2018.01.09

Statistics, Inference, Sampling

Objectives

- Become familiar with the terminology of experiments
- Differentiate correlation and causation
- Define Parameters and statistics and their relationship to inference
- Understand Sampling and Sampling Distributions

Experiments

Why would someone perform a study or an experiment?

- We have questions and want answers!
- Experimentation is fundamental to learning (eg. fire hot)

See slides, complete the two examples. Handouts should be printed.

We have some terminology drop out of this exercise.

Experimental Units	The material that is assigned treatment.
Sample Size	Number of experimental units
Treatment	The variables the researcher changes. Also called independant variable, exposure, explanatory variable, predictor variable, etc.
Outcome	The variables the researchers measures. Also called response, dependant variable

Statistics, Inference, and Sampling

McCaig Statistics Group
Bryce A. Besler

Experiments

The following examples are fictionalized for the purpose of differentiating randomized experiments from observational studies.

Example 0.1. The Graduate Student Association (GSA) offers workshops on scholarship writing for students. They are interested in how effective these workshops are at helping students secure funding. In the last 5 years, 2198 students applied to national funding (NSERC, CIHR, SSHRC). Of those students, 916 attended a workshop on scholarship writing. The GSA is able to attain the amount of money each student was awarded in national funding. They discover that students who attend the scholarhips workshop secure an average of $5,680 more than their peers who do not attend the workshop. They conclude that workshops are extremely effective at helping students secure funding.

Let's identify the information in Example 0.1.

Sample Size	2198
Researcher(s)	GSA
Method	Review history of attending workshops and amount of money awarded in national competitions.
Results	Students who attend workshops attain $ 5,680 more than their peers on average.
Independant Variable	Attending a workshop
Dependant Variable	Scholarship funding
Experimental Units	Graduate Students

Example 0.2. The MTC is interested in increase the number of scholarships McCaig trainees are awarded. The MTC randomly funds 41 trainees to attend a workshop on scholarship writing. The remaining 40 trainee do not attend the workshop. All trainees apply for national funding and report their award value to the MTC. The MTC finds that there is no difference in the amount of funding secured by trainees who attended the workshop compared to those who do attend. The MTC find that scholarship workshops have no effect on funding success and stops financially supporting the workshop.

Let's identify the information in Example 0.2.

Sample Size	81
Researcher(s)	MTC
Method	Randomly send some students to the workshop and not others
Results	Attending the workshop does not increase the amount of money awarded
Independant Variable	Attending a workshop
Dependant Variable	Scholarship funding
Experimental Units	McCaig Trainees

We have the following terminology for experiments:

Experimental Unit	The material that is assigned treatment.
Sample Size	Number of experimental units.
Treatment	The variables the researcher changes. Also called independant variable, exposure, explanatory variable, predictor variable, etc.
Outcome	The variables the researcher measures. Also called response, dependant variable.

Correlation and Causation

What do you think? Do workshops increase the amount of money students are awarded?

$$\boxed{\text{Workshop}} \xrightarrow{?} \boxed{\text{Funding}}$$

What are the differences between these two studies?

Sample Size	The GSA had many more students
Assigning Treatment	The MTC assigned trainees to attend or not attend workshops.
Population	McCaig trainees are a subset of all graduate students

The GSA example is an observational study

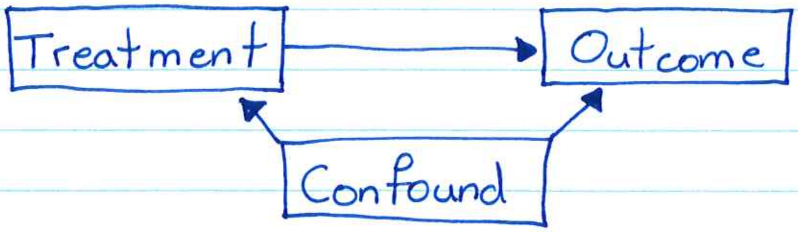
→ Researchers (GSA) could not assign treatment (workshop) to experimental units (graduate students).

The MTC example is an randomized experiment

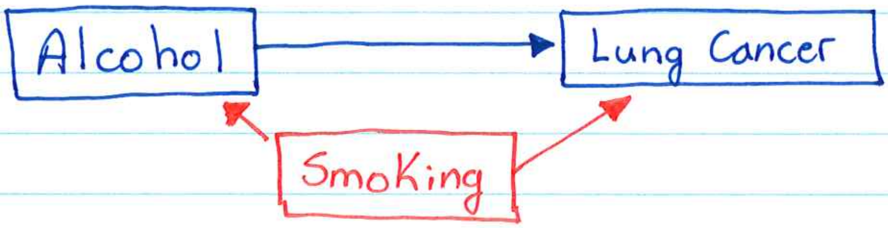
→ Researchers (MTC) ~~could~~ assigned treatment (workshop) randomly to experimental units (g McCaig Trainees).

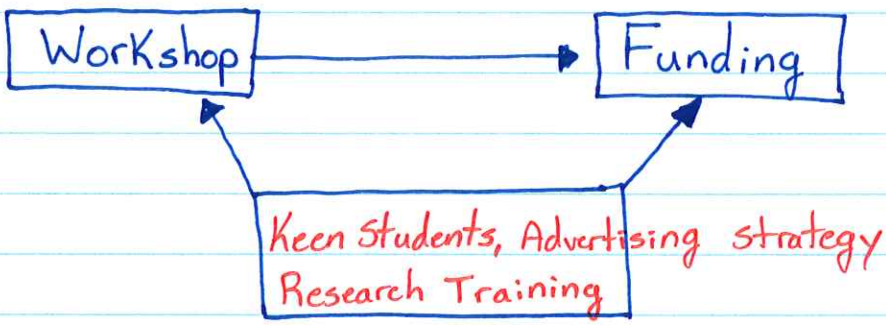
In general, causation can only be established from randomized experiments (not observational studies). This is because randomization leads to a mixing of confounding variables between treatment groups.

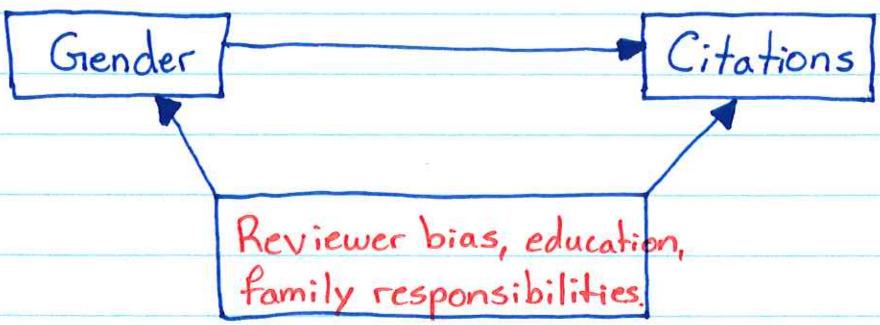
A confounding variable is defined as a variable related both to the treatment and the outcome.



Does treatment cause outcome?
Or, does confound cause treatment and outcome?







Observational studies — Causation cannot be ~~rule~~ established because confounding variables cannot be ruled out.

Randomiz~~ed~~ Experiment — Statistical methods account for chance through uncertanty.

See spurious correlations.

Parameters, Statistics, and Inference

We have discussed experiments. Now, we want to answer our question (hypothesis) by using data from the experiment. That is, we seek a statistical inference.

Statistical inference is defined as

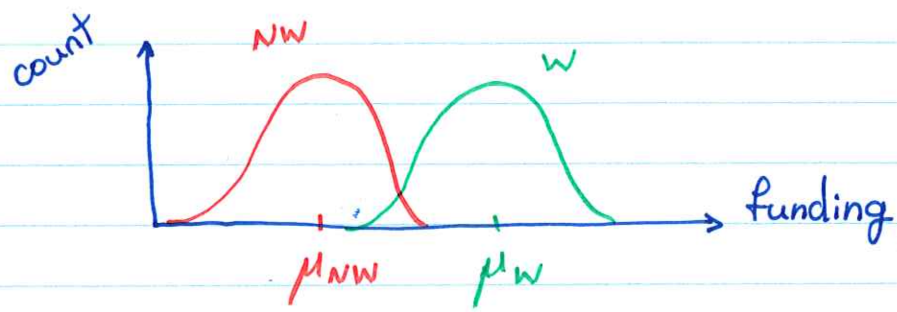
1) A conclusion that patterns in the data are present in some broader context (inference)
2) the conclusion is justified by a probability model. (statistical)

Workshop Example

$W \equiv$ Attend Workshop

$NW \equiv$ No Workshop

$\mu \equiv$ Ave. Funding



parameter μ_W is the average funding of everyone who attends the workshop. We want to estimate this so we draw N samples $x_1, x_2, \ldots, x_N$ and compute the average

statistic

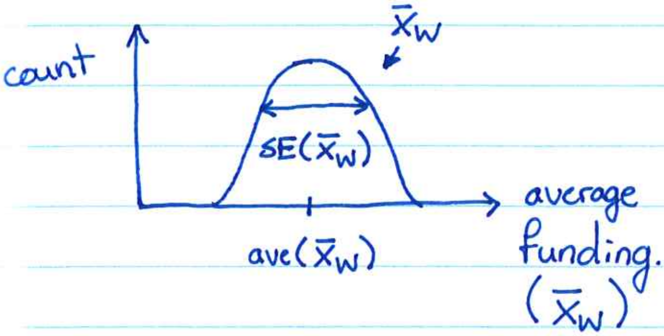
$$\bar{x}_W = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$\bar{x}_W$ is an estimate of μ_W

Is μ_W random? No!

Is x_i random? Yes!

Is $\bar{x}_W$ random? Yes!



If we repeated the experiment multiple times with a random sampling, we would get different estimates $\bar{x}_W$!

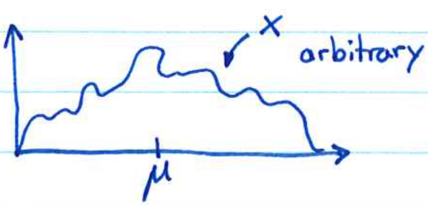
That means $\bar{x}_W$ is
1) Random (random variable)
2) Has a distribution (sampling distribution of $\bar{x}_W$)

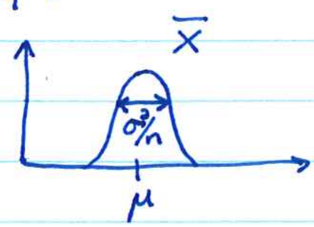
This implies we can measure
1) Absolute $\bar{x}_W$
2) Precision $SE(\bar{x}_W)$ Standard Error

~~Furethermore, there is a proof (central limit theorem) that an the average of a sample taken from a finite variance population distribution is dist~~

Furthermore, there is a theorem (central limit theorem) that the average of a sample taken from an arbitrary population with mean μ and finite variance σ^2 is distributed normally $N(\mu, \sigma^2/\sqrt{n})$!

If $x_i \sim D_{\mu, \sigma^2}$
and $\bar{x} = \frac{1}{N} \sum_{i=1}^{n} x_i$
then $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

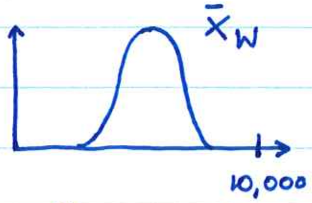




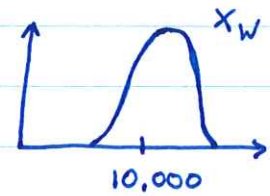
Let's say we are interested in the question:

Does $\mu_W = \$10,000$? (Null hypothesis)

We execute the experiment and get two distributions for $\bar{x}$.



Does $\mu_W = 10,000$?



Does $\mu_W = 10,000$?

Which case is more likely?

Every statistical test is based on the ideas of sampling and sampling distribution.

For all presenters, I want to see the following:

1) What is the model? ex. $\mu_W = 10,000$
2) What are the parameters? ex. μ_W
3) What are the statistics? ex. $\bar{x}_W$
4) How is precision measured? ex. $SE^2(\bar{x}_W) = \sigma^2/n$

Summary

1) What is the difference between an observational study and a randomized experiment?

2) What is a confounding variable?

3) What is the difference between a parameter and a statistic?

4) What is a sampling distribution?