# 00 Course Outline for Stat 614

David Gerard

2018-12-07

# 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.

# Statistics

Statistics — the field of answering questions using data.

Some 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 — 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 week of the semester.

# Data Inference (Probability)

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

- Last thirteen weeks of the semester.
- 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 the effectiveness of a drug on a all adults.

**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.
- Compare the mortality rate between groups that took and did not take a drug.

# 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."

# 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."
- In this class, we will learn what these statements mean and how to make our own inference statements.