SOC6707: Intermediate Data Analysis

Monica Alexander

Week 1: Introduction

Overview of today

- Overview of course
- Why learn statistics?
- Review concepts
- R and the `tidyverse`
- (Short break)
- Lab

Overview of course

Instructor and TA

Instructor: Monica Alexander (she/her)

- Email: monica.alexander@utoronto.ca
- Office hours: TBA

TA: Julia Igenfeld (she/her)

- Email: julia.ingenfeld@mail.utoronto.ca
- Office hours: TBA

A bit about me

Me:

- statistics $\cap$ chemistry $\rightarrow$ social science $\cap$ statistics
- 50/50 Statistical Sciences and Sociology departments
- Not Canadian (Australia $\rightarrow$ USA $\rightarrow$ Canada)

What I work on: a mix of demography, applied stats, epidemiology and computational social science

Mode of delivery

We will start in **online synchronous** format (for the first 4 lectures)

- Lectures and office hours online through Zoom (links on Quercus)
- Lectures will be recorded such that they can be accessed at a later date

After week 4, ???????

- Hopefully in person in room 240
- Updates as we get them

Objectives

This course introduces statistical techniques and methods to analyze data to draw inferences about social processes. You will learn

- How to read in, describe, plot and analyze data in a statistical software that uses a programming language (R)
- Some important methods of statistical analysis to explore relationships between social phenomena
- How to assess and evaluate the suitability and performance of statistical methods in different contexts

Objectives

Practical objectives (building on from first semester):

- Getting more comfortable using R in the context of the whole research cycle (getting data, reading it in, exploring, modeling)
- Learning how to analyze binary outcomes (maybe more?)
- Statistical literacy

General philosophies

- Understand your data (EDA! Plot!)
- Reproducibility (or close to)
- Knowing when but also when not to use methods (both in terms of your own analysis but also other people's analyses)

This course will be very hands-on with coding and data munging. The learning curve for R is steep but will (hopefully?!) pay off.

Textbooks and other resources

There is no required textbook for this class. Some resources that might be useful:

- R for Data Science: https://r4ds.had.co.nz/ (free)
- Telling stories with data: https://www.tellingstorieswithdata.com/index.html
- Gelman, Andrew; Hill, Jennifer, and Vehtari, Aki. 2020. 'Regression and Other Stories' (This is around $70 on Book Depository).

Software

We will be using the programming language R in this course, through RStudio.

- You should already have these installed from last semester!
- But if not, more info on how to install these on Quercus
- Emphasis on `tidyverse` and RMarkdown, which you may not have used (?) more later

Assessment

- Three assignments (3x15% $=$ 45%)
- Research project (55%)

Assignments

- Data analysis with R
- Interpretation
- Hand in code, instructor/TA should be able to run without errors

Research Project

Choose a data set, research question, and analysis approach

In the research project you will

- Develop a research question based on data set of choice
- Analyze data using methods learned in class
- Present, interpret and summarize findings

Research Project

Worth a total of 55%, but will be graded in four parts:

1. Research question, variables to be used (5%) due with A1
2. EDA (15%) due with A2
3. Analysis (10%) due with A3
4. Final report, which incorporates 1-3 (20%), due at end of semester.
5. Short presentations in final meeting (5%)

More detail in course outline, and as we go along.

Research Project

For you to start thinking about:

- In class we will be covering regression techniques that will allow you to investigate the association between an outcome of interest and multiple covariates (independent variables):
  - outcome could be CONTINUOUS or BINARY (time permitting: more than 2 categories)
  - covariates could be continuous, binary, categorical
- Think of a question of interest –> try and find data –> if you can't find data probably easiest to think of another question :)
- A large part of the project will be
  - coming to terms with using RMarkdown
  - presentation of data exploration and analysis in a clear concise way
  - practice writing a (short) scientific article

Lecture + Lab

- Each week I will lecture for about 1-1.5 hours, then we will have a lab with hands-on practice in R.

Course Policies

- **Communication**: First, see if you can answer your question by checking the syllabus. Second, try to ask questions during class, tutorials, or office hours. Third, there will be a discussion board on Quercus. Fourth, email myself or your TA (please include the course number in the subject line)
- **Accessibility**: visit http://studentlife.utoronto.ca/accessibility as soon as possible.

We’re all out here doing our best

- The current situation makes both learning and teaching challenging
- Try to be understanding of everyone’s sub-optimal situation
- Communication is key

Why learn statistics?

Why learn statistics?

As sociologists, we are trying to understand different aspects of society.

Statistical techniques give us a means to investigate and test research questions and policy impacts across different areas of people's lives.

Example research questions could include

- How is population mobility changing in the era of Covid-19?
- How do people cope with financial hardship?
- How does paid maternity leave affect women's workforce participation?
- Does volunteering increase your sense of wellbeing?

Why learn statistics

It's not just learning what you could do with data, it's learning what not to do with data

- How biases and selection can give misleading conclusions
- When is it inappropriate to use certain techniques

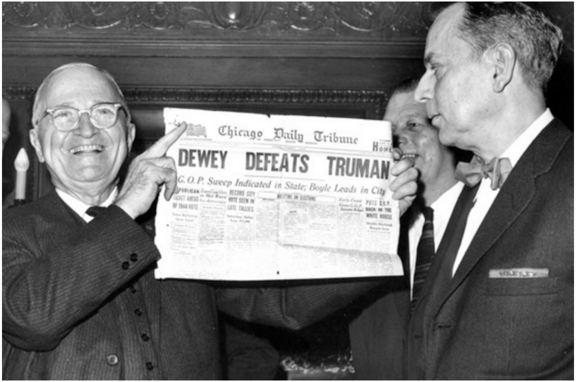
It's not just to support your own arguments, it's learning how to assess other people's arguments

- Statistics, data analysis and visualization is an art form
- Cutting through the lies, damned lies and (misused) statistics

Qualitative and quantitative research

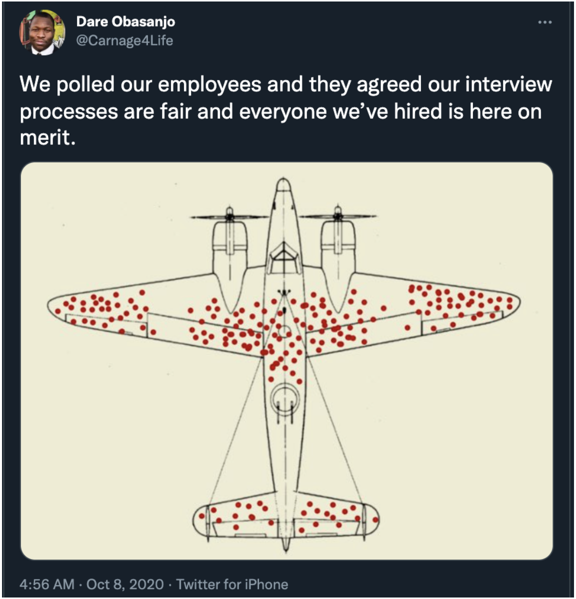
- Different methods of collecting, analyzing, and evaluating evidence to test formal hypotheses
- Often at different scales
- Not mutually exclusive! Mixed methods research is often the hardest and most powerful

Misleading statistics



- Truman won the Presidential election in 1948
- This is a photo of Truman holding an up an erroneous headline
- Based on phone survey which predicted overwhelming win for Dewey
- What went wrong?

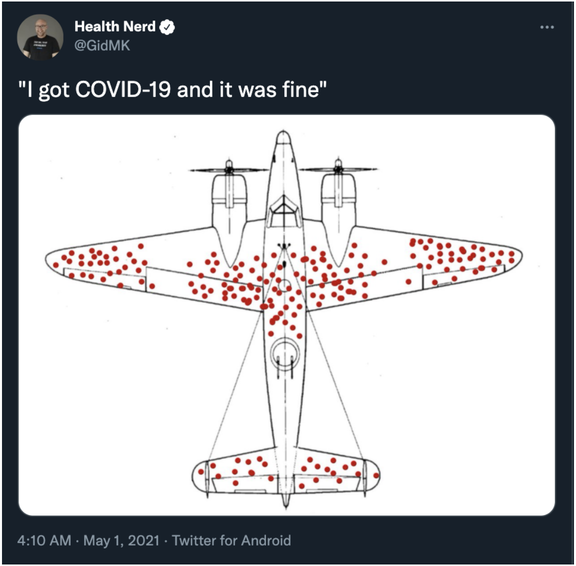
Bad stats becomes a meme



Dare Obasanjo
@Carnage4Life

We polled our employees and they agreed our interview processes are fair and everyone we've hired is here on merit.

4:56 AM · Oct 8, 2020 · Twitter for iPhone

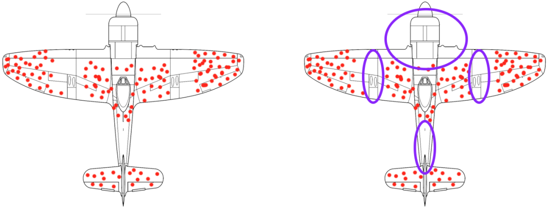


Health Nerd
@GidMK

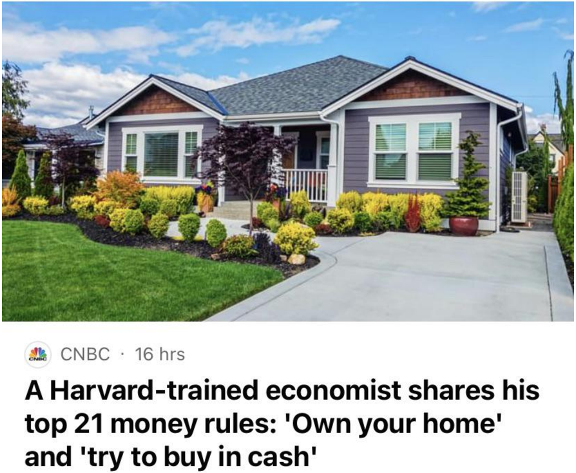
"I got COVID-19 and it was fine"

4:10 AM · May 1, 2021 · Twitter for Android

Misleading statistics

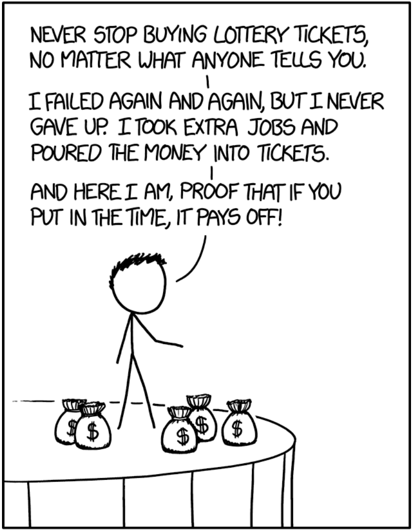


- Abraham Wald in WWII
- Want to place armor on planes in most effective place
- Gathered data from planes returning from battle and observed bullet holes
- Most holes in the fuselage, not so many in the engines
- Where should armor go?
- Can you think of other examples?



CNBC · 16 hrs

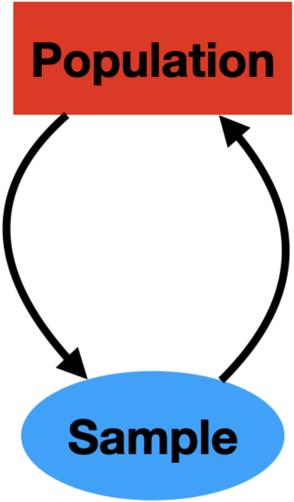
A Harvard-trained economist shares his top 21 money rules: 'Own your home' and 'try to buy in cash'


NEVER STOP BUYING LOTTERY TICKETS, NO MATTER WHAT ANYONE TELLS YOU.
I FAILED AGAIN AND AGAIN, BUT I NEVER GAVE UP. I TOOK EXTRA JOBS AND POURED THE MONEY INTO TICKETS.
AND HERE I AM, PROOF THAT IF YOU PUT IN THE TIME, IT PAYS OFF!
$
$
$
$
$

EVERY INSPIRATIONAL SPEECH BY SOMEONE SUCCESSFUL SHOULD HAVE TO START WITH A DISCLAIMER ABOUT SURVIVORSHIP BIAS.

Review

What it all comes down to



Populations

At the core of statistical methods is wanting to say something about a **population** of interest.

What is a population? Depends on the context of study

- Everyone enrolled in university in Canada
- Everyone at UofT
- Everyone studying graduate-level sociology at UofT
- Everyone in this class

Samples

Say we want to study the relationship between hours studied and job placement for all graduate university students in Canada.

- Not really plausible to get data on this for the whole of Canada.
- In reality, we would collect data on a **subset** or **sample** of the population and try and generalize to the whole of Canada.
- With statistics, we are going to make conclusions based on what we see in the sample that we hope will be true for the population.

Samples

Our example: the relationship between hours studied and job placement for all university students in Canada.

- We could plausibly measure the hours studied and job placement for those student who took SOC6707
- This class would be a sample of the population of interest, because you are all graduate university students in Canada.

Is it a good sample? What do I mean by good?

Sampling techniques

Include:

- **Simple Random Sampling (SRS)**: A random sample is sometimes defined as a sample in which all possible elements have an equal chance of occurring.
- **Stratified Sampling**: based on variable of interest
- **Cluster Sampling**: SRS within clusters (e.g. districts within a province, schools within districts)
- **Convenience Sampling**

What is the sampling method in our example?

Two main domains of statistics

- **Descriptive statistics**: uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.
- **Inferential statistics**: makes inferences and predictions about a population based on a sample of data taken from the population in question.

We will cover both types in this course. Understanding patterns in descriptive statistics is essential to doing good inferential statistics.

Variables

Traits, characteristics, outcomes that we are interested in. e.g.

- hours of study
- course grade
- industry of job placement
- province of residence
- age
- self-reported health

Variables

Often we are interested in studying the relationship between two or more variables.

- The **outcome** of interest is the **dependent variable**
- Variables **used to explain the outcome** can be called
  - independent variables
  - explanatory variables
  - covariates
  - predictors

I will use these terms interchangeably.

What are the independent and dependent variables in our example?

Types of measurement of variables

- **Quantitative**: has a numeric meaning
  - **Continuous**: any possible number
  - **Discrete**: possible values can assume only certain values, usually the counting numbers
- **Qualitative**: categorical, no numeric meaning

What are the types of variables in our example?

Random variables

- A **random variable** is a variable whose values depend on the outcomes of a random process.
- For our purposes, the “random process” is taking a random sample of a population
- For example, consider the variable annual income:
  - We randomly select someone from the population and note their income.
  - The value of this depends on the person who was selected
  - If we randomly selected someone else again, it’s likely that the income value would be different

RVs are the basis of probability and statistical inference!

Some common symbols and notation

I will try to keep notation to a minimum, but there are some common notation that will come up over and over. To start with

- Population size $= N$
- Sample size $= n$
- A particular individual in a sample denoted by index i
- Random variables (note these are captials!):
  - Dependent variable: Y
  - Independent variables X
- A set of random variables for individuals $i = 1, 2, \ldots, n$: $X_1, X_2, \ldots, X_n$
- Specific values or outcomes of the corresponding random variables:
  - Dependent variable: y
  - Independent variables x

Summary measures of quantitative data

Summary measures of quantitative data

Pretend you have a set of observations of a quantitative variable, e.g. everyone's height in this class. We often want to **summarize** our set of observations with one or more numbers. Often interested in:

- Measures of **central tendency**, i.e. what would we expect someone's height to be, what's the most common height
- Measures of **spread**, i.e. what are the ranges of heights observed, what is the deviation of heights away from the expected height?

Measures of central tendency

- **Mean**: the average
  - Population mean usually denoted as μ
  - Sample mean denoted with a bar e.g. $\bar{x}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Median**: the value for which 50% of the sample is below and 50% of the sample is above. It is the 50% percentile. To calculate
  - Order set of values from smallest to largest
  - find the middle number
- **Mode**: the value that occurs the most frequently

Question for you: do you know how to calculate these for a set of numbers in R?

Measures of variability

- **Range**: The difference between the minimum and maximum value
- **Interquartile range**: The difference between the 25% and 75% percentiles. To calculate
  - Order set of values from smallest to largest
  - Separate into quarters
  - Find the first quarter (Q1) and third quarter (Q3)
  - IQR = Q3 - Q1

Measures of variability

- **Variance**: average of the squares of the deviations
  - Population variance: σ^2
  - Sample variance: s^2

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- **Standard deviation**: average of the deviations
  - Population standard deviation: σ
  - Sample standard deviation: s

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{s^2}$$

Introduction to R and R Markdown

R and RStudio

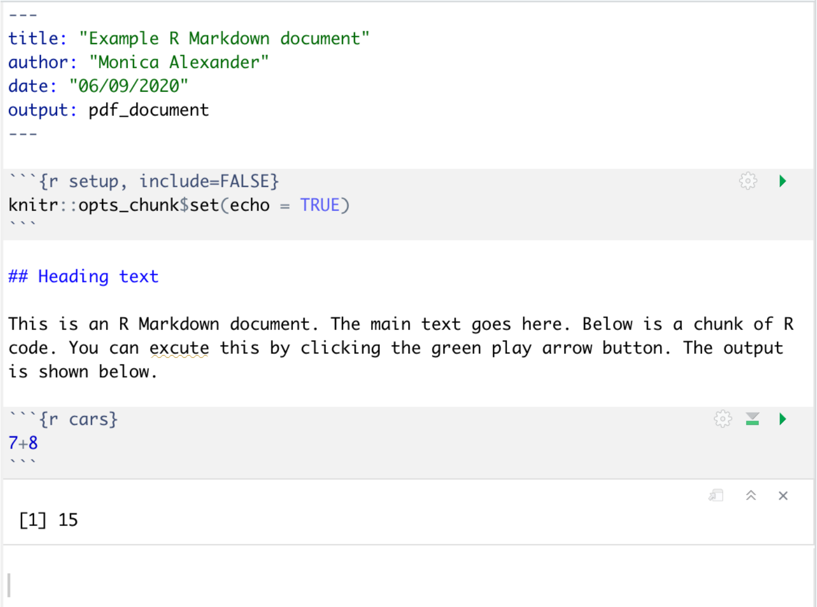
- You will need to download and install both R and RStudio
- Assuming you have these already from last semester, but more info on Quercus
- Please do this as soon as possible if you haven't already

Writing code in R

1. R Console: Executes each line of code as you go; does not save code for later use
2. R Script: Saves code and comments in a file so you can select some or all of the code in a script file to run; does not include output
3. R Markdown: A file which combines text and chunks of R code (which can be executed independently). This allows you to see output without "knitting" the whole file.

We will focus on number 3.

R Markdown documents



~~~
---
title: "Example R Markdown document"
author: "Monica Alexander"
date: "06/09/2020"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

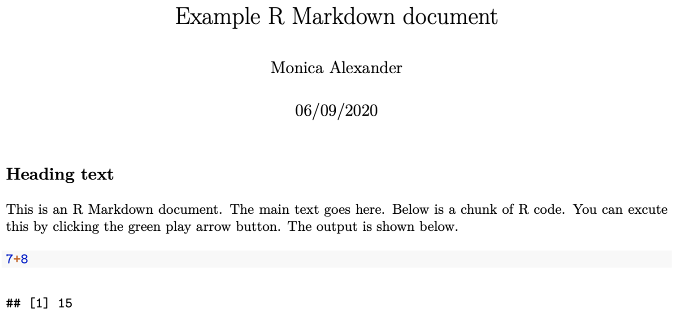
Heading text

This is an R Markdown document. The main text goes here. Below is a chunk of R
code. You can excute this by clicking the green play arrow button. The output
is shown below.

```{r cars}
7+8
```
```

```
[1] 15
```
```

# Knitted to a PDF



## Example R Markdown document

Monica Alexander

06/09/2020

### Heading text

This is an R Markdown document. The main text goes here. Below is a chunk of R code. You can excute this by clicking the green play arrow button. The output is shown below.

```
7+8
```

```
## [1] 15
```

R Packages

- A lot of people have written **R packages**, which are add ons to base R that increase the functionality

Think of a phone analogy:

- R/RStudio is a phone
- R packages are apps

We will be using a few different R packages quite a lot during the course, e.g.

- `dplyr` (data manipulation)
- `ggplot2` (graphing)

These and other packages can be downloaded through downloading the `tidyverse` package (you will do this in the lab)

Reading in and manipulating real data

```r
library(tidyverse)
gss <- read_csv(file = "data/gss.csv")
```

- `read_csv` is a **function** from the `tidyverse` package
- We are assigning the contents of the file to an object called `gss`
- The `gss` object is a data frame or **tibble** that contains all the GSS data
- We can now use other functions to manipulate and analyze the GSS data in R

Manipulating data with the tidyverse

- In this class we will be using the tidyverse “grammar” of R coding
- Tidyverse styling coding centers on a set of functions that allow you to do stuff to your dataset
- These actions are threaded together in a “sentence” using a function called the “pipe” (which is like saying “and then”)

Selecting a column with `select()`

```
select(gss, age)
```

```
## # A tibble: 20,602 x 1
##      age
##    <dbl>
##  1  52.7
##  2  51.1
##  3  63.6
##  4  80
##  5  28
##  6  63
##  7  58.8
##  8  80
##  9  63.8
## 10  25.2
## # ... with 20,592 more rows
```

The pipe %>%

```
gss %>%
  select(age)
```

```
## # A tibble: 20,602 x 1
##      age
##    <dbl>
##  1  52.7
##  2  51.1
##  3  63.6
##  4  80
##  5  28
##  6  63
##  7  58.8
##  8  80
##  9  63.8
## 10  25.2
## # ... with 20,592 more rows
```

- Read as “and then”
- So above we are taking the gss data **and then** selecting the age column

More than one pipe / Important functions

Arrange

```
gss %>%
  select(age) %>%
  arrange(age)
```

```
## # A tibble: 20,602 x 1
##      age
##    <dbl>
##  1  15
##  2  15
##  3  15
##  4  15
##  5  15
##  6  15
##  7  15
##  8  15.1
##  9  15.1
## 10  15.1
## # ... with 20,592 more rows
```

Important functions for manipulating data

- The pipe %>%
- `select` (columns)
- `filter` (rows)
- `arrange`
- `mutate`
- `summarize`

->

-> -> -> -> -> ->

Why you should learn R

- Free, open, reproducible, large community
- Used in industry
- Relatively easier to implement powerful stat methods
- Once you get the hang of it, can use to make
  - slides
  - websites (e.g. mine)
  - interactive applications (e.g. here)

Where to get help

- Intro to R:
  - R4DS is the most relevant textbook for learning "tidyverse" R
  - Telling stories with data: https://www.tellingstorieswithdata.com/01-03-r_essentials.html
- Lab time and office hours
- Google, google, google
  - Don't expect you will be able to code for memory to start off with
  - Think about what you want to do and then if you don't know how to do it, Google key terms
  - Googling errors is also very helpful
  - I mostly learned R from the formidable teacher, Stack Overflow