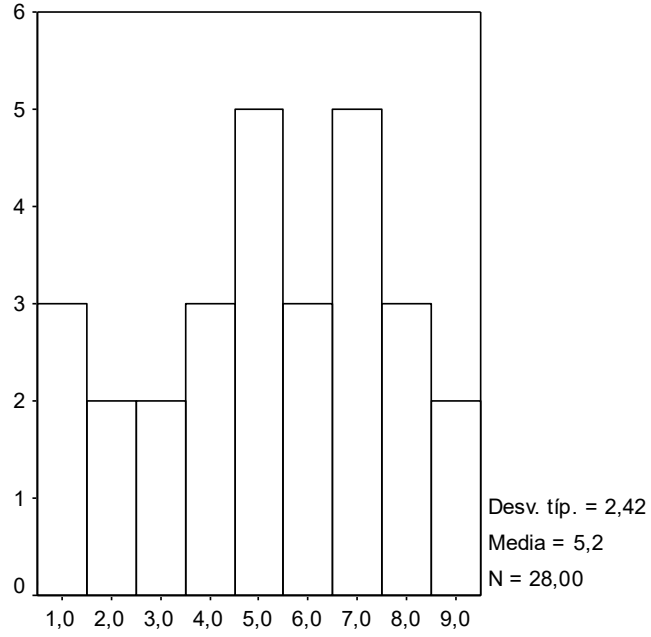
Histogram examples with different bias and kurtosis



V2



V4



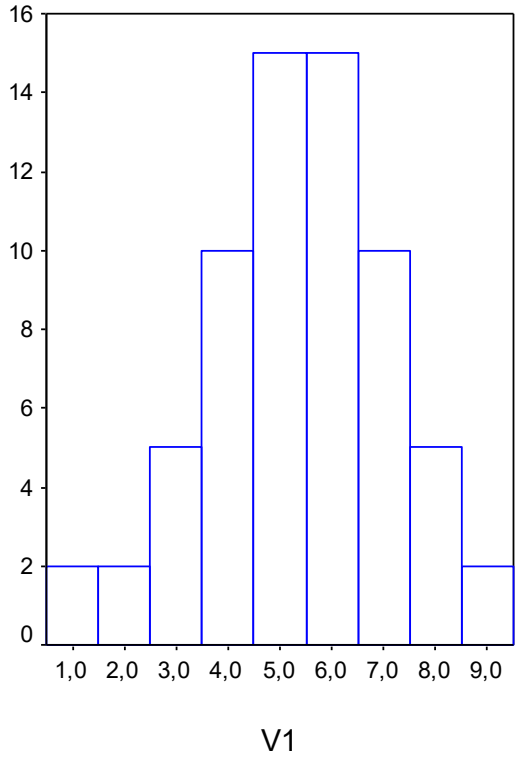
V5

Example for a particular set of data

Data

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

Histogram



Descriptive measures

Mean	3,9
Median	4
Mode	4
Standard Deviation	1,67
Variance	2,78
kurtosis	-0,43
Skewness	-0,02
Range	6
Minimum	1
Maximum	7
Count	30

Mean, Standard Deviation, Coefficients of Asymmetry and Pointing for Pooled Data

Frequency table for the quantitative variable

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**		

Let c_j be the class mark (or class center) and f_j the relative frequency of class j, where j = 1, 2,..., k.

1) The Mean for grouped data is equal to the sum of the goods of the class marks by their relative frequencies, of the form:

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

The standard deviation for pooled data is given by:

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

The Coefficient of Asymmetry for grouped data is given by:

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

The Pointing Coefficient for pooled data is given by:

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

Linear association measures: covariance

Quantitative Data

Covariance: It is a measure of **Joint Variability** between two variables (x_1 , x_2) or (x , y)

x	y
$x_{(1)}$	$y_{(1)}$
$x_{(2)}$	$y_{(2)}$
$\vdots$	$\vdots$
$x_{(n)}$	$y_{(n)}$

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- If Cov(x,y) is positive: the association between x and y is directly proportional, i.e., when *x* increases, *y* also increases; and vice versa.
- If Cov(x,y) is negative: the association between *x* and *y* is inversely proportional, i.e., when *x* increases, *y* decreases, and vice versa.
- If Cov(x,y) is zero: there is no association between *x* and *y*.

Linear association measures: correlation

Quantative Data

Correlation: It refers to the degree of association between two variables (x1 , x2) or (x , y)

Pearson Correlation Coefficient (r): measures the degree of Linear Association between two quantitative variables

x	y
$x_{(1)}$	$y_{(1)}$
$x_{(2)}$	$y_{(2)}$
$\vdots$	$\vdots$
$x_{(n)}$	$y_{(n)}$

$$r = \frac{\text{cov}(x,y)}{s_x s_y} \quad \cdots\cdots\blacktriangleright \quad r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} \qquad -1 \leq r \leq 1$$

- If r is positive: the association between x and y is directly proportional, i.e., when x increases y it also increases; and vice versa.
  - If r=1: the linear association is perfect.
- If r is negative: the association between x and y is inversely proportional, i.e., when x increases and decreases; and vice versa.
  - If r=-1: the linear association is perfect.
- If r is zero: there is no association between x and y.

Example: graphical representation of variables x and y