2pm Tuesday

Eng Maths 2

Applied Statistics

Lecture 1

David Barton & Sabine Hauert

Department of Engineering Mathematics

Why applied statistics?

Why applied statistics?

Engineers have long considered statistics to be evil *but that is changing!*

- High performance engineering — error margins are getting smaller!
- Uncertainty quantification (UQ) is a hot topic (particularly aerospace)
- Products/machines/people are complicated to model
- Data is pouring out of our ears...

This course aims to give you a grasp of the *basics* and an idea of what can go wrong

Statistics in the real world

Moderate coffee drinking is safe, and three to four cups a day may have some health benefits, according to a large review of studies, in the BMJ.

The real story — www.popsci.com/coffee-drinkers-live-longer

Other examples?

Calling Bullshit: Data Reasoning in a Digital World

Logistics

Course: INFO 198 / BIOL 106B, University of Washington
To be offered: Autumn Quarter 2017
Credit: 3 credits, graded
Enrollment: 180 students
Instructors: Carl T. Bergstrom and Jevin West
Synopsis: Our world is saturated with bullshit. Learn to detect and defuse it.

callingbullshit.org

Calling Bullshit — Learning Objectives

From their website —

Our learning objectives are straightforward. After taking the course, you should be able to:

- Remain vigilant for bullshit contaminating your information diet.
- Recognize said bullshit whenever and wherever you encounter it.
- Figure out for yourself precisely why a particular bit of bullshit is bullshit.
- Provide a statistician or fellow scientist with a technical explanation of why a claim is bullshit.
- Provide your crystals-and-homeopathy aunt or casually racist uncle with an accessible and persuasive explanation of why a claim is bullshit.

We will be astonished if these skills do not turn out to be among the most useful and most broadly applicable of those that you acquire during the course of your college education.

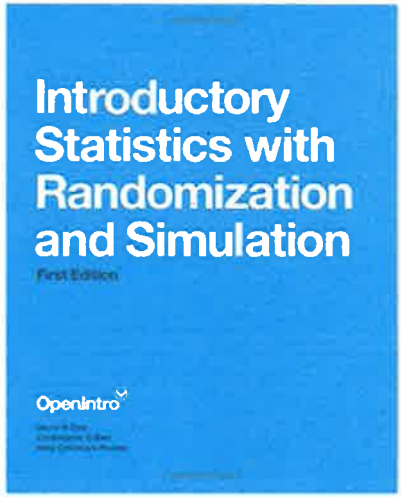
Housekeeping

- 18 lectures
  - Weeks 16, 17, 19–22 (break for reading week)
- Exercise sheets
  - Exercises from the text book
  - Multiple exercise sheets (extra — not essential)...
- Marked homework
- Exam
  - Section A — 5× 8 mark questions (compulsory)
  - Section B — 6× 20 mark questions (answer 3)
    - Must answer one on each topic!
- I'm assuming you know probability...
  - Will do some recapping as we go along

Textbooks

OpenIntro Statistics

Introductory Statistics with Randomization and Simulation



www.openintro.org — free to download!

What are statistics?

From Wikipedia(!)

Statistics

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data.

Statistic

A statistic (singular) or sample statistic is a single measure of some attribute of a sample (e.g., its arithmetic mean value).

→ A single number that describes data

The Royal Statistical Society UK statistic of the year (2018) is 27.8%

Types of statistic

Numerical

The usual statistics: mean, median, mode, standard deviation, ...

Discrete e.g., integer values — median number of people in the queue

Continuous e.g., real values — average number of people in the queue

Categorical

Counting how many individuals fall into a category, e.g.,

Category 1: number of people waiting to buy a coffee

Category 2: number of people waiting to buy something else

Data: individuals, samples, and populations

Individual The basic unit of observation

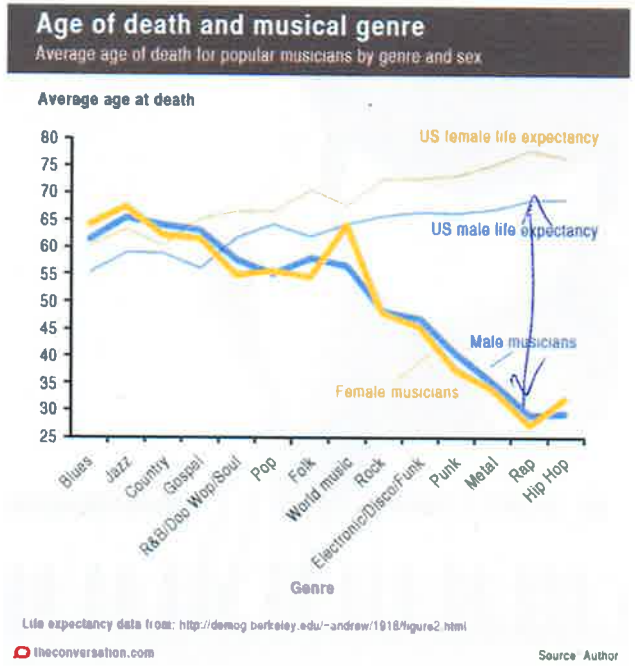
Population All the individuals

Sample All the individuals you can (or are willing to) observe

Identify the individual, population, and an *appropriate* sample to answer the questions —

1. How long on average does it take to graduate from Bristol University?
2. What is the average lifespan of an iPhone?
3. How likely am I to get the flu?

Problems with sampling



https://www.youtube.com/watch?v=sxYrzzy3cq8

Basic principles of experimental design

Question (city planning) — does a proposed change to traffic light phasing improve the traffic flow?

Controlling. Minimise systematic differences between treatment groups.

↳ Day of week → impact on traffic → weekends vs. weekdays.
↳ School holidays.
↳ Remove as much variation as possible. → 1. before the change

Randomisation. Randomise the individuals between treatment groups.

↳ Randomise which weekday day (Mon-Fri) ↳ 2. after the change

Replication. Include as many individuals in the sample as possible.

↳ Replicate while randomising.
↳ often most expensive!

Basic principles of experimental design

Blocking. Divide the sample into sub-groups based on known (significant) differences between individuals; randomly allocate equal numbers from each sub-group into each treatment group.

↳ Randomise everything within the weekdays group.

Basic principles of experimental design

Question (smartphones) — on average, does an Android phone outlast an iPhone?

// vague?!

Individuals? Population? Treatment groups? Controls? Randomisation? Blocking? (Assuming as much replication as possible.)

Individual: Measuring the lifespan of the phone.
Population: All broken Android or iphones.

Profession.
Age group.
Daily use.

Answering statistical questions

From an experiment with 10 different phones we have the following phone lifetimes (in years)

Android	iPhone
5.0	5.3
3.9	5.8
7.6	6.2
4.8	11.1
8.9	3.2
6.04	6.32

Average

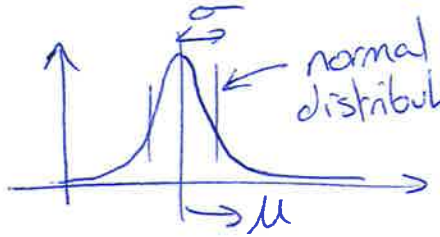
What can be concluded from this?

→ is this difference statistically significant?

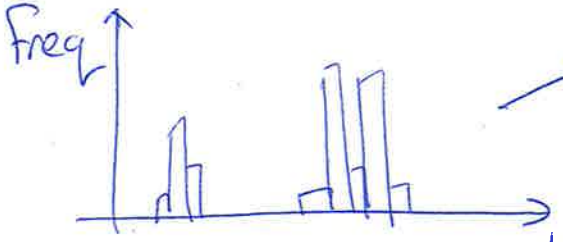
Parametric models vs non-parametric models

Parametric models — use the data to estimate the parameters of the population (mean, variance, etc) assuming some distribution → Good for small amounts of data.

→ Histogram of the data.



Non-parametric models — the data is the model; assumes the data is sufficiently representative (e.g., you have lots of data)



→ not normal!

requires a computer

Quote of the day

Mark Twain

Facts are stubborn things, but statistics are more pliable.

Exercises

- One question from each of the subsections in §1.9 of OpenIntro Statistics.