Quantitative Methods – I (Statistics)

A. Y. 2022-23

Prof. Lorenzo Cavallo

Chapter1

Introduction

This course: Quantitative Methods

- 12 weeks (12 credits): 6 with me, 6 with prof. Stefano Grassi
- Weekly appointments: 3 classes (2 hours each)

THIS MODULE

- Classes

Tuesday-Wednesday-Thursday (P2), 17:00-19:00

- Office Hours:

Ask for an appointment by email.

- You can reach me here:

lorenzo.cavallo.480084@uniroma2.eu

Quantitative Methods: final exam

- **Exam structure**

  - Closed-book written exam, including open questions, multiple choices questions, exercises ***on the entire program*** ***(Quantitative Methods 1 & 2)***

- **Exam rules**

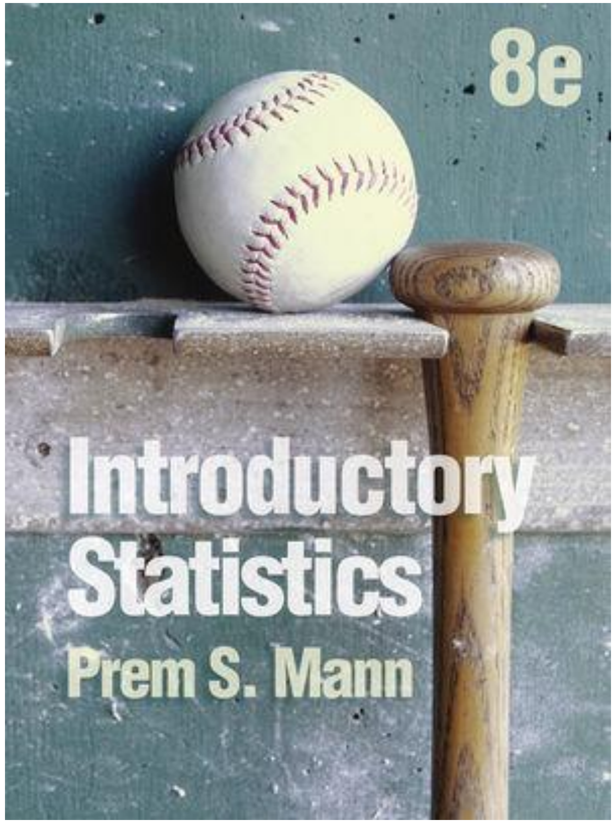
  - 6 dates, you can attend only 1 in each session
  - No mid-term
  - Remember to book for the exam: students not booked will not be allowed to take the exam
  - Final marks on Delphi (typically within one week)

Quantitative Methods: classroom rules

- Students chatting and talking will be invited to leave the room

Quantitative Methods: program

- 4 sub-parts:
  1. Part I: Descriptive Analysis
  2. Part II: Probability and main distributions
  3. Part III: Inference
  4. Part IV: Regression Analysis



- Textbook: Introductory Statistics (8th Edition), by P. S. Mann, edited by Wiley

- Slides and other material on the website. Access granted for those signed to the newsletter

Quantitative Methods: program

The course provides the basics of data analytics for economics and business, dealing with understanding, summarizing and representing statistical information, both univariate and multivariate, measuring association, and the essentials of probability theory and calculus.

The latter is at the basis of statistical inference and learning, which will be the topic of Quantitative Methods 2.

Quantitative Methods: Syllabus

1. Descriptive Statistics and data analysis
   - Data structures and sources, variables and their measurement scales
   - Tables and plots: frequency distributions and graphical representation of data
   - Measures of central tendency and dispersion. Measures of association of two variables
2. Probability theory
   - Basic concepts and set theory. Definition of probability, axioms and theorems. Conditional probability and independence. Bayes' theorem
   - Random variables and probability distributions. Discrete random variables, continue random variables, multiple random variables.

Statistics: meanings

1. **Statistics** as a numerical fact

2. **Statistics** as the science of collecting, analyzing, presenting, and interpreting data, as well as of making decisions based on such analyses (educated guess).

3. **Statistics** has been defined as the art (or science) of learning from data.

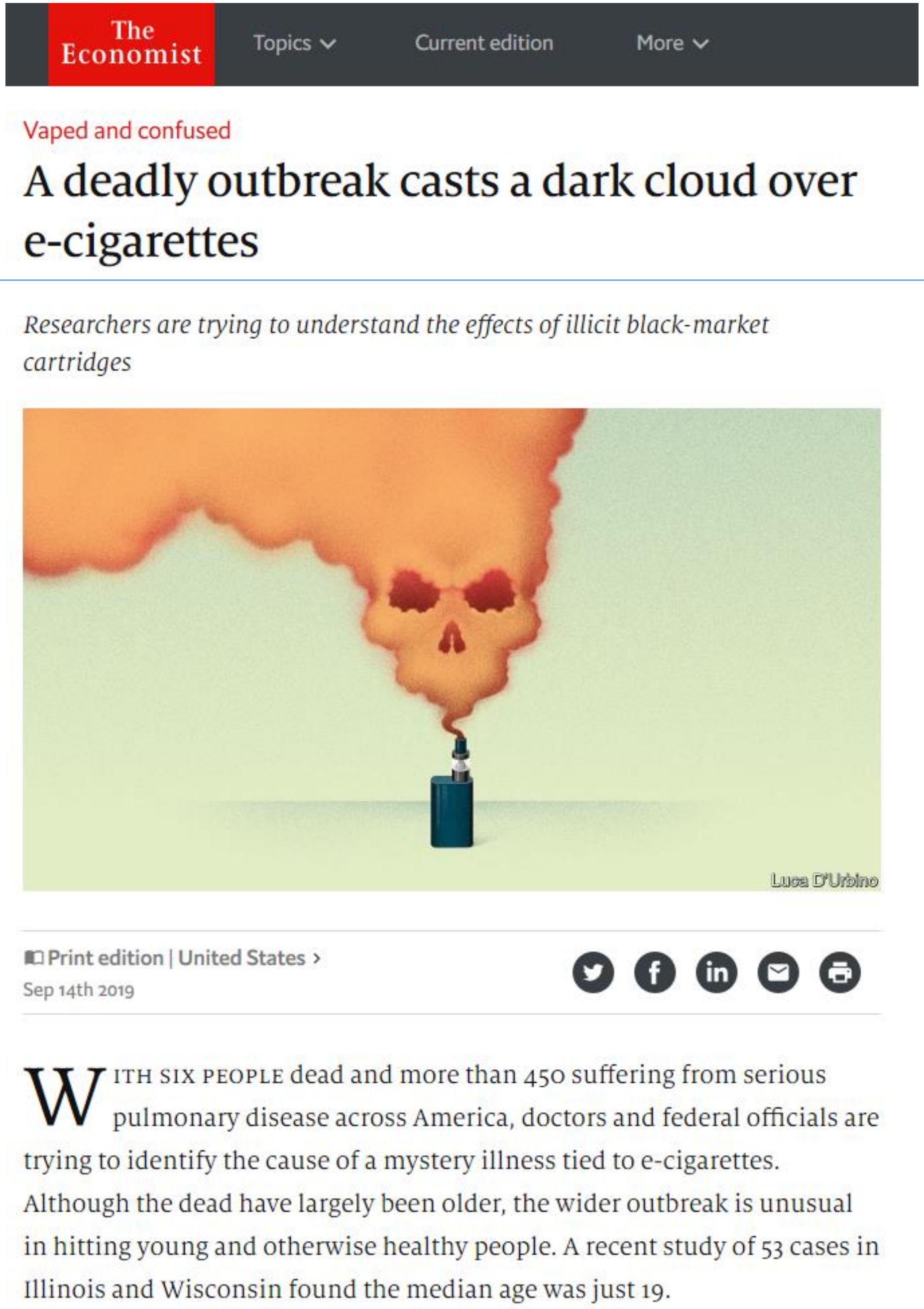
Statistics is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions (S. Ross, Introductory Statistics, 2nd ed., Elsevier, 2005).

Statistics: meanings

1. Statistics as a numerical fact

…***six*** people dead and more than ***450*** suffering from serious pulmonary disease across America…

…a recent study of 53 cases in Illinois and Wisconsin found the median age was just ***19***.



Vaped and confused

A deadly outbreak casts a dark cloud over e-cigarettes

Researchers are trying to understand the effects of illicit black-market cartridges

Luca D'Urbino

Print edition | United States ›

Sep 14th 2019

WITH SIX PEOPLE dead and more than 450 suffering from serious pulmonary disease across America, doctors and federal officials are trying to identify the cause of a mystery illness tied to e-cigarettes. Although the dead have largely been older, the wider outbreak is unusual in hitting young and otherwise healthy people. A recent study of 53 cases in Illinois and Wisconsin found the median age was just 19.

Statistics everywhere

Statistics has become more popular than ever - newspapers and magazines are filled with statistics and graphs summarizing data - and is now used in almost all professions

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy

https://www.indeed.com/jobtrends/q-%22Data-Scientist%22.html

The Economist

Theresa May v Brussels

Ten years on: banking after the crisis

South Korea's unfinished revolution

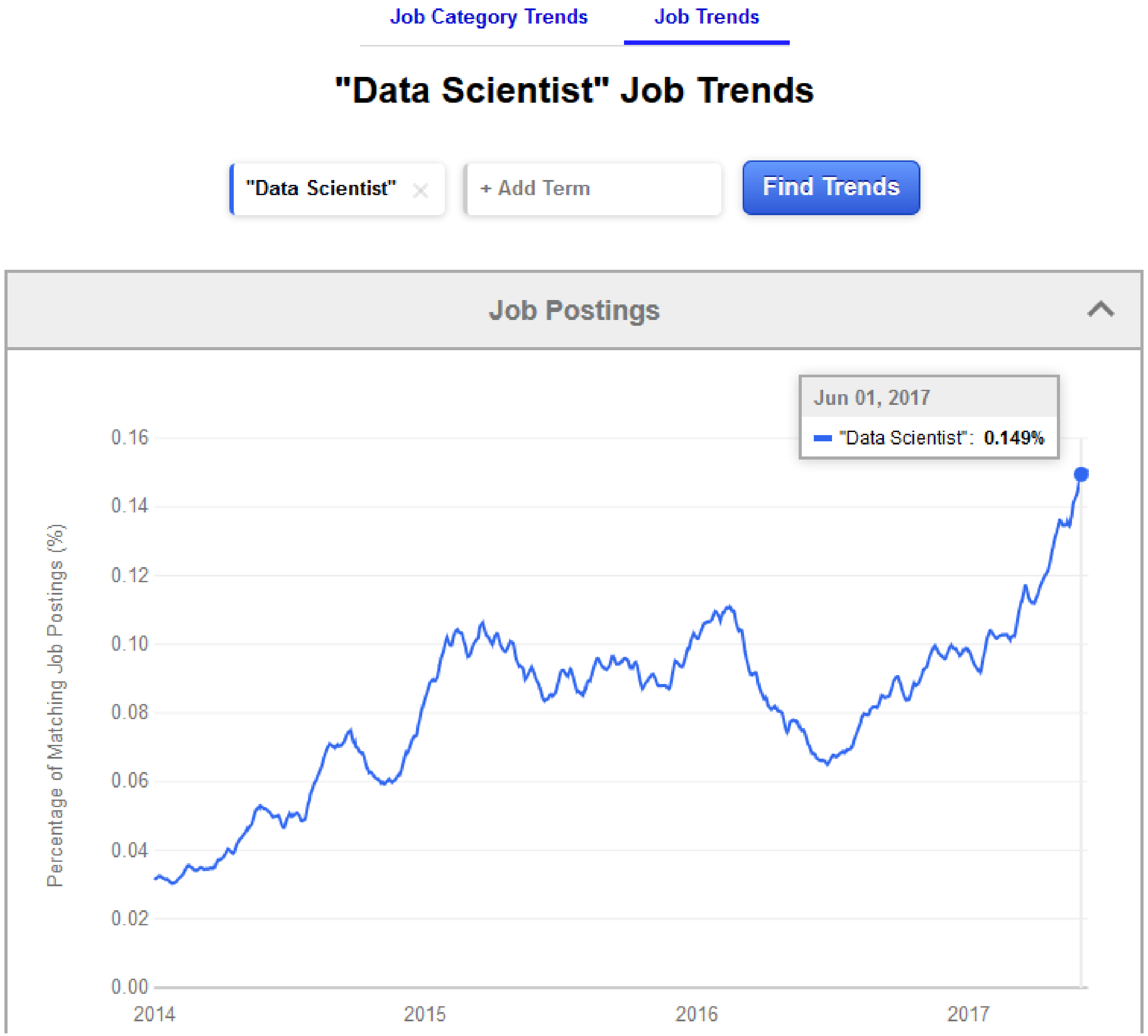
Biology, but without the cells

MAY 6TH–12TH 2017

The world's most valuable resource

Data and the new rules of competition

amazon
Google



"Data Scientist" Job Trends

"Data Scientist"

Job Postings

Jun 01, 2017
"Data Scientist": 0.149%

Statistics as the science: types

1. **Descriptive Statistics** consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures

2. **Probability** concerns the measurement of the likelihood that an uncertain outcome will occur

3. **Inferential Statistics** consists of methods that use ***sample*** results to help make decisions or predictions about a ***population***.

Descriptive Statistics: Data and data structures

Descriptive statistic, or data analysis, deals with presenting, organizing and summarizing data.

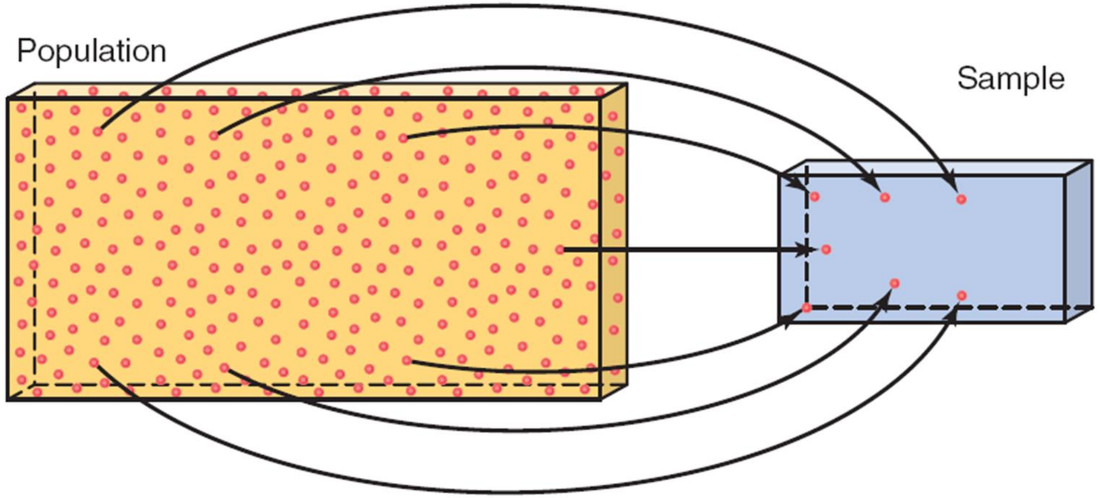
We will assume that data are available or have been collected for a research question.

Note that statistics is also concerned with data collection: the procedures by which data are collected is crucial to be able to address the main research question. The design of surveys and censuses is integral part of the statistical methodology.

Population and Samples

- **Population:** all elements – individuals, items, or objects – whose characteristics are being studied (also called **target population**).

- **Sample:** portion of the population selected for study



Example:

Population ➔ this class

Sample ➔ a random subsample of 10 of you

Populations, samples, units and their characteristics

We are interested in obtaining information about a total collection of elements, which we refer to as a **population** (e.g., All the students enrolled in the second year of BAE, Waiting times for bus n. 20, House prices in Rome municipality during 2021).

A subgroup of the population that is selected for data collection is a **sample**. A representative sample is tipically obtained through randomization, so that each unit has a non zero known probability of being selected.

The elementary units making up the population or the sample are the statistical **units**.

Populations, samples, units and their characteristics

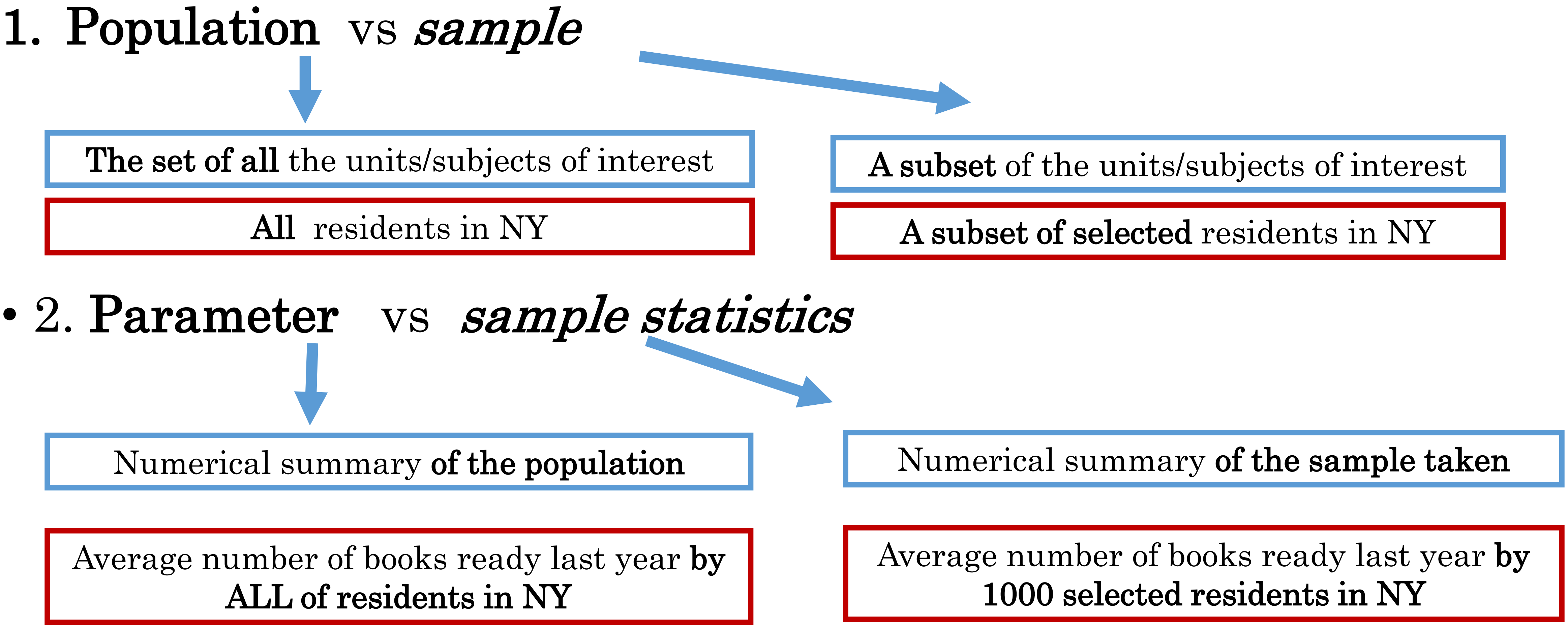
Let C denote a **population** or a sample and let u denote the **units**.

The profile of the units u_1, u_2, . . . , is provided by their characteristics, which is a set of measurements that we call **variables**.

In sum, we consider a set of units, making up a population or a sample, described by a set of variables.

The most commonly available data structure is a matrix, a rectangular array of measurements according to rows and columns, where the rows pertain to the statistical units and the columns to the variables.

Summing up



1. **Population** vs ***sample***

The set of all the units/subjects of interest

All residents in NY

A subset of the units/subjects of interest

A subset of selected residents in NY

- 2. **Parameter** vs ***sample statistics***

Numerical summary **of the population**

Average number of books ready last year **by ALL of residents in NY**

Numerical summary **of the sample taken**

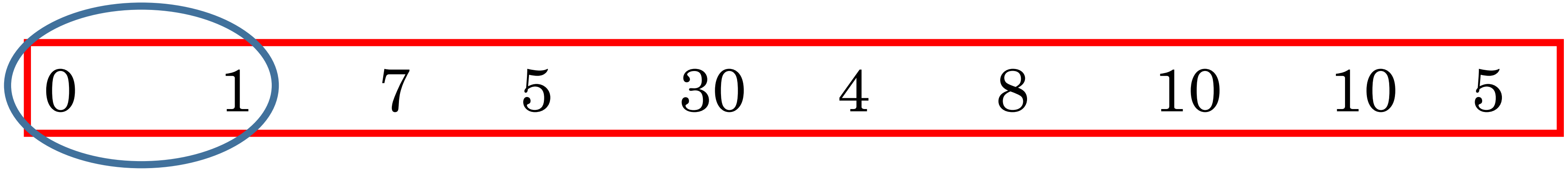
Average number of books ready last year **by 1000 selected residents in NY**

Parameters and sample statistics

Many samples might be taken from the same population ➔ many sample statistics might result.

Ex: parameter is number of books read last year

Suppose this is the population.. ➔ parameter is: 8



0 1 7 5 30 4 8 10 10 5

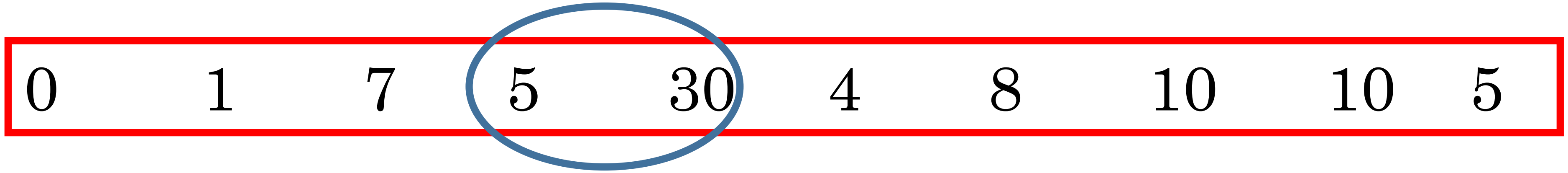
..and this possible sample ➔ sample statistics is: 0.5

Parameters and sample statistics

Many samples might be taken from the same population → many sample statistics might result.

Ex: parameter is number of books read last year

Suppose this is the population.. → parameter is: 8



0 1 7 5 30 4 8 10 10 5

..and this possible sample → sample statistics is: 17.5

Parameters and sample statistics

Many samples might be taken from the same population → many sample statistics might result.

Ex: parameter is number of books read last year

Suppose this is the population.. → parameter is: 8

0 1 7 5 30 4 8 10 10 5

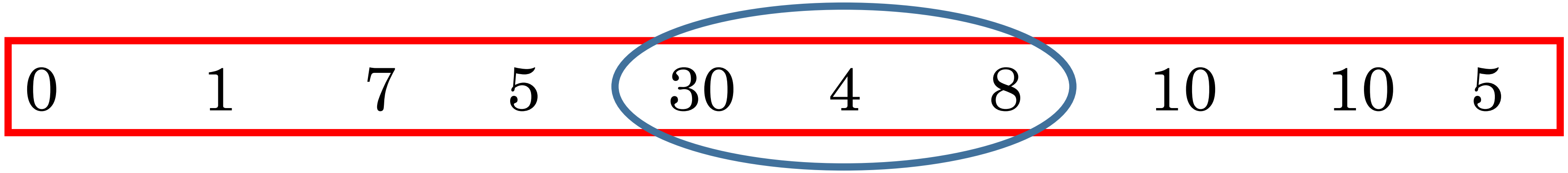
..and this possible sample → sample statistics is: 7.5

Parameters and sample statistics

Many samples might be taken from the same population ➔ many sample statistics might result.

Ex: parameter is number of books read last year

Suppose this is the population.. ➔ parameter is: 8



0 1 7 5 30 4 8 10 10 5

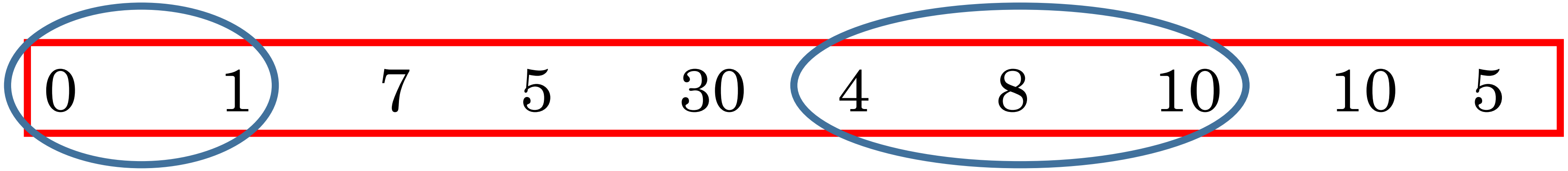
..and this possible sample ➔ sample statistics is: 14

Parameters and sample statistics

Many samples might be taken from the same population ➔ many sample statistics might result.

Ex: parameter is number of books read last year

Suppose this is the population.. ➔ parameter is: 8



0 1 7 5 30 4 8 10 10 5

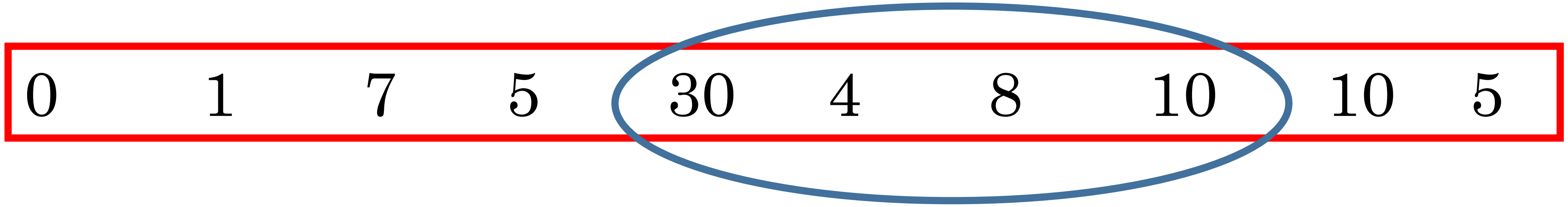
..and this possible sample ➔ sample statistics is: 4.6

Parameters and sample statistics

Many samples might be taken from the same population ➔ many sample statistics might result.

Ex: parameter is number of books read last year

Suppose this is the population.. ➔ parameter is: 8



0 1 7 5 30 4 8 10 10 5

..and this possible sample ➔ sample statistics is: 13

➔ the sample statistics **CHANGES** with the sample!

Survey

Collection of information (***variables***) on the elements of population or sample (***observations***).

Census: data collected on every member of the population (not necessarily people, it could be agricultural, housing, banking…)

See e.g.: https://www.istat.it/en/censuses

Sample Survey: data collected on a subsample of members member of
See e.g.: http://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/

Samples

Representative: it reflects the main characteristics of the population

Simple Random Sample: each element of the population has same chance of being selected

With replacement: each time the element drawn from the population is put back in the population ➔ if we do several draws, we may select the same item more than once

Without replacement: the selected element is not put back in the population (each time, the size of the population is reduced by one) ➔ if we do several draws, we never select the same item more than once

Samples: examples

Consider a box that contains 25 pencils of different colors: you draw a pencil, record its color, and put it back in the box before drawing the next pencil.

Every time we draw a pencil from this box, the box contains 25 pencils ➔ This is sampling **with replacement**.

Q: consider the rolling a die many times. Is this with or without replacement?

Samples: examples

Consider a box that contains 25 pencils of different colors: you extract a pencil, record its color, and DO NOT put it back in the box before extracting the next pencil.

Before the second extraction, the box contains 24 pencils, before the third, 23, etc.. Every time we extract a pencil from this box, the box contains one pencil less ➔ This is sampling **without replacement**.

Q: you drink a bottle of soda from a 12-bottle box and you record the taste. Is this extraction with or without replacement?

Statistical dataset

A statistical dataset or database is a set of measurements taken on the statistical units making up our population or sample.

The information is obtained by measuring variables on the units.

Let X denote a variable (gender, age, income, etc.). A measurement involves attributing a value of X according to some rule and with a given content. In other words, we make the association

$$u \rightarrow x$$

where x is the value of X associated with unit u.

The values of X are exhaustive and mutually exclusive.

Data and measurement scales

To analyze a dataset, you first need to determine what type of data you’re dealing with.

Fortunately, to make this easier, all types of data fit into one of four broad categories: **nominal, ordinal, interval, and ratio data.**

While these are commonly referred to as ‘data types,’ they are really different **scales or levels of measurement**.

The nominal and ordinal data are known as **categorical data**.

The interval and ratio data are both types of **numerical data**.

Categorical Data

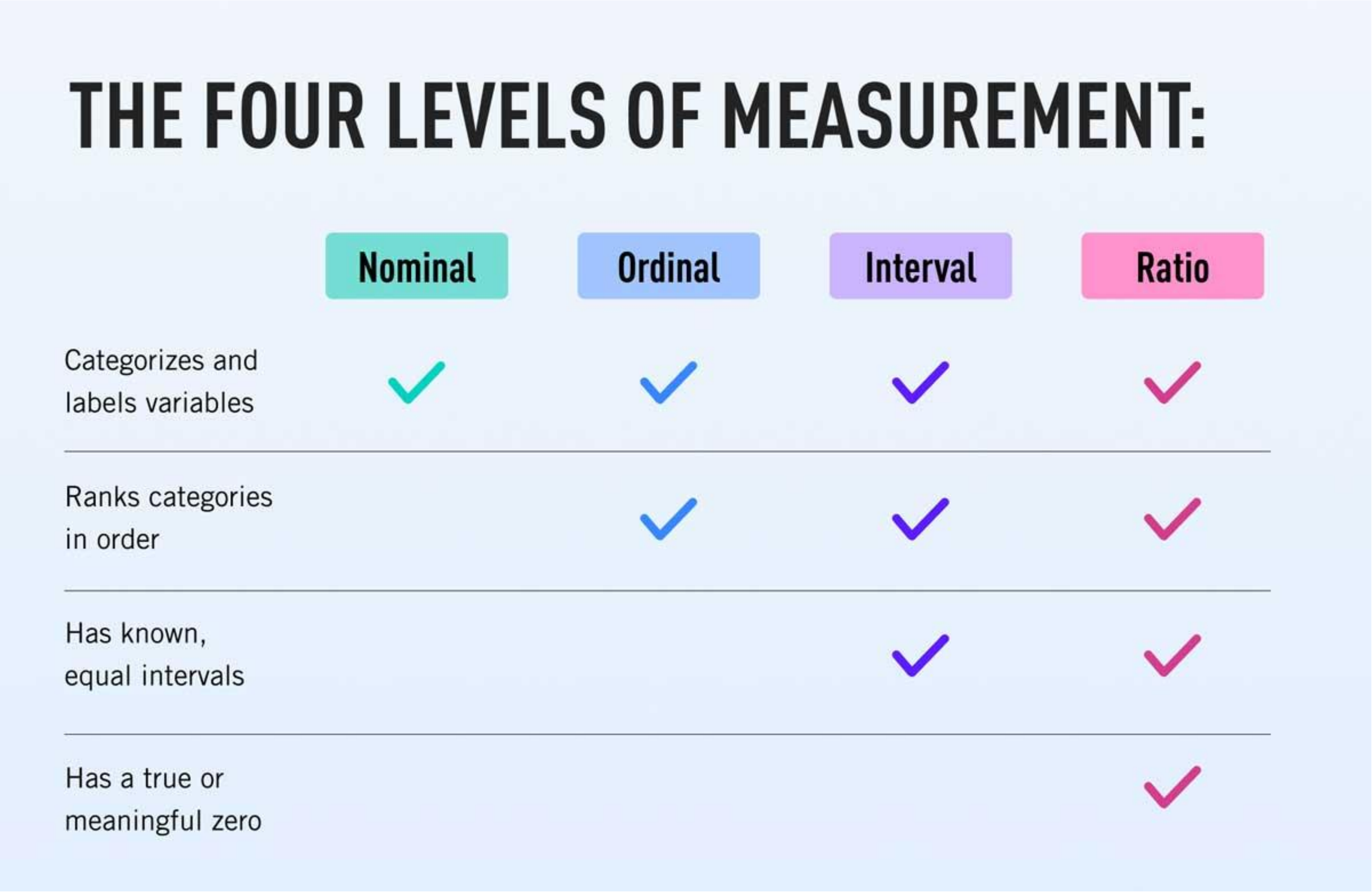
Nominal data is the simplest data type. It classifies data purely by labeling or naming values e.g. measuring marital status, hair, or eye color. It has no hierarchy to it.

Ordinal data classifies data while introducing an order, or ranking. For instance, measuring economic status using the hierarchy: 'wealthy', 'middle income' or 'poor.' However, there is no clearly defined interval between these categories.

Numerical Data

Interval data classifies and ranks data but also introduces measured intervals. A great example is temperature scales, in Celsius or Fahrenheit. However, interval data has no true zero, i.e. a measurement of 'zero' can still represent a quantifiable measure (such as zero Celsius, which is simply another measure on a scale that includes negative values).

Ratio data classifies and ranks data, and uses measured intervals. However, unlike interval data, ratio data also has a true zero. When a variable equals zero, there is none of this variable. A good example of ratio data is the measure of height - you cannot have a negative measure of height.



THE FOUR LEVELS OF MEASUREMENT:

	Nominal	Ordinal	Interval	Ratio
Categorizes and labels variables	✓	✓	✓	✓
Ranks categories in order		✓	✓	✓
Has known, equal intervals			✓	✓
Has a true or meaningful zero				✓

Variables

- **Variables**: characteristics of interest collected on the elements of population or sample (***observations***).

Example: colour of the eyes, number of brothers and sisters, credit on the mobile phone, height, weight, mark at the last exam, etc…

Types of Variables

1. **Qualitative (or Categorical)** : cannot be measured numerically

   Ex: colors of the eyes, gender, brand of a car, place of birth, etc..

2. **Quantitative** : can be measured numerically

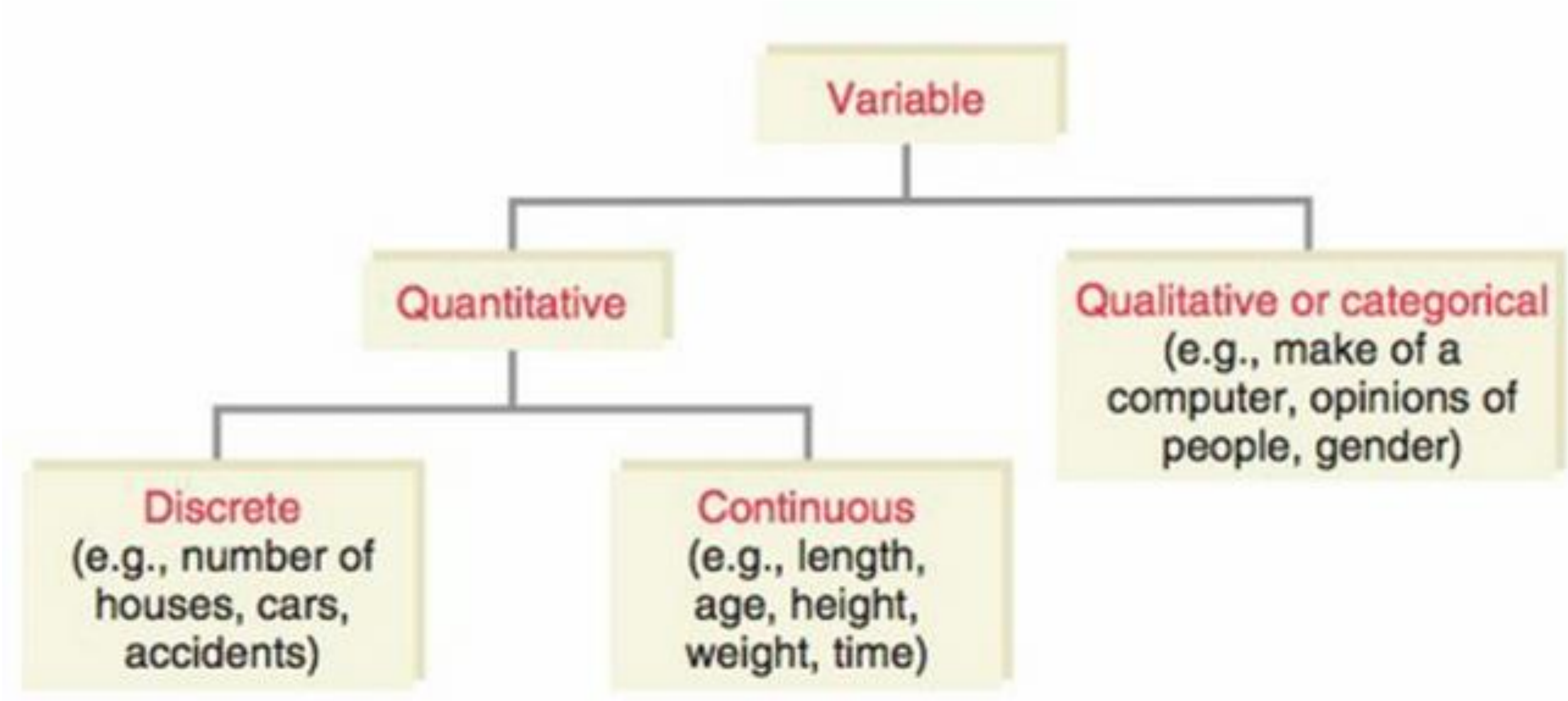
   - **Discrete**: countable values (no intermediate values between integers)

     Ex: nr. of cars held, nr. of people visiting a bank branch every day, number of students in a class, …

   - **Continuous** : assume any numerical values

     Ex: height, weight, time to reach TV, income (and all monetary amounts),…

Types of Variables



Variable
Quantitative
Qualitative or categorical (e.g., make of a computer, opinions of people, gender)
Discrete (e.g., number of houses, cars, accidents)
Continuous (e.g., length, age, height, weight, time)

Discrete and continuous quantitative variables

Let X be a quantitative variable. We can further distinguish between

- **Discrete variables:** the values x that X can take are either finite or countably infinite, i.e., nr. of children per family, nr. of students in a class, nr. of citizens of a country.
- **Continuous variables:** X can take on an uncountable set of values x, i.e., the amount of time required to complete a project, the height of children, the amount of rain, in inches, that falls in a storm, the square footage of a two-bedroom house, the weight of a truck, the speed of cars.

Types of Variables: Exercise

Classify the following variables:

VARIABLE	TYPE
The time a student spends studying for an exam (in hours)	
The amount of rain last year in Rome (in ml)	
The arrival status of an airline flight at an airport (early, on time, late, canceled)	
The price of a BigMac in 10 different cities	

Types of Variables: Exercise

Classify the following variables:

VARIABLE	TYPE
The time a student spends studying for an exam (in hours)	Quant. Continuous
The amount of rain last year in Rome (in ml)	Quant. Continuous
The arrival status of an airline flight at an airport (early, on time, late, canceled)	Qualitative
The price of a BigMac in 10 different cities	Quant. Continuous

Types of Variables: Exercise

Classify the following variables:

VARIABLE	TYPE
The amount of gasoline put into a car in a gas station	
The number of clients in a bank branch	
The brand of a car	
The number of coffees drunk in a day	

Types of Variables: Exercise

Classify the following variables:

VARIABLE	TYPE
The amount of gasoline put into a car in a gas station	Quant. Continuous
The number of clients in a bank branch	Quant. Discrete
The brand of a car	Qualitative
The number of coffees drunk in a day	Quant. Discrete

Types of Variables: Stocks and Flows

According to the time stamp of the measurement we distinguish also:

- **Stock variables:** they can be measured only with reference to a specific time point (residents in a country, value of capital stock, wealth, marital status, employment status, etc.)

- **Flow variables:** they can be measured only with reference to a time interval (production, sales, income, etc.)

Stocks and Flows variables

- Wealth is a **stock**, income is a **flow**
- Kilowatt hours (e.g. stored in a Tesla battery) are a **stock**, Kilowatts are a **flow** (e.g. current charging or discharging the battery)
- The amount of gold in a reserve is a **stock**, the mining of it is a **flow**
- The population of a country is a **stock**, birth rates, death rates and migration rates are **flows**
- The inventory in a warehouse is a **stock**, orders taken from it are **flows**

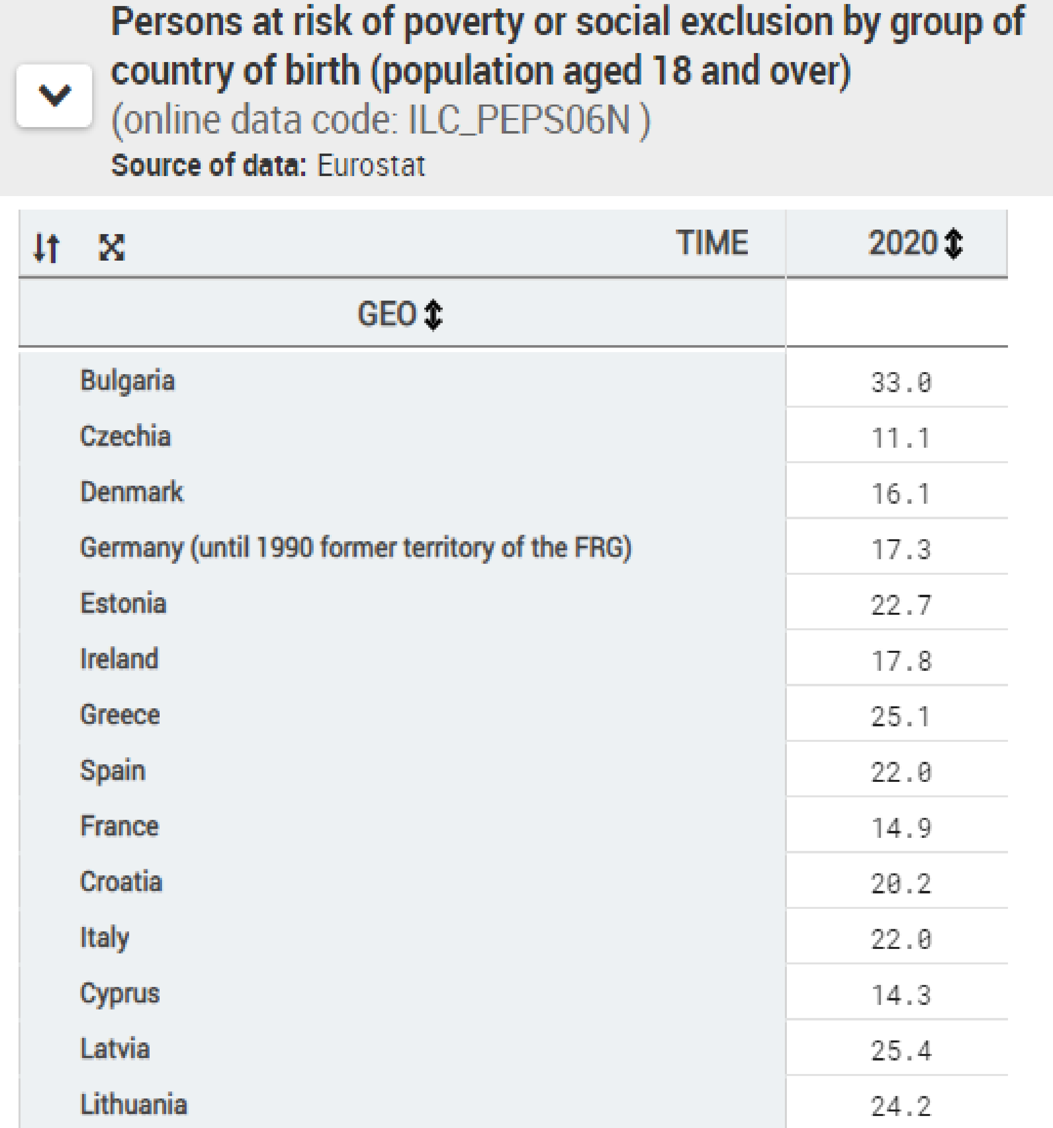
Types of Data

1. **Cross-Section**: data are collected for many elements ($n = 1, 2, 3, \ldots, N$) in the same period of time ($t=1$)

2. **Time-Series**: data are collected for the same element ($n =1$) at different points in time ($t= 1,2,3,\ldots,T$)

3. **Panel Data**: data are collected for many elements ($n = 1, 2, 3, \ldots, N$) at different points in time ($t= 1,2,3,\ldots,T$)

Cross-section: Example



Persons at risk of poverty or social exclusion by group of country of birth (population aged 18 and over)
(online data code: ILC_PEPS06N)
Source of data: Eurostat

GEO \ TIME	2020
Bulgaria	33.0
Czechia	11.1
Denmark	16.1
Germany (until 1990 former territory of the FRG)	17.3
Estonia	22.7
Ireland	17.8
Greece	25.1
Spain	22.0
France	14.9
Croatia	20.2
Italy	22.0
Cyprus	14.3
Latvia	25.4
Lithuania	24.2

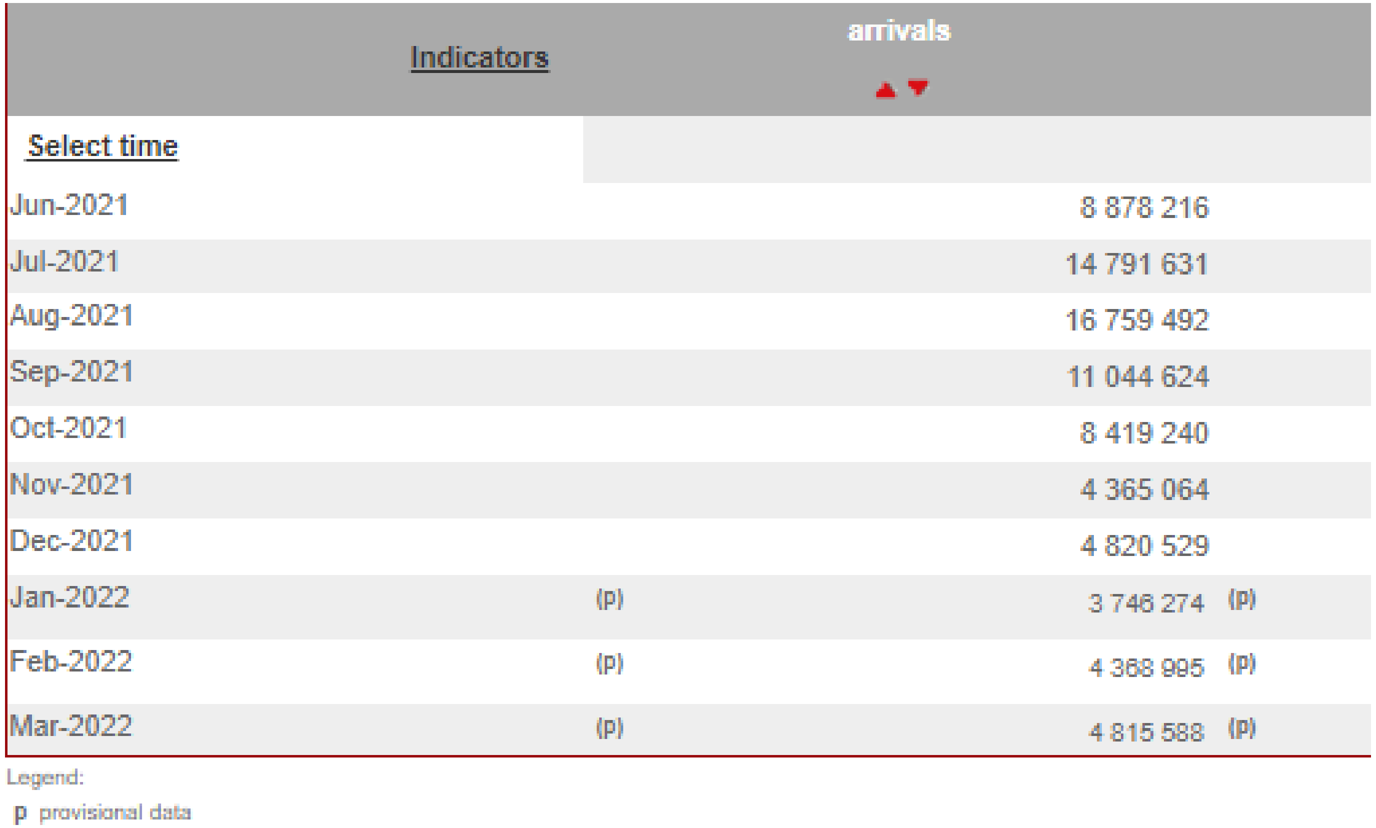
1.**Cross-Section**: data are collected for many elements (n = 1, 2, 3, …, N) in the same period of time (t=1)

Percentage of people at risk of poverty or social exclusion by group of country of birth (population aged 18 and over), European countries (n=1,…,27), t=2020

https://ec.europa.eu/eurostat/databrowser/bookmark/626f6b49-a809-46a2-883d-169d19359f3c?lang=en

Time-Series: Example

2. Time-Series: data are collected for the same element (n=1) at different points in time (t= 1,2,3,…,T)



Indicators		arrivals ▲▼	
Select time			
Jun-2021		8 878 216	
Jul-2021		14 791 631	
Aug-2021		16 759 492	
Sep-2021		11 044 624	
Oct-2021		8 419 240	
Nov-2021		4 365 064	
Dec-2021		4 820 529	
Jan-2022	(p)	3 746 274	(p)
Feb-2022	(p)	4 368 995	(p)
Mar-2022	(p)	4 815 588	(p)

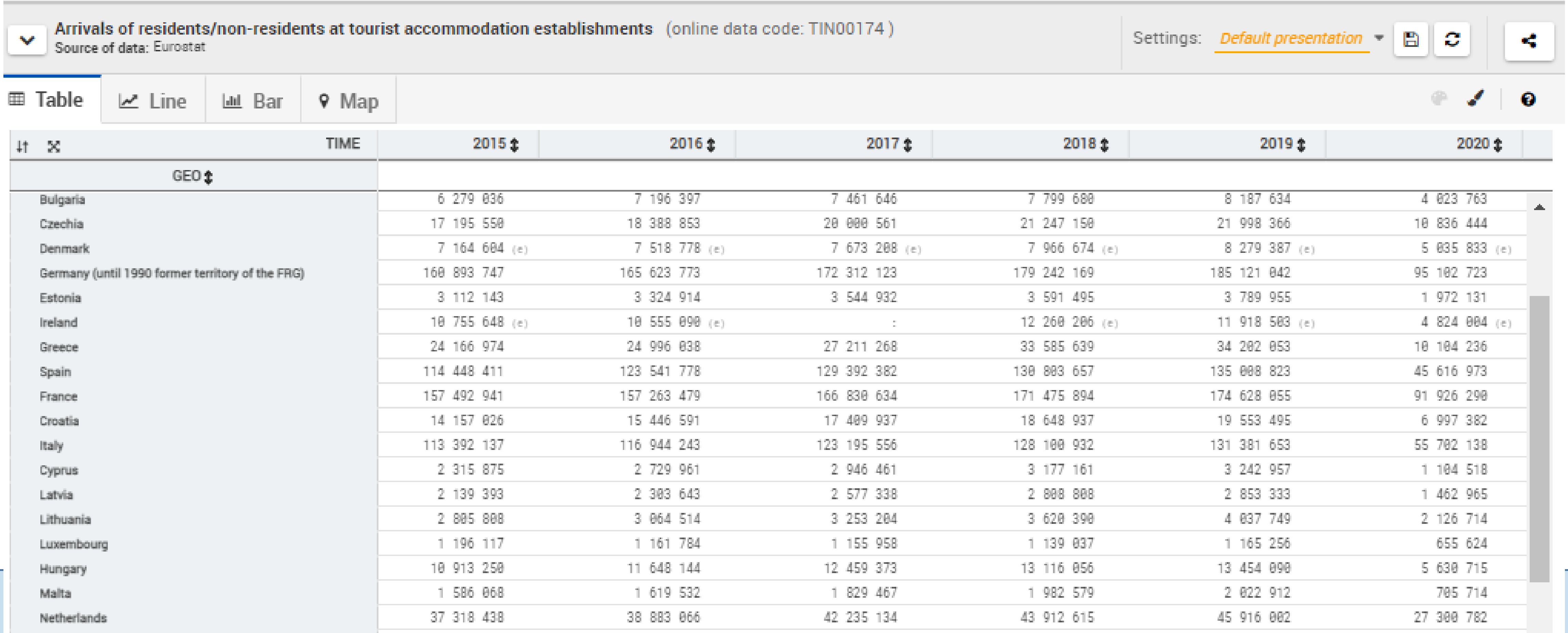
Legend:
p provisional data

Monthly data on arrivals in accommodation establishments in Italy (n=1), from 2021-06 to 2022-06 (t=2021-06,2021-07,…, 2022-06)

Ex. http://dati.istat.it/?lang=en&SubSessionId=2c85fd67-7c39-4c50-8bfa-bed39a447064

Panel Data: Example

3. **Panel Data**: data are collected for many elements (n=1, 2, 3, …, N) at different points in time (t= 1,2,3,…,T) Ex.https://ec.europa.eu/eurostat/databrowser/bookmark/fe477a47-8142-4d9b-ba80-ba2e6b55a71d?lang=en



Arrivals of residents/non-residents at tourist accommodation establishments (online data code: TIN00174)
Source of data: Eurostat

Settings: Default presentation

TIME GEO	2015	2016	2017	2018	2019	2020
Bulgaria	6 279 036	7 196 397	7 461 646	7 799 680	8 187 634	4 023 763
Czechia	17 195 550	18 388 853	20 000 561	21 247 150	21 998 366	10 836 444
Denmark	7 164 604 (e)	7 518 778 (e)	7 673 208 (e)	7 966 674 (e)	8 279 387 (e)	5 035 833 (e)
Germany (until 1990 former territory of the FRG)	160 893 747	165 623 773	172 312 123	179 242 169	185 121 042	95 102 723
Estonia	3 112 143	3 324 914	3 544 932	3 591 495	3 789 955	1 972 131
Ireland	10 755 648 (e)	10 555 090 (e)	:	12 260 206 (e)	11 918 503 (e)	4 824 004 (e)
Greece	24 166 974	24 996 038	27 211 268	33 585 639	34 202 053	10 104 236
Spain	114 448 411	123 541 778	129 392 382	130 803 657	135 008 823	45 616 973
France	157 492 941	157 263 479	166 830 634	171 475 894	174 628 055	91 926 290
Croatia	14 157 026	15 446 591	17 409 937	18 648 937	19 553 495	6 997 382
Italy	113 392 137	116 944 243	123 195 556	128 100 932	131 381 653	55 702 138
Cyprus	2 315 875	2 729 961	2 946 461	3 177 161	3 242 957	1 104 518
Latvia	2 139 393	2 303 643	2 577 338	2 808 808	2 853 333	1 462 965
Lithuania	2 805 808	3 064 514	3 253 204	3 620 390	4 037 749	2 126 714
Luxembourg	1 196 117	1 161 784	1 155 958	1 139 037	1 165 256	655 624
Hungary	10 913 250	11 648 144	12 459 373	13 116 056	13 454 090	5 630 715
Malta	1 586 068	1 619 532	1 829 467	1 982 579	2 022 912	705 714
Netherlands	37 318 438	38 883 066	42 235 134	43 912 615	45 916 002	27 300 782