Introduction

Statistical Learning

Welcome to Statistical Learning

Statistical learning comes in two variants:

- Statistical Learning M [F8204B033M]. This is a self-contained course of 6 CFU
- Statistical Learning [F8204B015]. This is part of the course Data Science (12 CFU), which is composed by Data Mining (6 CFU) and Statistical Learning (6 CFU)

Course objectives

Acquisition of modern statistical methods for inference on complex data set, and improvement of

- problem solving
- programming skills in R

In particular:

- Formulate quantitative models to address scientific questions
- Apply a range of statistical methods for inference
- Organize and perform a complete data analysis, from data exploration, to analysis, to synthesis, to communication

The expected workload is $\approx 1/3$ applied and $\approx 2/3$ theorical.

Pre-requisites

This course is designed for 2nd year CLAMSES students. It assumes knowledge of the topics of the courses *Probabilità e Statistica Computazionale M* (modules Probabilità Applicata and Statistica Computazionale) and *Statistica Avanzata M* (modules Statistica Multivariata and Teoria dell'Inferenza Statistica).

Course material

You can download the material from the course webpage

https://github.com/aldosolari/SL

Please keep in mind that

- the material is subject to updates: check the last version
- the material will be uploaded progressively

Main references

- [EH] Efron and Hastie (2016) Computer-Age Statistical Inference: Algorithms, Evidence, and Data Science, Cambridge University Press
- [GI] Giraud C. (2015) Introduction to High-Dimensional Statistics. Chapman and Hall/CRC

- etc.

Exam

The exam is divided in two parts:

1. Data analysis assignment
2. Oral exam: 1 question about 1. and 3 questions about topics discussed during the course

In particular, the data analysis assignment requires that before the oral exam, each student must submit

- A write-up of their data analysis in a synthesized format, with numbered figures and references. (You may also include supplementary material for detailed additional calculations/analyses)
- A reproducible .Rmd file that produces all of the numbers, figures and results in your write-up.

All documents should be submitted electronically. The grades will be broken down according to the following characterization of your data analysis.

1. Did you answer the scientific question?
2. Did you use appropriate statistical methods?
3. Was your write-up simple, clear, and precise?
4. Was your code reproducible?

Keep in mind that this is a data science class. In some cases standard methodology will be sufficient to answer the question of interest, in some cases you will need to go beyond the course.

You may speak to your fellow students about specific statistical questions related to the project, but the overall idea, analysis, and write-up should be your own individual work.

In general the goal is to *answer the question* and provide an estimate of *uncertainty*.

One data set, many analysts

In 2014, 29 teams of researchers were recruited and they were asked to answer the same research question with the same data set.

Figure 1: Image from Silberzahn and Uhlmann (2015)

Research question

> Are football (soccer) referees more likely to give red cards to players with dark skin than to players with light skin?

This question touches on broad issues, such as how prejudice affects sports and how well the effects of prejudice, as detected in laboratory settings, show up in the real world.

Figure 2: Mario Balotelli, playing for Manchester City, is shown a red card during a match against Arsenal. Ph. by MICHAEL REGAN/GETTY

Data

All teams were given the same large data set collected by a sports-statistics firm across four major football leagues.

The data can be downloaded from

https://osf.io/47tnc/

A description of the data is

https://github.com/aldosolari/SL/tree/master/assignment/README.txt

It included referee calls, counts of how often referees encountered each player, and player demographics including team position, height and weight.

It also included a rating of players' skin colour. As in most such studies, this ranking was performed manually: two independent coders sorted photographs of players into five categories ranging from 'very light' to 'very dark' skin tone.

Results

Teams approached the data with a wide array of analytical techniques, and obtained highly varied results.

Of the 29 teams, 20 found a statistically significant correlation between skin colour and red cards.

The median result was that darkskinned players were 1.3 times more likely than light-skinned players to receive red cards.

But findings varied enormously, from a slight (and non-significant) tendency for referees to give more red cards to lightskinned players to a strong trend of giving more red cards to dark-skinned players.

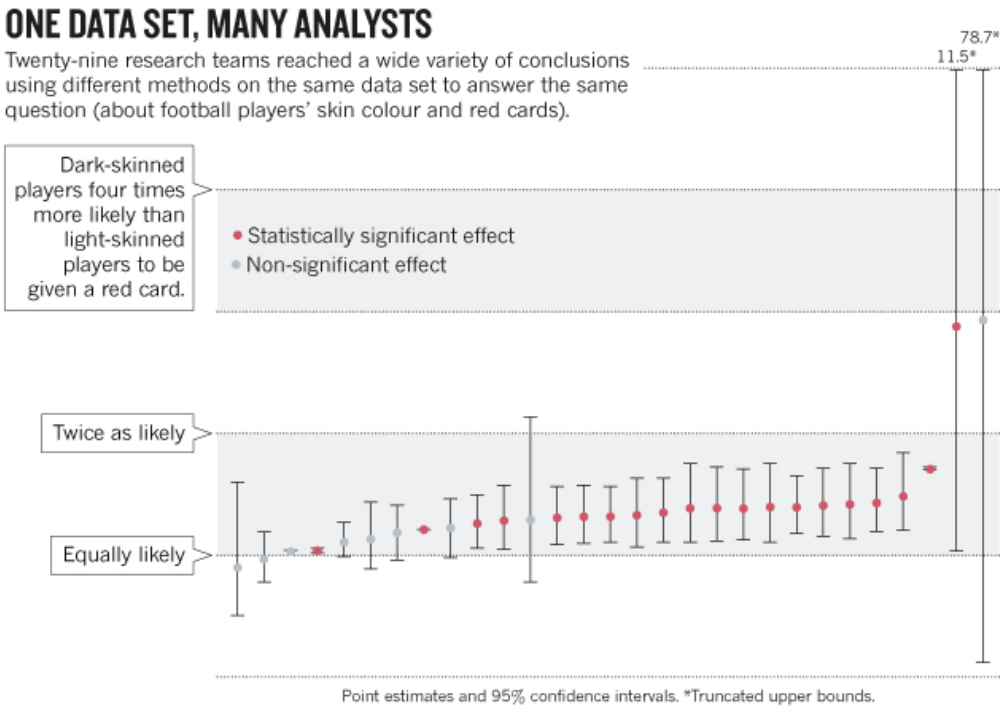

Figure 3: Figure from Silberzahn and Uhlmann (2015)

References

- Silberzahn and Uhlmann (2015) Crowdsourced research: Many hands make tight work. Nature 526, 189–191
- Jeff Leek (2015) Advanced Data Science course webpage http://jtleek.com/advdatasci/index.html