TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

MSc European Economy and Business Law

Statistics Pre-course 2022

Topic 1: Descriptive Statistics

Francesca Centofanti

francesca.centofanti@students.uniroma2.eu

Outline

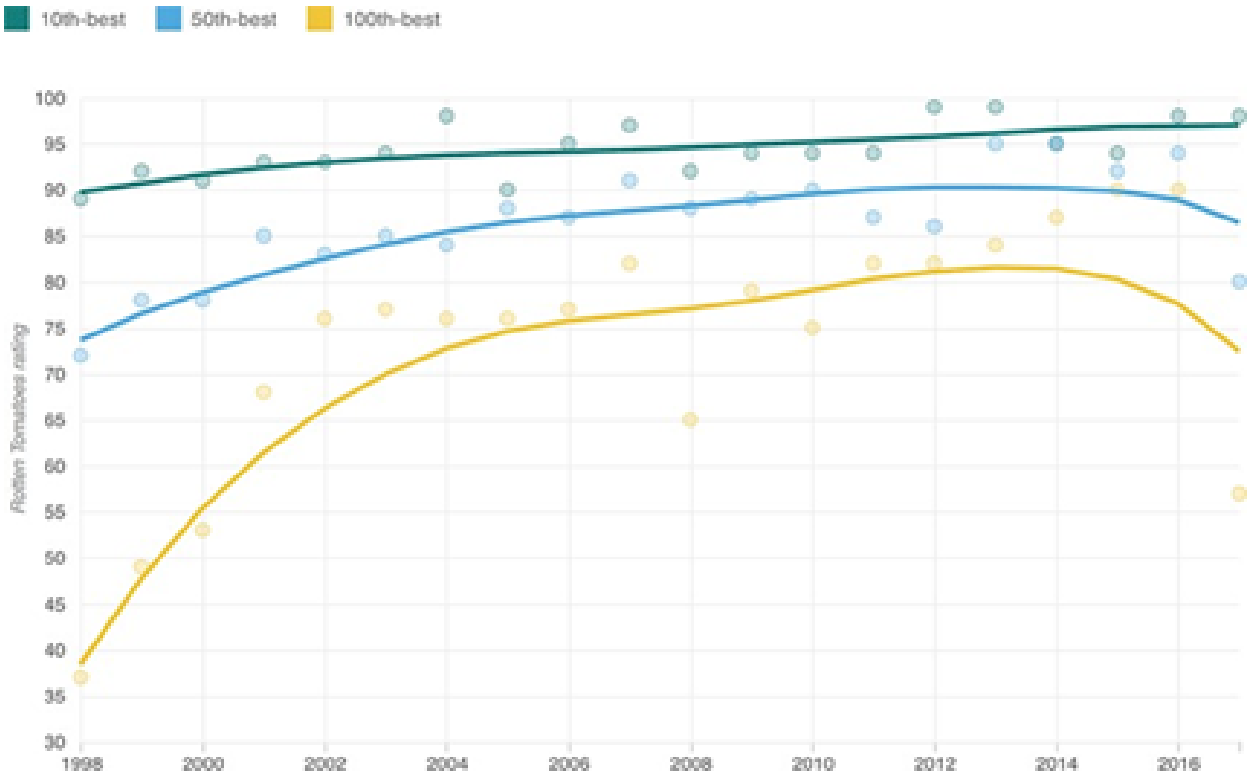
- Introduction
- Statistical variables
- Distributions
- Graphical Representation
- Descriptive statistics
- Appendix: Descriptive Statistics with two variables

Introduction: What is Statistics?

Statistics is the art of learning from data

- How to make sense of the data: **describe** (formulating an hypothesis) and **infer** (validating an hypothesis)
  1. **Descriptive statistics** refers to methods for summarizing the collected data (where the data constitutes either a sample or a population). The summaries usually consist of plots and numbers such as averages and percentages.
  2. **Inferential statistics** refers to methods of making decisions or predictions about a population, based on data obtained from a sample of that population.

Example



- **Descriptive Statistics**: there is an increasing trend in the ratings of "good" movies.
- **Inferential Statistics**: can we conclude that movies are getting better or is this trend appearing just by coincidence?

Statistical variables: What is a variable?

A **variable** y is any **characteristic** observed in a study

- The **values** we observe for a variable are the **data** and we call them **observations**
- We can distinguish between **population** and **sample** data:
  1. **Population**: The collection of all individuals, families, groups, organizations, and units that we are interested in finding out about
  2. **Sample**: The subset of the population we observe
- Each member of a population is identified by the unit i and the set $x_1, ..., x_k$ in which the statistical units take values is called *modality*

Example

- **Variable**: Grade in Statistics of Tor Vergata's Students
- **Observations**: 18, 22, 26, 30, 22, 21, ...
  1. Population: all the students of Tor Vergata
  2. Unit: the single student y_i
  3. Sample: a class $y_1, ..., y_N$
  4. Modality: 18:30

Statistical variables: Types of variables

- **Qualitative** (Categorical):
  1. Nominal
  2. Ordinal
- **Quantitative** (Numerical)
  1. Discrete
  2. Continuous

Statistical variables: Categorical

A variable is called **categorical** if each observation belongs to one of a **set of distinct categories**

- **Nominal**: Categories are disconnected

Examples: Hair Color, Religion, Race, McDonald's menu item, supported Football Team

- **Ordinal**: Categories are ranked

Examples: Amazon's rating, Level of Education

BE CAREFUL! A variable using numbers as labels for its categories is still a categorical variable and is not quantitative.

Statistical variables: Numerical

A variable is called **quantitative** if observations on it take **numerical values** that represent different magnitudes of the variable

- **Discrete**: the variable assume values in a countable set (*how many*)

Examples: Number of Episodes in a series, Grades

- **Continuous**: the variable assume values in a continuous set (*how much*)

Examples: Time, most Physical Measurements

Distributions: categorical/numerical discrete variables frequency

Given a sample $y_1, ..., y_N$ taking values in a set $x_1, ..., x_k$ we define:

- **Absolute (Raw) Frequency** n_i: how many times the $i-th$ modality appears in the sample

$$n_i = \sum_{j=1}^{N} 1(y_j = x_i)$$

- **Relative Frequency** f_i: proportion of the how many $i-th$ modality appears in the sample

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

Frequency distribution: tabular summary of data showing the frequency of items in each of several modalities of the variable of interest.

Example

- **Frequency Table**

Modalities x	Absolute Frequency n	Relative Frequency f
x_1	n_1	f_1
$\dots$	$\dots$	$\dots$
x_K	n_K	f_K

Distributions: Cumulative Relative Frequency

If the variable is *categorical ordinal* or *numerical discrete* we can define **Cumulative Relative Frequency**, as the proportion of observation that take a given value

Let $x_1, ..., x_k$ be the modalities increasingly ordered, then

$$F_i = \sum_{j \leq 1} f_j$$

Example: x_1 = Very Unsatisfied, x_2 = Mildly unsatisfied, x_3 = Neutral, x_4 = Mildly Satisfied, x_5 = Very Satified

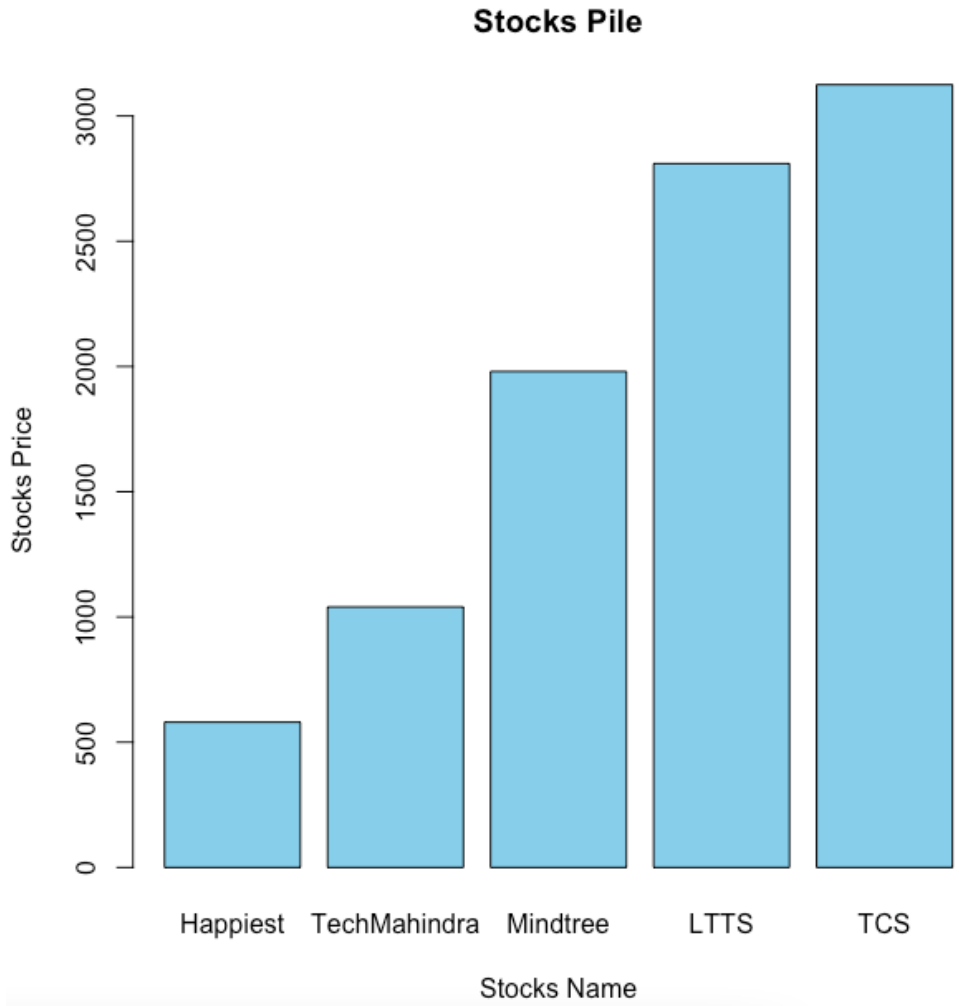
- What is the proportion of customer which is not happy with the service? F_3

Graphical Representation

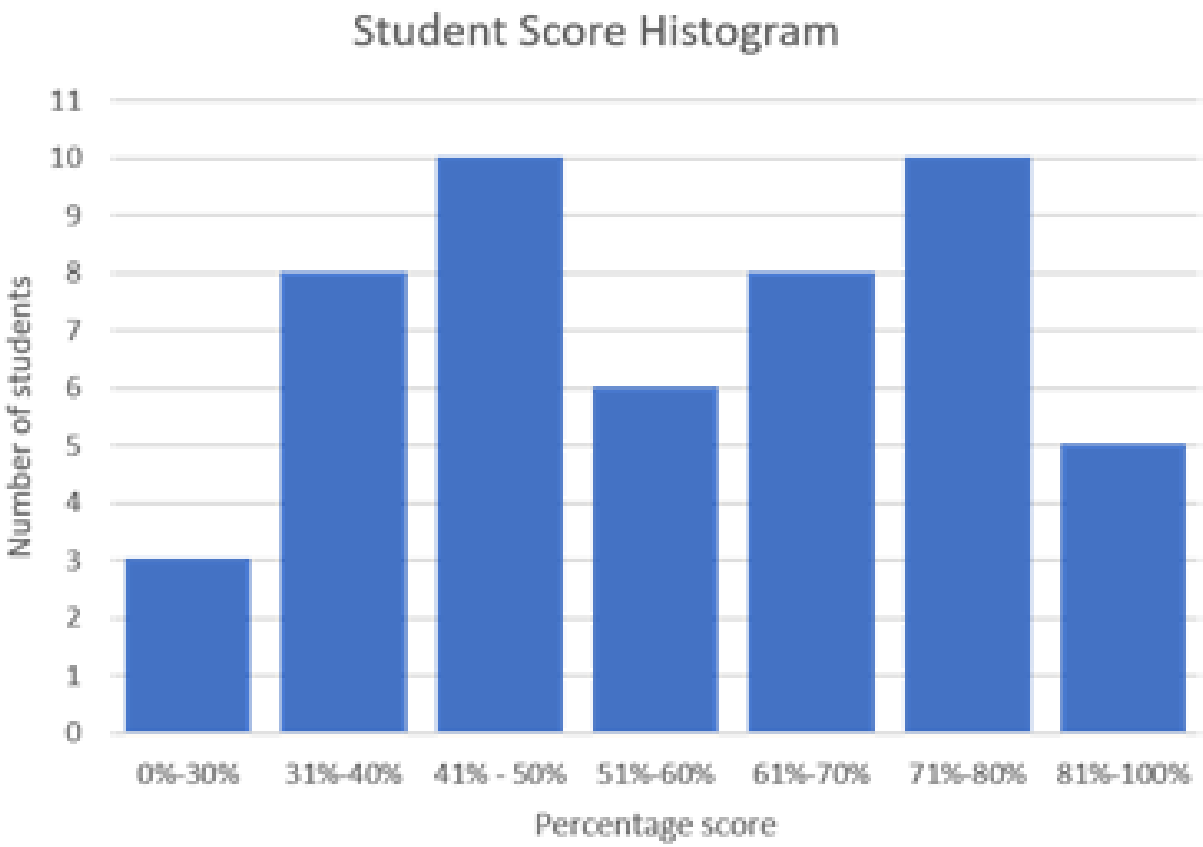
Graphical representations give us "a feel" for the data, telling us about the "shape" and the spread of the data and allowing us to compare different variables

- **Barplot** (*Categorical / Discrete variables*): each modality is associated to a bar, whose height corresponds to the absolute frequency
- **Histogram** (*Continuous variables*): each class is associated to a bar whose area correspond to its density

Example: Barplot



Example: Histogram



Descriptive Statistics: Summary Statistics

Summary Statistics are numerical summaries which describe a feature of the variable with **one number**.

- Most important (useful) features are:
  1. **Centrality**: describing what is a "typical" value for the observations
  2. **Variability**: describing whether the observations take similar values or they are quite different from one another

Descriptive Statistics: Centrality Measures

- From a *frequency table* we can compute:
  1. **Mode**: the value that appears more often (it can be more than one!)
  2. **Median**: the value that splits in half the distribution
  3. **Mean**: the balance point of the distribution

Descriptive Statistics: Mode

The *mode* is the modality with the **highest observed frequency**

$$y_{mode} = \{x_j \ \ such \ \ that \ \ f_j \geq f_i \ \ \forall i \neq j\}$$

- This is the most general notion of center and applies to all types of data, both categorical and numerical data.
- If data are grouped (i.e. they are divided into classes), then the notion of mode becomes the modal class

Descriptive Statistics: Median

The *median* is the **middle value** of the observations when the observations are ordered (in whichever direction)

- Let $y_{(1)}, ..., y_{(N)}$ be the ordered sample:
  1. if N is **odd**, the median is the value

$$y_{med} = y_{(\frac{N+1}{2})}$$

  2. if N is **even**, the median is the value

$$y_{med} = \frac{1}{2}(y_{(\frac{N}{2})} + y_{(\frac{N}{2}+1)})$$

- The median can be computed for all numerical variables and for categorical ordinal variables.
- The median is a center in the sense that it splits the data in two, half the data below it and half above it

Median extension: Quartiles

- The median y_{med} tell us which is the level reached by at least 50% of the population
- Its extension, the **percentile** of level p, tell us which is the level reached by at least $p \times 100\%$ of the population, and it is defined as the first modality x_k for which $F_k \geq p$.
- The percentiles of level $p = 0.25$ and $p = 0.75$ have a special role and are called 1st and 3rd **quartile** respectively.

Descriptive Statistics: Mean

The *(arithmetic) mean* is the sum of the observations divided by the number of observations

$$\overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

- It is interpreted as the **balance point** of the distribution because it minimizes the following quantity:

$$\overline{y} = \underset{c}{arg\,min} \sum_{i=1}^{N} (y_i - c)^2$$

Mean Properties

1. **Internality**:

$$y_{(1)} \leq \overline{y} \leq y_{(N)}$$

2. **Linearity**: if $z_i = ay_i + b$ then

$$\overline{z} = a\overline{y} + b$$

3. **Zero-deviation**:

$$\sum_{i=1}^{N}(y_i - \overline{y}) = 0$$

4. **Associativity**: if $\overline{y}$, $\overline{x}$ are the arithmetic means of two samples of size N and M, respectively, then the mean of the combined sample is

$$\overline{z} = \frac{N \times \overline{y} + M \times \overline{x}}{N + M}$$

Descriptive Statistics: Final Remarks on Centrality

- Mean from the frequency distribution (alternative definition):

$$\overline{y} = \frac{1}{N}\sum_{j=1}^{K} x_j n_j = \sum_{j=1}^{K} x_j \frac{n_j}{N} = \sum_{j=1}^{K} x_j f_j$$

- Mean vs Median: the mean takes into account all observations (including possible anomalous ones), while the median takes into account only order of the observations, regardless of their values

REMEMBER! Mean and Median coincide only when the distribution of the variable is symmetric

Descriptive Statistics: Variability Measures

- Observations are "similar" between each other according to:
  1. **Ranges**: observations varies in a small interval
  2. **Variance**: observations are all close to the same value (typically the mean)

Descriptive Statistics: Ranges

- Based on the idea of *variability* as the size of the interval in which the observations lay, we can define two measures of spread:
  1. **(Global) Range of Variation**: difference between the maximum and the minimum value observed in the sample

$$RV = y_{(N)} - y_{(1)}$$

  2. **Interquartile Range**: difference between the 3rd and the 1st quartile and it gives us the smallest interval in which 50% of the observations lay

$$IQ = y_{Q_1} - y_{Q_3}$$

Descriptive Statistics: Variance

The *variance* is base on the idea that the larger the deviations from the mean $(y_i - \overline{y})$, regardless of their sign, the larger the variability

- Because of the *zero-deviation property* of the mean, we cannot use just the average of the deviations to exploit the idea of deviation from the center
- To solve this issue, we have to use the **average of the squared deviations**:

$$s^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y})^2}{N}$$

- To compute the variance, we can use two strategies:
  1. Compute directly all the deviations $(y_i - \overline{y})^2$, then average them
  2. Exploit the alternative definition of the variance:

$$s^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \overline{y}^2 \quad with \quad \frac{1}{N}\sum_{i=1}^{N} y_i^2 = \frac{1}{N}\sum_{i=1}^{K} x_j^2 n_j = \sum_{i=1}^{K} x_j^2 f_j$$

Descriptive Statistics: Final Remarks

- Variance is a *scale-dependent* index, i.e. it depends on the scale on which the data are measured (it is sometimes called **scale parameter**)
- **Standard Deviation** (square root of the variance): same scale as the data, making it more interpretable than the variance

$$s = \sqrt{s^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2}$$

- To compare the variability of variables measured on different scales we have to compute the **variation coefficient**

$$CV = \frac{s}{|\overline{y}|}$$

REMEMBER! Differences are relative: to compare variables with different magnitudes we have to use a *standardization* process Z

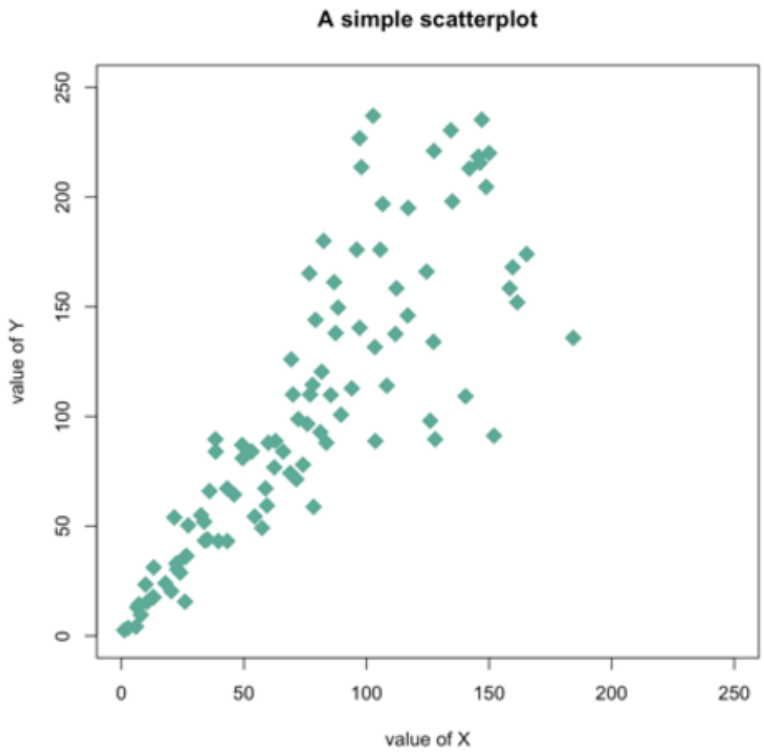
$$Z = \frac{X - \mu}{\sigma}$$

Appendix: Descriptive Statistics with two variables

- In real life we usually have more than one variable measured on each unit, that may or may not be related
- We focus on the case of both variables being *numerical* (although association measures exist for categorical and mixed type variables as well)
- We want to know whether there is a relationship (**association**) between them:
  1. **Positive Association**: as x goes up, y tends to go up
  2. **Negative Association**: as x goes up, y tends to go down

Association: Graphical Representation

- A two dimensional variable (X_1, X_2) is usually represented through a **scatterplot**
  1. each axis correspond to one of the two variables X, Y
  2. each point represent a units and its coordinates correspond to values of the two variables observed on it



Association Measures: Covariance and Correlation

The *covariance* is an indicator measuring the **strength** of the association between two variables

$$cov_{x,y} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

BE CAREFUL! The covariance measure only *linear association* (i.e. the case where the relationship between the two variables x and y is of the type $y_i = ax_i + b$) and depends on the scale of the data (we have no general reference to determine if the observed covariance is large or small)

- Rescaled Covariance: **Correlation**

$$r_{x,y} = \frac{cov_{x,y}}{s_x s_y}$$

Association Measures: Pros and Cons of Correlation

The *correlation* is between -1 and 1: the closer $|r_{x,y}|$ is to 1, the stronger the linear association in the observations

- **Pros**:
  1. Correlation does not depend on the variables'unit, i.e. it is not affected by the scale of the observations
  2. Correlation is symmetric with respect to the two variables, i.e. it does not treat favourably one variable over the other
- **Cons**:
  1. Correlation still measure only linear dependence. That is $|r_{x,y}| = 1$ when data lay on a straight line. However, if there is some more complicated form of dependence, even something simple like $y_i = ax_i^2 + b$, correlation may not capture it
  2. Correlation *does not* mean **causation**! There may be a strong correlation between two variables even when there is no relation between them

Example: Spurious Correlation

- X = People who died by falling out of their bed
- Y = Lawyers in Puerto Rico
- $r_{x,y} = 0.957087$