

# Types of Data

Rosario Barone

Tor Vergata University of Rome

Statistical tools for decision making

Undergraduate Degree in Global Governance

A.Y. 2024/2025

# Basic Concepts in Statistics

- **Population:** The entire group of interest from which data is collected.
- **Sample:** A subset of the population used to represent the whole.
- **Variable:** A characteristic or property being measured or observed.
- **Data:** The values collected from the population or sample.
- **Descriptive Statistics:** Methods to summarize and describe data.
- **Inferential Statistics:** Methods to make predictions or inferences about a population based on sample data.

# Data

- Different types of data are classified based on their characteristics.
- Understanding data types is crucial for choosing appropriate analysis methods.

# Qualitative data

- It can't be expressed as a number and can't be measured. It consist of words, pictures, and symbols.
- Also called categorical data because the information can be sorted by category.

# Examples of qualitative data

- Colors e.g. the color of the eyes
- Holiday destination
- Names
- Ethnicity

There are 2 general types of qualitative data: **nominal** data and **ordinal** data

# Nominal data

- Nominal data is used just for labeling variables, without any type of quantitative value.
- The nominal data just name a thing without applying it to order. It could just be called "labels."

# Examples of nominal data

- Gender
- Hair color
- Marital status
- Ethnicity

Note that there is no intrinsic ordering to the variables:

Eye color is a nominal variable having a few categories (Blue, Green, Brown) and there is no way to order these categories from highest to lowest.

# Ordinal data

- Ordinal data shows where a number is in order. This is the crucial difference from nominal types of data.
- Ordinal data is data which is placed into some kind of order by their position on a scale. Ordinal data may indicate superiority.
- You cannot do arithmetic with ordinal numbers because they only show sequence.
- Ordinal variables are considered as “in between” qualitative and quantitative variables.

# Ordinal data

- In other words, the ordinal data is qualitative data for which the values are ordered.
- In comparison with nominal data, the second one is qualitative data for which the values cannot be placed in an ordered.
- We can also assign numbers to ordinal data to show their relative position. But we cannot do math with those numbers. For example: “first, second, third. . . etc.”

# Examples of ordinal data

- The first, second and third person in a competition.
- Rating classes: A, B, C, and etc.
- Status: low, medium and high.

# Quantitative data

- Quantitative data seems to be the easiest to explain. It answers key questions such as “how many, “how much” and “how often” .
- Quantitative data can be expressed as a number or can be quantified. Simply put, it can be measured by numerical variables.
- Quantitative data are easily amenable to statistical manipulation and can be represented by a wide variety of statistical types of graphs and charts such as line, bar graph, scatter plot, and etc.

# Examples of quantitative data

- Scores on tests and exams e.g. 85, 67, 90 and etc.
- The weight of a person or a subject.
- Your shoe size.
- The temperature in a room.

There are 2 general types of quantitative data: **discrete** data and **continuous** data.

# Discrete data

- Discrete data is a count that involves only integers. The discrete values cannot be subdivided into parts.
- For example, the number of children in a class is discrete data. You can count whole individuals. You can't count 1.5 kids.
- To put in other words, discrete data can take only certain values. The data variables cannot be divided into smaller parts.
- (Most of the times) It has a limited number of possible values e.g. days of the month.

# Examples of discrete data

- The number of students in a class.
- The number of workers in a company.
- The number of goal, passes or shoots in a football game.
- The number of test questions you answered correctly.

# Ordering and plotting data

## Frequency

The number of times an event or value occurs in a dataset.

## Frequency Table

A table that displays the frequencies of different values or categories in a dataset.

# Frequency

- **Absolute frequency:** number of times that a value appears. It is represented as  $f_i$  where the subscript represents each of the values. The sum of the absolute frequencies is equal to the total number of data, represented as  $N$ .

$$f_1 + f_2 + \dots, f_n = N$$

or equivalently

$$\sum_{i=1}^n f_i = N$$

- **Relative frequency:** the result of dividing the absolute frequency of a certain value by the total number of data. It is represented as  $n_i$ . The sum of the relative frequencies is equal to 1. We can prove this easily by factorizing

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

# Frequency

- **Cumulative frequency:** the sum of absolute frequencies of all the values equal to or less than the considered value (only for ordinal data). This is represented as  $F_i$ .
- **Relative cumulative frequency:** the result of dividing the cumulative frequency by the total number of information, which is represented by  $N_i$ .

# Creating a Frequency Table

- 1 Identify the values or categories in your dataset.
- 2 Count how many times each value or category appears.
- 3 Organize the data into a table.

## Example: Rating class modelling

Our sample comprises 25 European corporations operating in the Energy sector.

<b>Corporation</b>	<b>Country</b>	<b>Rating class</b>
ItaEnergy	Italy	D
GoEng	Germany	A
Ena	France	B
SolarEn	Italy	A
⋮	⋮	⋮
PowerFS	Spain	C

Table: Simulated Rating Class Data

## Example: Rating class analysis

We can summarize the information with a frequency table!

Rating class	$f_i$
A	5
B	8
C	3
D	2
E	7

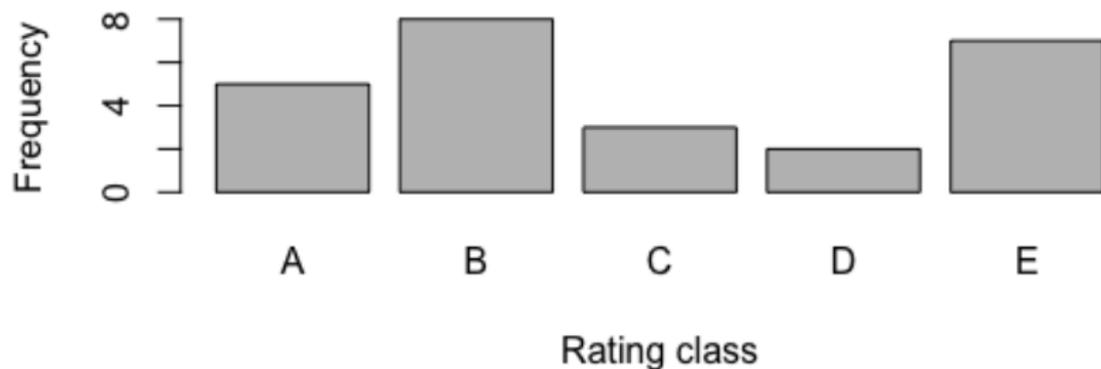
Table: Frequency Table of Ratings

## Example: Rating class analysis

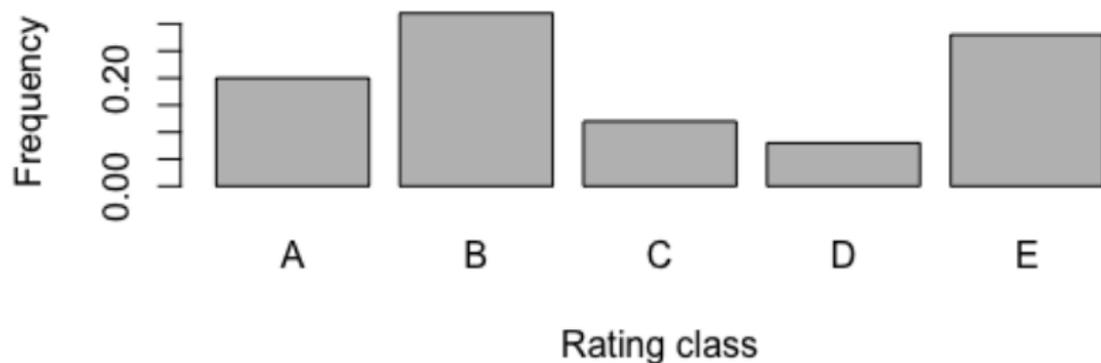
Rating class	$f_i$	$n_i$
A	5	$5/25 = 0.20$
B	8	$8/25 = 0.32$
C	3	$3/25 = 0.12$
D	2	$2/25 = 0.08$
E	7	$7/25 = 0.28$

Table: Frequency Table of Ratings

# Rating class modelling: Barplot Absolute Frequency



# Rating class modelling: Barplot Relative Frequency



## Example: Rating class analysis Absolute Frequencies

Rating class	$f_i$	$n_i$
A	5	$5/25 = 0.20$
B	8	$8/25 = 0.32$
C	3	$3/25 = 0.12$
D	2	$2/25 = 0.08$
E	7	$7/25 = 0.28$

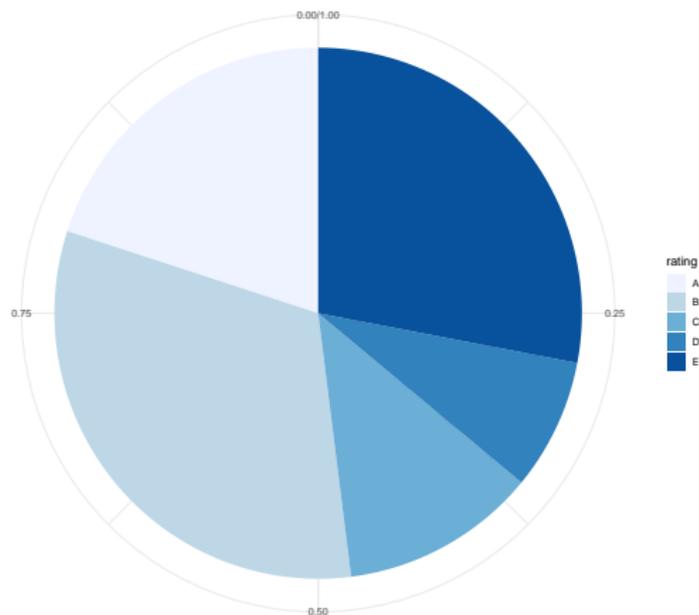
Table: Frequency Table of Ratings

## Example: Rating class analysis Absolute Frequencies

Rating class	$f_i$	$n_i$	$F_i$	$N_i$
A	5	$5/25 = 0.20$	5	0.2
B	8	$8/25 = 0.32$	13	0.52
C	3	$3/25 = 0.12$	16	0.64
D	2	$2/25 = 0.08$	18	0.72
E	7	$7/25 = 0.28$	25	1

Table: Frequency Table of Ratings

# Rating class modelling: Pie chart



## Example: Crimes in Italian municipalities in 2022

	<b>Municipality</b>	<b>Crime</b>	<b>date</b>
1	Rome	robbery	01/01/2022
2	Milan	murder	01/01/2022
3	Florence	brawl	01/01/2022
4	Naples	robbery	01/01/2022
⋮	⋮	⋮	⋮
2183043	Pavia	domestic violence	31/12/2022
2183044	Rome	brawl	31/12/2022
2183045	Rome	theft	31/12/2022

Table: Simulated Crime data

\*NOTE: The data shown is not real\*

## Example: Crimes in Italian municipalities in 2022

- Does it make sense to analyze the frequencies city by city?

## Example: Crimes in Italian municipalities in 2022

- Does it make sense to analyze the frequencies city by city?
- In Italy there are 7901 municipalities: no, it doesn't make much sense.

## Example: Crimes in Italian municipalities in 2022

- Does it make sense to analyze the frequencies city by city?
- In Italy there are 7901 municipalities: no, it doesn't make much sense.
- What can we do?

## Example: Crimes in Italian municipalities in 2022

	<b>Municipality</b>	<b>Area</b>	<b>Number of Crimes</b>
1	Abano Terme	North	9
2	Abbadia Cerreto	North	7
3	Abbadia Lariana	North	13
3	Abbadia San Salvatore	Center	3
⋮	⋮	⋮	⋮
7899	Zumpano	South	4
7900	Zungoli	South	12
7901	Zungri	South	7

Table: Crime data

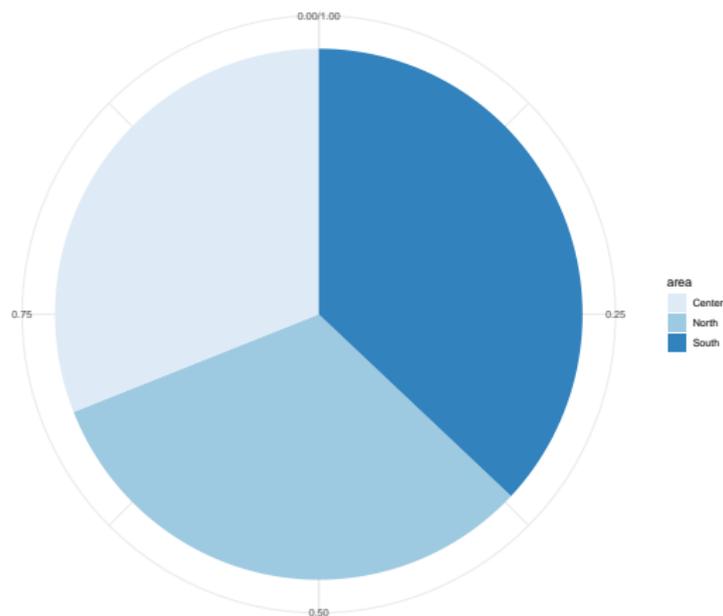
\*NOTE: The data shown is not real\*

## Example: Crimes in Italy in 2022

<b>Rating class</b>	<b><math>f_i</math></b>	<b><math>n_i</math></b>
North	698574	0.32
Center	676744	0.31
South	807727	0.37

Table: Frequency Table of Crimes in Italy

# Example: Crimes in Italy in 2022



# Continuous data

- Continuous data is information that could be meaningfully divided into finer levels. It can be measured on a scale or continuum and can have almost any numeric value.
- For example, you can measure your height at very precise scales — meters, centimeters, millimeters and etc.
- You can record continuous data at so many different measurements — width, temperature, time, and etc. This is where the key difference from discrete types of data lies.
- The continuous variables can take any value between two numbers. For example, between 50 and 72 meters, there are infinite of possible lengths.

# Examples of continuous data

- Time (Waiting time).
- Height and Weight.
- Stock Price.
- Income, Inflation and other economic variables.

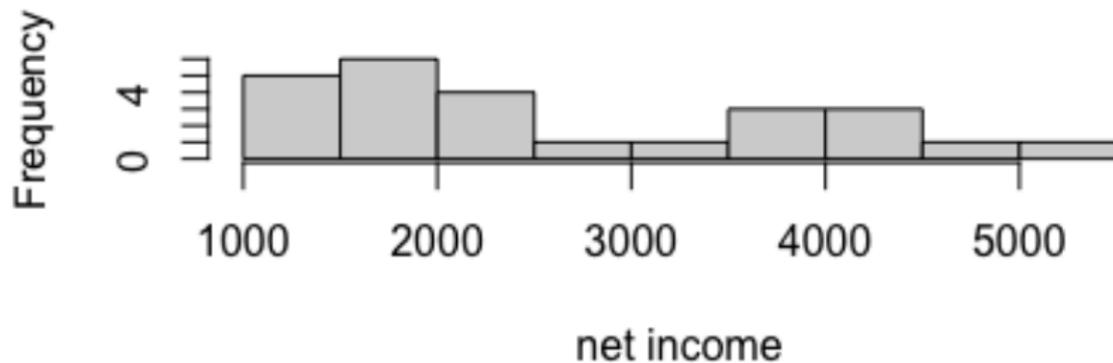
## Example: Net Income analysis

<b>Corporation</b>	<b>Country</b>	<b>Rating class</b>	<b>Net Income (M)</b>
ItaEnergy	Italy	D	1400
GoEng	Germany	A	4300
Ena	France	B	3300
SolarEn	Italy	A	4570
⋮	⋮	⋮	⋮
PowerFS	Spain	C	2700

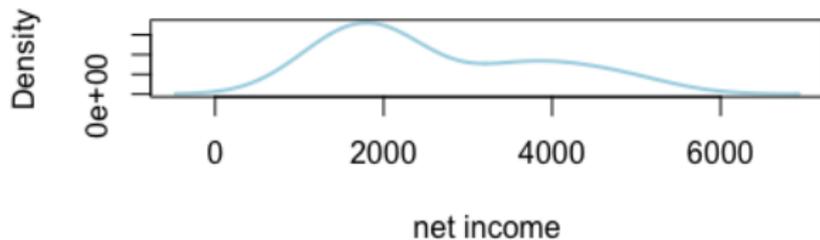
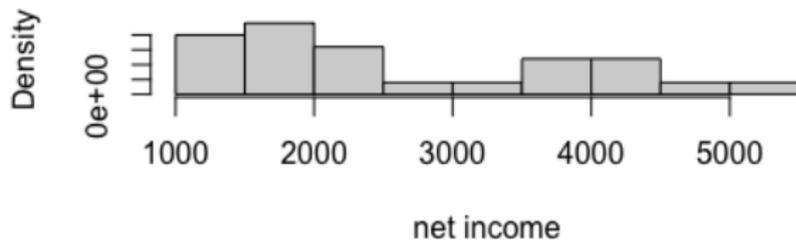
Table: Simulated Rating Class Data

We want to analyse the net income.

## Example: Net Income analysis



# Example: Net Income analysis



## Example: Net Income analysis

What if we want to summarise the **Net Income** distribution?

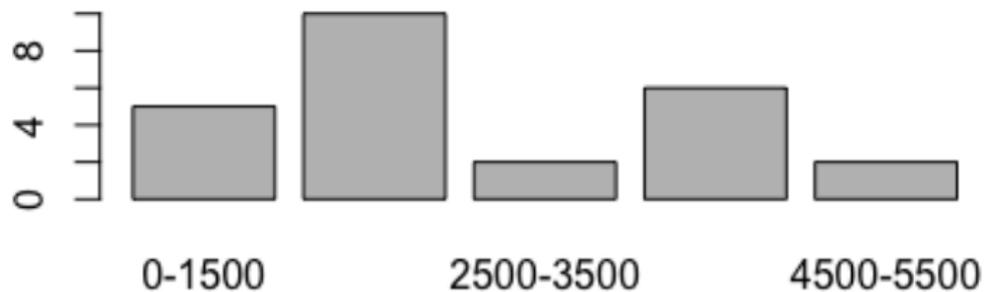
## Example: Net Income analysis

What if we want to summarise the **Net Income** distribution?

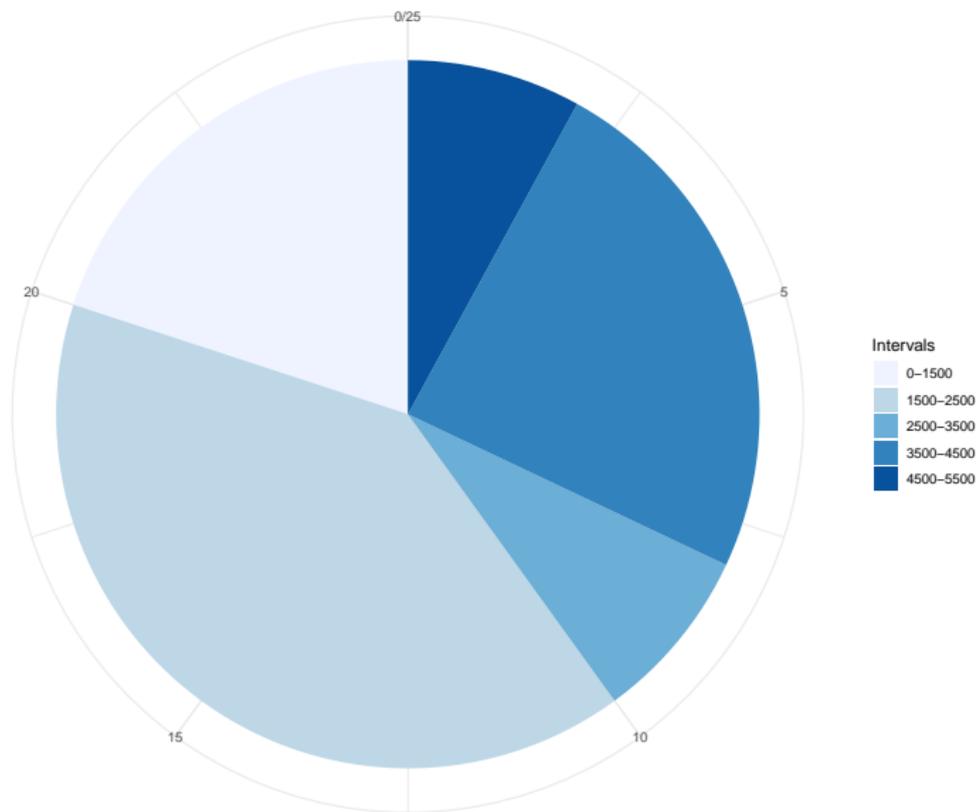
We can split in intervals the **Net Income** values perform the analysis.

<b>Net Income range (M)</b>	<b>Number of Corporations</b>
[0, 1500]	5
]1500, 2500]	10
]2500, 3500]	2
]3500, 4500]	6
[4500, 5500]	2

# Net Income analysis: Pie Chart



# Net Income analysis: barplot



# Summary on the type of data

- Type of data:
  - ▶ Qualitative:
    - ★ Nominal
    - ★ Ordinal
  - ▶ Quantitative
    - ★ Discrete
    - ★ Continuous
- Frequency distributions: visual displays that organise and present frequency counts so that the information can be interpreted more easily.