Stat 234: Statistical Models and Methods I

David Gerard

2017-09-21

Learning Objectives

- Three aspects of Statistics
- Population/Sample

Statistics

Statistics — the field of answering questions using data.

Data — Numerical or qualitative descriptions of people/places/things that we want to study.

My own personal projects:

- Phylogenetic inference
  - Data: DNA sequences from individuals from three subpopulations of ratsnakes.
  - Question: Is one of these three populations a hybrid of the other two?
- Gene expression analysis
  - Data: Measurements of disease status and how "turned on" or "turned off" many genes are for many individuals.
  - Question: Which genes are more "turned on" when you have a disease?

Statistics

Statistics — the field of answering questions using data.

Some more examples

- Google search
  - Data: Billions of search queries and user satisfaction of the results.
  - Question: What results does a user want from a query?
- Fraud Detection
  - Data: Collection of financial records for a large corporation.
  - Question: Is there evidence for fraud?

Statistics

Three aspects:

1. Data Design
2. Data Description
3. Data Inference

Statistics

Three aspects:

2. Data Description
3. Data Inference

Statistics

Three aspects:

2. Data Description
3. Data Inference — informed by Probability

Data Design

Where do we get data?

- What is the proper way to collect data?
- When can we claim a causal connection between variables? (e.g. Does smoking contribute to cancer? Does better self esteem make students learn better?)
- What are some sources of bias (unwanted systematic tendencies in the data collection)?
- Only touched on in this course.

Data Description

How do we describe the data we have?

- Numerical summaries — use numbers to describe the data.
- Graphical summaries — use pictures to describe the data.
- Exploratory data analysis — play with the data to get a "feel" for it.
- Lots of R.
- First two weeks of the course.

Data Inference (Probability)

How can we tell if our conclusions from the exploratory data analysis are real?

- Last eight weeks of the quarter.
- Probability — subdiscipline of Mathematics that provides a foundation for modeling chance events.
- Inference — describing a population (probabilistically) by using information from sample.

Population

Statisticians (among others) are interested in characteristics of a large group of people/countries/objects

- Characterize/describe income of U.S. residents.
- Characterize/describe success rate of startups.
- Characterize/describe tastiness of burgers.

population

A population is a group of individuals/objects/locations for which you want information.

Sample

It is usually expensive/impossible to measure characteristics of every case in a population.

Sample

A sample is a subgroup of individuals/objects/locations of the population.

- Measure income from 50 US adults.
- Look up time to IPO of 100 startups.
- Eat 3 burgers.

Inference

From the sample, describe the population using probability.

- "Using a procedure that would capture the true average income of U.S. residents 95% of the time, we say the mean income is somewhere between 51,502 and 52,498."
- "Using a procedure that is only wrong 5% of the time, we reject the hypothesis that startups are more likely to succeed than fail."
- "With overwhelming confidence ($p < 0.001$), we say burgers taste good."

Inference

From the sample, describe the population using probability.

- "Using a procedure that would capture the true average income of U.S. residents 95% of the time, we say the mean income is somewhere between 51,502 and 52,498."
- "Using a procedure that is only wrong 5% of the time, we reject the hypothesis that startups are more likely to succeed than fail."
- "With overwhelming confidence ($p < 0.001$), we say burgers taste good."
- In this class, we will learn what these statements mean and how to make our own inference statements.

Books: Both are FREE

- OpenIntro:
  https://www.openintro.org/stat/textbook.php?stat_book=os
  - For data description and data inference.
  - Low level.
- From Algorithms to Z-Scores: Probabilistic and Statistical Modeling in Computer Science:
  http://heather.cs.ucdavis.edu/probstatbook
  - Haven't decided if we'll use it yet.
  - For probability.
  - Slightly higher level.

Resources for learning R

YOU WILL NEED TO LEARN R ON YOUR OWN FOR THIS COURSE

- swirl: http://swirlstats.com/.
  - R package with nice interactive tutorials for learning R basics.
- Code School: http://tryr.codeschool.com/.
  - Another interactive introduction to R.
- R Cheat Sheets with important functions: https://www.rstudio.com/resources/cheatsheets/.
- R Base Graphics Cheat Sheet: https://dcgerard.github.io/stat234/base_r_cheatsheet.html.
- R-tutorial: This Wednesday and Thursday from 6pm to 8pm in Eckhart 133. You should go to one if you can.

Administrative Slide

- Math prerequisites: Read through Part 0 of math supplement. Do a few exercises.
- My office hours: Tuesdays 11:00 am – 12:20 pm in Jones 226.
- Get everything else from the syllabus on Canvas.