Statistics

Vicent Giner-Bosch
Universitat Politècnica de València, Spain

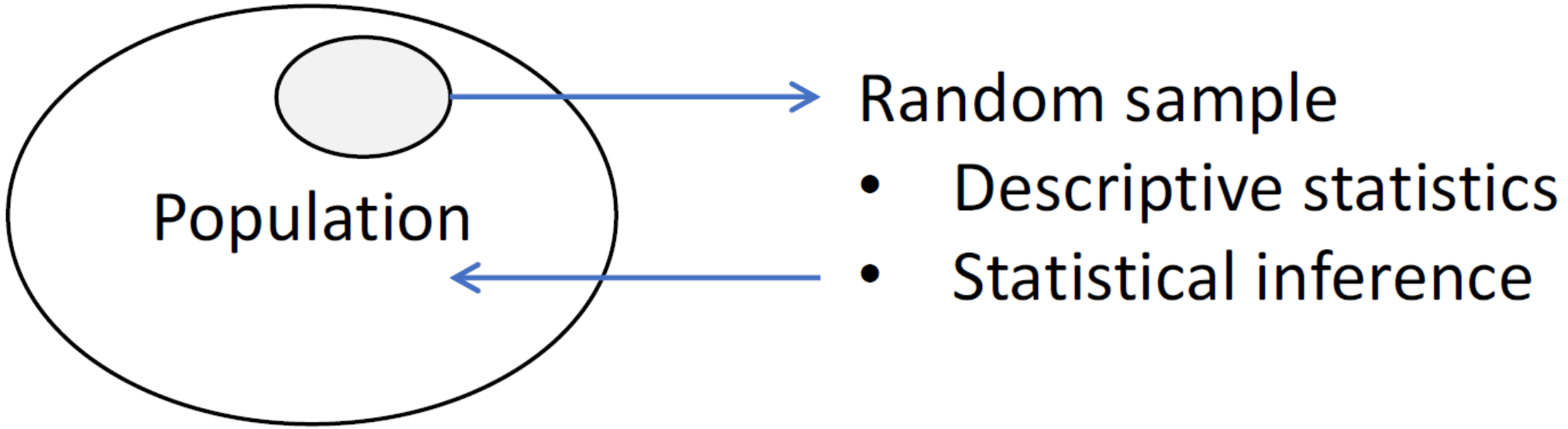
About me

- Associate Professor at the UPV (Valencia, Spain).
- 20+ years teaching Statistics and Operational Research.
- More about me at vigibos.webs.upv.es.

What and what for

- In science and engineering, we need to collect and interpret data in order to have a better knowledge of the system, product, process, situation, population... under study.
- Statistics provides us with a perfect toolbox for achieving this aim.
- It mainly deals with random phenomena.
- It plays a major role in all the stages of the data mining process.

Population vs sample



Data Mining process

Cleaning	Handling missing data and noisy data, including outliers and inconsistencies.
Integration	Combining multiple, heterogeneous data sources (databases, files, etc.).
Reduction	Dimensionality reduction, sample reduction, data compression.
Transformation	Smoothing, aggregating, normalising, discretisation.
Mining	Identifying patterns and knowledge from large datasets, using supervised (prediction, classification and forecasting) and unsupervised (clustering) learning techniques.
Evaluation	Identifying the truly interesting patterns representing knowledge based on different types of interesting measures.
Representation	Communicating mined knowledge in the form of reports, tables and dashboards.

What people are saying

"Being able to apply Statistics is like having a secret superpower."

— Evan Miller, statistical software developer (@EvMill)

"Maths and, especially, Statistics are the science behind the science."

— Anabel Forte-Deltell, Universitat de València (@AnaBayes)

What people are saying

"While big data promise much to scientists, entrepreneurs and governments, they are doomed to disappoint us if we ignore some very familiar statistical lessons."

— Tim Harford, Economics communicator (@TimHarford)

"There are a lot of small data problems that occur in big data. They don't disappear because you've got lots of the stuff. They get worse."

— David Spiegelhalter, University of Cambridge (@d_spiegel)

Program schedule

FIRST PART: Descriptive Statistics and Statistical Inference

- UNIT 1: Descriptive Statistics.
- UNIT 3: Probability distributions.
- UNIT 4: Statistical inference.

SECOND PART: Advanced features

- UNIT 5: Design of experiments and ANOVA.
- UNIT 6: Regression.
- UNIT 7: Time series.

Lesson schedule: 2023

2023	MON 23 OCT	TUE 24 OCT	WED 25 OCT	THU 26 OCT	FRI 27 OCT	2023
8–9						8–9
9–10						9–10
10–11						10–11
11–12						11–12
12–13						12–13
13–14						13–14
14–15						14–15
15–16						15–16
16–17						16–17
17–18						17–18

Methodology

- Class & lab sessions.
- Practical exercises.
- All done with R (statistical software).
  - It is both a programming language and a software environment.
  - It is free. Many people (researchers, scientists) are contributing to its development by sharing their own pieces of code.
  - It is within the top 3 tools for analytics, data science and machine learning, together with Python and Rapid Miner (KDnuggets, 2019).
  - It is a must-know for a data scientist (KDnuggets, 2018).

Evaluation

- 60% Lab sessions.
  - Short assignments to do at class using R.
  - In pairs.
- 40% Final exam.
  - Written exam containing R screenshots.
  - You will be allowed to use a single A4 note sheet handwritten both sides.

FINAL EXAM 2023: Thursday, 2nd November, 10:30.

Behaviour policy

THE THREE P's

Three basic rules to follow at class:

1. Punctuality. Try to be on time at class.
2. Paying attention. Try to be silent and to listen when either the teacher or your classmates are talking.
3. Plane mode. Avoid distractions! Switch on the airplane mode on your phone.

Let’s begin!