Deisgn of Experiments & Hypothesis Testing

Felipe José Bravo Márquez

May 7, 2021

Motivation

In the first lecture we discussed the three major goals of statistics:

1. Describe
2. Decide
3. Predict

- In this lecture we will introduce the ideas behind the use of statistics to make decisions.
- In particular, decisions about whether a particular **hypothesis** is supported by the data. [Poldrack, 2019]

Null Hypothesis Statistical Testing (NHST)

- The specific type of hypothesis testing that we will discuss is known null hypothesis statistical testing (NHST).
- If you pick up almost any scientific research publication, you will see NHST being used to test hypotheses.
- Learning how to use and interpret the results from hypothesis testing is essential to understand the results from many fields of research.
- NHST is usually applied to **experimental** data.
- Thus, we need to introduce basic concepts on the design of experiments.

Experiments and Inference About Cause

- In the previous lecture we studied how to infer characteristics of a population from sample data using surveys or polls.
- A second type of inference is when we want to infer **cause-effect relationships** between two or more variables (e.g, does smoking cause cancer) from experimental data.
- Example [Watkins et al., 2010]: Children who drink more milk have bigger feet than children who drink less milk.

Experiments and Inference About Cause

- There are three possible explanations for this association:
  1. Drinking more milk causes children's feet to be bigger.

  2. Having bigger feet causes children to drink more milk.

  3. A **lurking variable** is responsible for both.

- We know that bigger children have bigger feet, and they drink more milk because they eat and drink more of everything than do smaller children.
- Hence, the right explanation is the third one, and the child's **overall size** is the lurking variable.
- A lurking variable is a variable that may or may not be apparent at the outset but, once identified, could explain the pattern between the variables.

Experiments and Inference About Cause

- Suppose you want to prove that explanation 1 is the right reason.
- Approach 1: take a bunch of children, give them milk, and wait to see if their feet grow.
- This won't prove anything, because children's feet will grow whether they drink milk or not.
- Approach 2: take a group of children, divide them randomly into two **groups**: 1) one group that will drink milk and 2) another group that will not, wait and compare the size of the feet of both groups.
- This approach is an **experiment**, and is the only way to establish cause and effect.

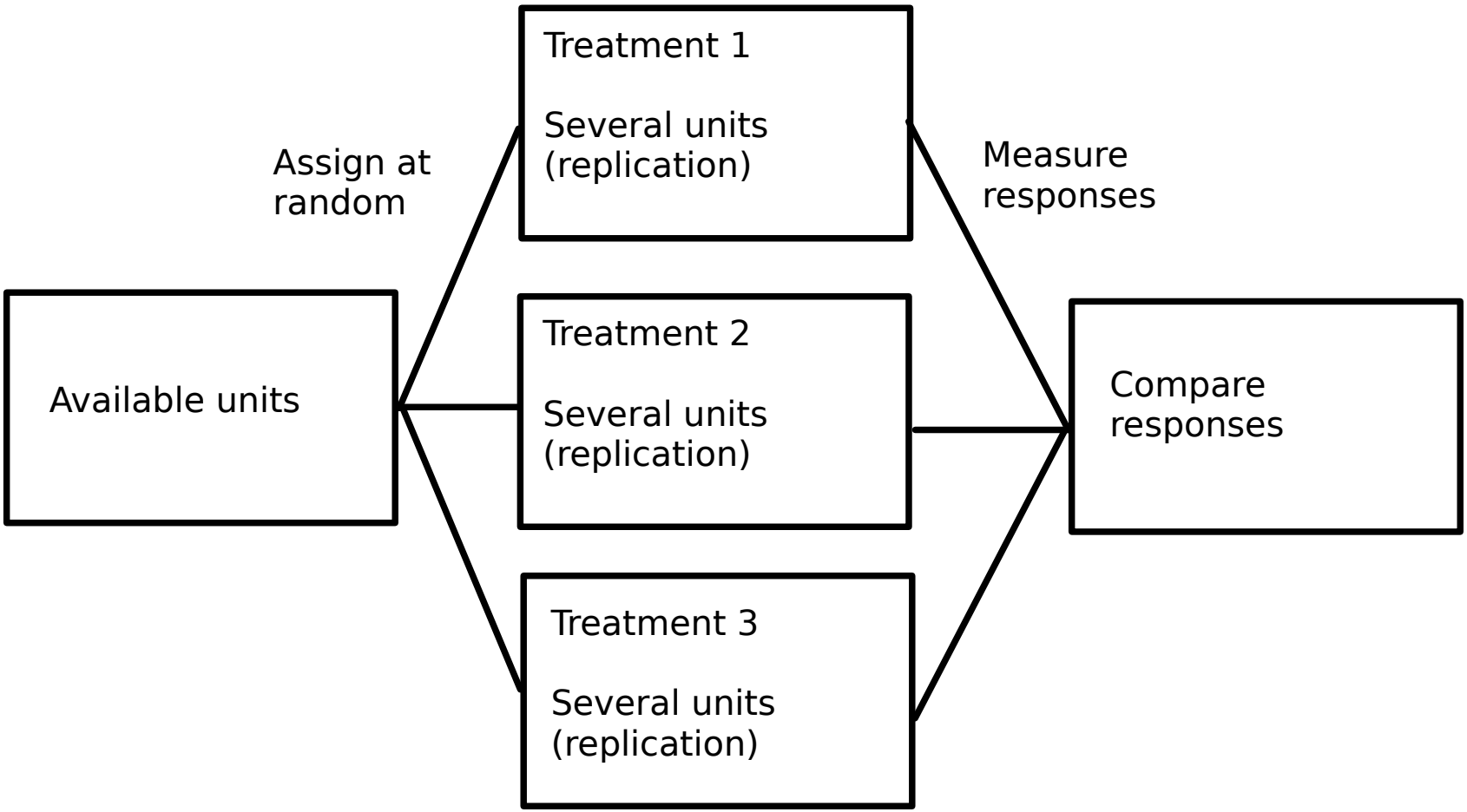
Main Concepts of Experimental Design

- **Experimental units**: the subjects on which we experiment (e.g, patients, users, laboratory animals). When the experiment units are people, we call them **subjects**.
- **Treatments**: the conditions on which we compare different unit groups. Examples: drinking milk vs. not drinking milk, smoking vs. not smoking, taking drug A vs. drug B.
- **Treatment or Experimental group**: a group of units that receives a particular treatment. Example: patients taking a new drug, software users seing a new layout.
- **Control group**: a group of units used for comparison that receives either a standard treatment or no treatment at all. Example: patients taking a placebo (a fake treatment), software users seing the standard layout.
- **Response variable**: the variable of interest used to measure the effect of the treatments on the units. Examples: weight, birth rate, click-rate, revenue, etc.

Main Concepts of Experimental Design

- **Randomization**: random assignment of treatments (including the control group) to units. This is very important since not all units are alike (e.g., people have different ages, weights, preferences). Randomization is the most reliable method of creating homogeneous treatment groups, without involving any potential biases or judgments.
- **Replication**: the repetition of an experiment on a large group of subjects. Replication reduces variability in experimental results.
- **Randomized Controlled Trial** (RCT): an experiment in which units are randomly assigned to one of several treatments and one of these groups is a control group.
- **Blind Experiment**: when the units (e.g., patients) don't know the treatment they are receiving.
- **Double-blind Experiment**: when neither the units (e.g., patients) nor the experimenters (e.g., doctors) know who is receiving a particular treatment.

Main Concepts of Experimental Design



Characteristics of a well-desgined experiment.

A/B Testing

- Data-driven companies like Amazon, Microsoft, eBay, Facebook, Google and Netflix often conduct experiments to make decisions [Kohavi et al., 2012].
- In this context, experiments are called **online controlled experiments** or **A/B tests**.
- The idea is the same, users (experimental units) are randomly exposed to one of two variants of a webpage or APP: Control (A), or Treatment (B).
- When the number of variants (treatments) is greater we have an A/B/n test.
- The response variable is called **Overall Evaluation Criterion** (OEC), which is a quantitative measure of the experiment's objective.
- OECs can be revenue, clickthrough-rate, user session duration, etc...

A/B Testing

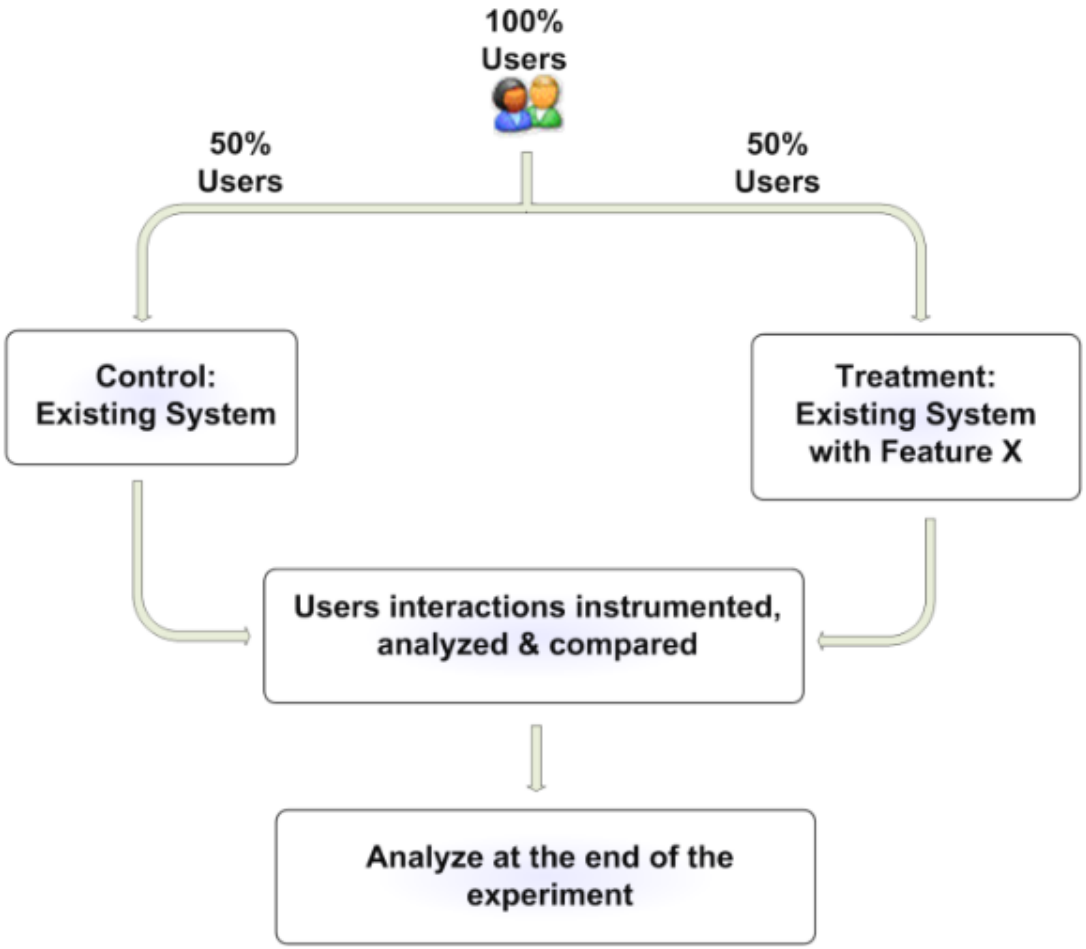


Image souce: [Kohavi et al., 2012]

Example: MSN Real Estate

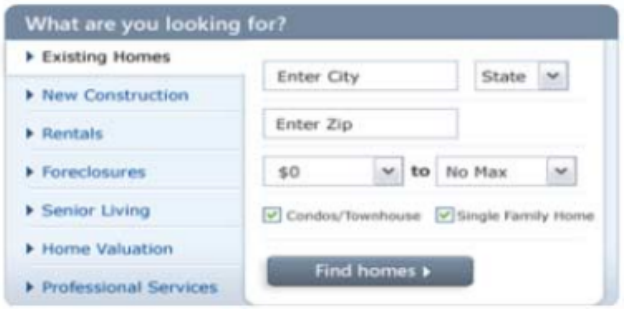
- The team running the MSN Real Estate site wanted to test different designs for the “Find a home” widget [Kohavi et al., 2009].
- Visitors who click on this widget are sent to partner sites, and Microsoft receives a referral fee.
- Six different designs of this widget, including the incumbent (control), were proposed.
- Users were randomly splited between the variants in a persistent manner (a user receives the same experience in multiple visits) during the experiment period.

Example: MSN Real Estate

Control

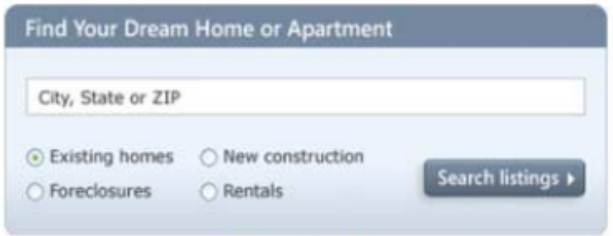
Treatment1

Treatment 2



Treatment 3

Treatment 4



Treatment 5

Example: MSN Real Estate

- Their interactions are instrumented and key metrics computed.
- In this experiment, the Overall Evaluation Criterion (OEC) was simple: average revenue per user.
- The winner, Treatment 5, increased revenues by almost 10% (due to increased clickthrough).
- The Return-On-Investment (ROI) for MSN Real Estate was phenomenal, as this is their main source of revenue, which increased significantly through a simple change.

Observational Studies and Confounding

- Sometimes we can't randomly assign units to the different treatments.
- For example, it would be unethical to design a randomized controlled trial deliberately exposing people to a potentially harmful situation.
- In an **obervational study** the conditions of interest are aready built into the units being studied.
- Observational studies are almost always worse than controlled experiments for determining cause-effect relationships.
- But sometimes is the only thing we can do.
- A phenomenum called **confounding** is the major treat to observational studies.
- Two possible influences on an observed outcome are **confounded** if they are mixed together in a way that makes it impossible to separate their effects on the responses [Watkins et al., 2010].

Example of Confounded Observational Study

- The thymus, a gland in your neck, behaves in a peculiar way.
- Unlike other organs of the body, it doesn't get larger as you grow—it actually gets smaller.
- Ignorance of this fact led early 20th-century surgeons to adopt a worthless and dangerous surgical procedure.

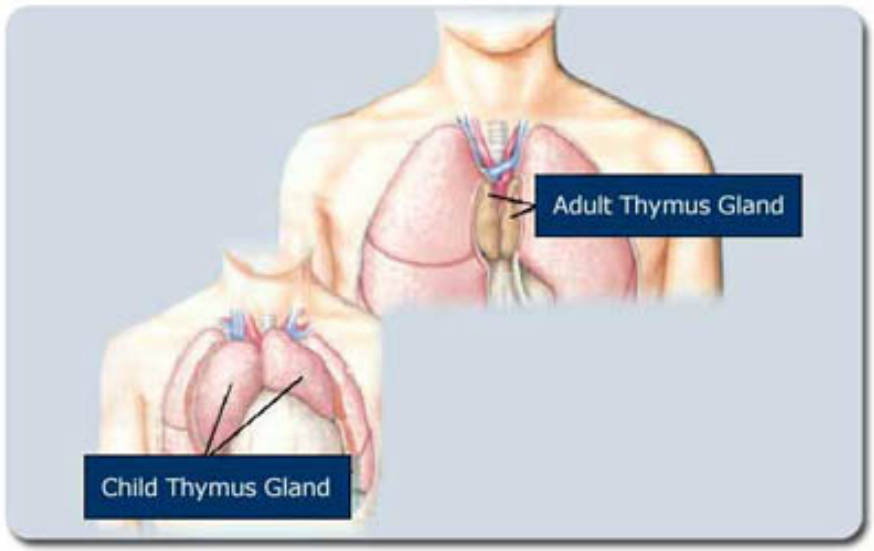


Figure: source: `http://esvc001414.wic005tu.server-web.com/tech_imm_bio_principle.htm`

Example of Confounded Observational Study

- Many infants were dying of what seemed to be respiratory obstructions.
- Doctors did autopsies on infants who died with respiratory symptoms and compared againts aotopsies made on adults who died of various causes.
- Most autopsies to infants show big thymus glands compared to adults.
- Doctors concluded that the respiratory problems were caused by an enlarged thymus.
- In 1912, Dr. Charles Mayo published an article recommending removal of the thymus to treat respiratory problems in children.
- This recommendation was made even though a third of the children who were operated on died.
- The doctors could not tell whether children with a large thymus tended to have more respiratory problems because they had no evidence about children with a smaller thymus.

Example of Confounded Observational Study

- Age and size of thymus were confounded.
- The thymus study is an example of an observational study, not an experiment.

	Age	
	Child	Adult
Thymus size		
Large	Problems	No evidence
Small	No evidence	No problems

- If Dr. Mayo had used a randomized experiment to evaluate surgical removal of the thymus, he would have seen that the treatment was not effective and many lives might have been spared.
- However, at the time, randomized experiments were not often used in the medical profession.
- These days, any new medical treatment (e.g., a COVID vaccine) must prove its value in an RCT.

Another Example of Confounding

- Suppose we want to compare student performance on standardized tests (e.g., SIMCE, PSU) in public and private schools.
- We know that the socioeconomic distribution of students is different in public and private schools.
- We also suspect that socioeconomic background may influence student performance on these tests.
- The type of school (public or private) and the socioeconomic background are confounded.

Randomized Paired Comparison (Matched Pairs)

- Randomized Paired Comparison or Matched Pairs is an approach to design experiments **controlling** for confounding variables.
- We sort the experimental units into pairs of similar units (matched pairs), where similarity is measured according to confounding variables.
- The two units in each pair should be enough alike that you expect them to have a similar response to any treatment.
- Randomly decide which unit in each pair is assigned which treatment.
- We are essentially building comparable Control and Treatment populations by segmenting the users by common confounds, similarly to stratified sampling.

Matched Pairs Example

- Suppose we want to study the relation between hypertension and end-stage renal disease (ESRD) [De Graaf et al., 2011].
- Obesity is a potential confounder as obesity is associated with both hypertension and ESRD.
- Matching approach: we ensure that the average body mass index (BMI) is the same in the group of patients exposed to hypertension and another group of patients unexposed to hypertension.
- This could be achieved by searching an obese patient without hypertension for each obese patient with hypertension.
- Other potential confouding variables like age or sex could also be considered in the matching.

Hypothesis Testing

- When we want to test whether some assumed **property** about a population is contrasted with a statistical sample we use a **hypothesis test**.
- The test consists of the following hypotheses:
  - **Null Hypothesis** H_0: Symbolizes the current situation. What has been considered real up to the present.
  - **Alternative Hypothesis** H_a: it is the alternative model that we want to consider.
- The idea is to find enough **statistical evidence** to reject H_0 and be able to conclude H_a.
- If we do not get enough statistical evidence **we fail to reject** H_0

Hypothesis Testing

Methodology to Perform a Hypothesis Test

- Choose a null hypothesis H_0 and alternative H_a.
- Set a test significance level α.
- Calculate a statistic T from the data.
- The statistic T is usually a standardized value that we can check in a distribution table.
- Define a rejection criterion for the null hypothesis. It is usually a critical value c.

Types of T-tests

https://en.wikipedia.org/wiki/Student%27s_t-test

- Single-sample t-test
- Unpaired two sample t-test: better using https://en.wikipedia.org/wiki/Welch%27s_t-test
- Paired two sample t-test

https://www.datanovia.com/en/lessons/types-of-t-test/#one-sample-t-test Tests can be one-sided or two-sided

Nice explanations of degrees of freedom:

https://crumplab.github.io/statistics/t-tests.html

Single-sample T-test

- Example: It is known that the average number of hours of monthly Internet use in Chile is 30 hours.
- Suppose we want to show that the average is different from that value.
- We would have that $H_0 : \mu = 30$ and $H_a : \mu \neq 30$
- Let's set $\alpha = 0.05$ and collect 100 observations.
- Suppose we get $\overline{X_n} = 28$ and $s = 10$
- One way to test is to construct a confidence interval for μ and see if H_0 is in the interval.

```
> 28-qt(p=0.975,99)*10/sqrt(100)
[1] 26.01578
> 28+qt(p=0.975,99)*10/sqrt(100)
[1] 29.98422
```

- The interval would be the acceptance zone of H_0 and anything outside of this would be my rejection region.
- Since 30 is in the rejection region, I reject my null hypothesis with 5% confidence.

Univariate T-test

- Another way to perform the test is to compute the statistic $T = \frac{\overline{X_n} - \mu_o}{\frac{s}{\sqrt{n}}}$
- In this case it would be

$$T = \frac{28 - 30}{\frac{10}{\sqrt{100}}} = -2$$

- Since $H_a : \mu \neq 30$, we have a two-sided test, where the acceptance region is.

$$t_{n-1,1-\alpha/2} < T < t_{n-1,\alpha/2}$$

```
 > qt(0.025,99)
[1] -1.984217
> qt(0.975,99)
[1] 1.984217
```

- Since T is in the rejection region, we reject the null hypothesis.

P-value

- Generally, in addition to knowing whether we reject or fail to reject a null hypothesis we want to quantify the evidence we have against it.
- A **p-value** is defined as the probability of obtaining an outcome at least as extreme as that observed in the data given that the null hypothesis is true.
- "Extreme" means far from the null hypothesis and favorable for the alternative hypothesis.
- If the **p-value** s less than the significance level α, we reject H_0
- Example:

```
> data(iris)
> mu<-3 # null hypothesis
> alpha<-0.05
> n<-length(iris$Petal.Length)
> xbar<-mean(iris$Petal.Length)
> s<-sd(iris$Petal.Length)
> se<-s/sqrt(n)
> t<-(xbar-mu)/(s/sqrt(n))
> pvalue<-2*pt(-abs(t),df=n-1)
> pvalue
[1] 4.94568e-07 # is less than 0.05 then we reject H0
```

Univariate T-test

- The elegant way to do it in R:

```
> t.test(x=iris$Petal.Length,mu=3)

One Sample t-test

data:  iris$Petal.Length
t = 5.2589, df = 149, p-value = 4.946e-07
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 3.473185 4.042815
sample estimates:
mean of x
    3.758
```

Errors

- We have two types of errors when we perform a hypothesis test
- Type I error: it is when we reject the null hypothesis when it is true.
- This error is equivalent to the significance level α.
- Type II error: is when the null hypothesis is false but we do not have statistical evidence to reject it.
- To mitigate type I errors we generally use smaller values of α.
- To mitigate type II errors we generally work with larger samples.
- There is a trade-off between type I and type II errors.

	Retain H_0	Reject H_0
H_0 true	✓	type I
H_1 true	type II error	✓

Statistical Power

Critics to Hypothesis Testing

FOUR CARDINAL RULES OF STATISTICS by Daniela Witten

- ONE: CORRELATION DOES NOT IMPLY CAUSATION. Yes, I know you know this, but it's so easy to forget! Yeah, YOU OVER THERE, you with the p-value of 0.0000001 — yes, YOU!! That's not causation.
- No matter how small the p-value for a regression of IQ onto shoe size is, that doesn't mean that big feet cause smarts!! It just means that grown-ups tend to have bigger feet and higher IQs than kids.
- So, unless you can design your study to uncover causation (very hard to do in most practical settings — the field of causal inference is devoted to understanding the settings in which it is possible), the best you can do is to discover correlations. Sad but true.
- TWO: A P-VALUE IS JUST A TEST OF SAMPLE SIZE. Read that again — I mean what I said! If your null hypothesis doesn't hold (and null hypotheses never hold IRL) then the larger your sample size, the smaller your p-value will tend to be.
- If you're testing whether mean=0 and actually the truth is that mean=0.000000001, and if you have a large enough sample size, then YOU WILL GET A TINY P-VALUE.
- Why does this matter? In many contemporary settings (think: the internet), sample sizes are so huge that we can get TINY p-values even when the deviation from the null hypothesis is negligible. In other words, we can have STATISTICAL significance w/o PRACTICAL significance

FOUR CARDINAL RULES OF STATISTICS by Daniela Witten

- Often, people focus on that tiny p-value, and the fact that the effect is of **literally no practical relevance** is totally lost.
- This also means that with a large enough sample size we can reject basically ANY null hypothesis (since the null hypothesis never exactly holds IRL, but it might be “close enough” that the violation of the null hypothesis is not important).
- Want to write a paper saying Lucky Charms consumption is correlated w/blood type? W/a large enough sample size, you can get a small p-value. (Provided there’s some super convoluted mechanism with some teeny effect size. . . which there probably is, b/c IRL null never holds)
- THREE: SEEK AND YOU SHALL FIND. If you look at your data for long enough, you will find something interesting, even if only by chance! In principle, we know that we need to perform a correction for multiple testing if we conduct a bunch of tests.
- But in practice, what if we decide what test(s) to conduct AFTER we look at data? Our p-value will be misleadingly small because we peeked at the data. Pre-specifying our analysis plan in advance keeps us honest. . . but in reality, it’s hard to do!!!
- Everyone is asking me about the mysterious and much-anticipated fourth rule of statistics. The answer is simple: we haven’t figured it out yet.... that’s the reason we need to do research in statistics

References I


De Graaf, M. A., Jager, K. J., Zoccali, C., and Dekker, F. W. (2011).
Matching, an appealing method to avoid confounding?
Nephron Clinical Practice, 118(4):c315–c318.


Kohavi, R., Crook, T., Longbotham, R., Frasca, B., Henne, R., Ferres, J. L., and Melamed, T. (2009).
Online experimentation at microsoft.
Data Mining Case Studies, 11(2009):39.


Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., and Xu, Y. (2012).
Trustworthy online controlled experiments: Five puzzling outcomes explained.
In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 786–794.


Poldrack, R. A. (2019).
Statistical thinking for the 21st century.
https://statsthinking21.org/.


Watkins, A. E., Scheaffer, R. L., and Cobb, G. W. (2010).
Statistics: from data to decision.
John Wiley & Sons.