Deisgn of Experiments & Hypothesis Testing

Felipe José Bravo Márquez

May 12, 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.

Figure: Image source: `https://www.dreamstime.com`

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.

- 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.
- 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.

Experiments and Inference About Cause

- The right explanation is the third one: the child's **overall size** is the lurking variable.
- However, suppose we want to prove that explanation 1 is the right reason with the following approaches.
- 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.

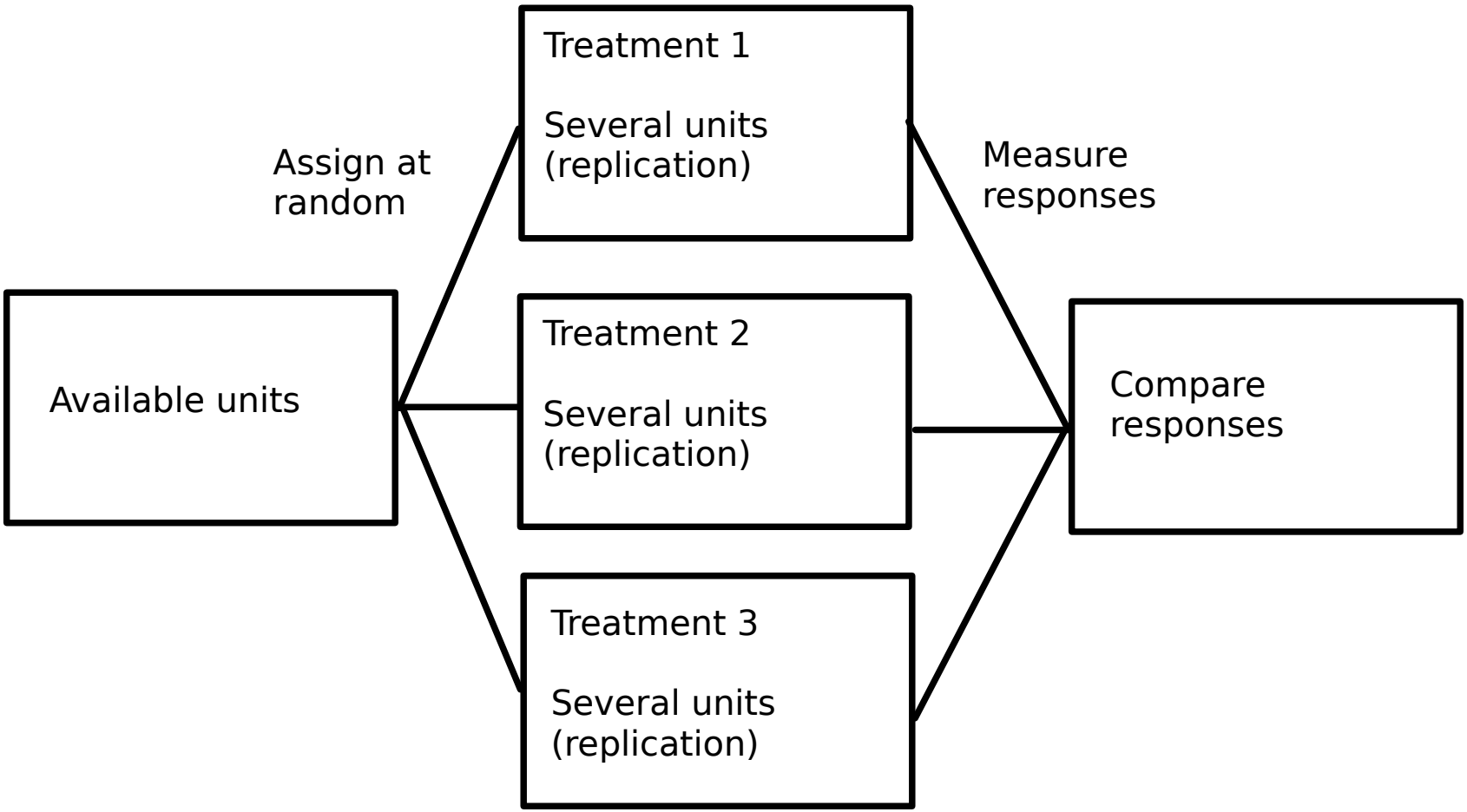
Main Concepts of Experimental Design

- **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 receiving a particular treatment. Example: patients taking a new drug, software users seeing a new layout.
- **Control group**: a group of units used for comparison receiving either a standard treatment or no treatment at all. Example: patients taking a placebo (a fake treatment), software users seeing the standard layout.
- **Response variable**: the variable of interest used to measure the effect of the treatments on the units. Examples: weight, birth rate, antibody levels, 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-designed experiment.

A/B Testing

- Data-driven companies like Amazon, Microsoft, eBay, Facebook, Google and Netflix constantly 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 the software: Control (A), or Treatment (B).
- An when there is more than one treatment 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

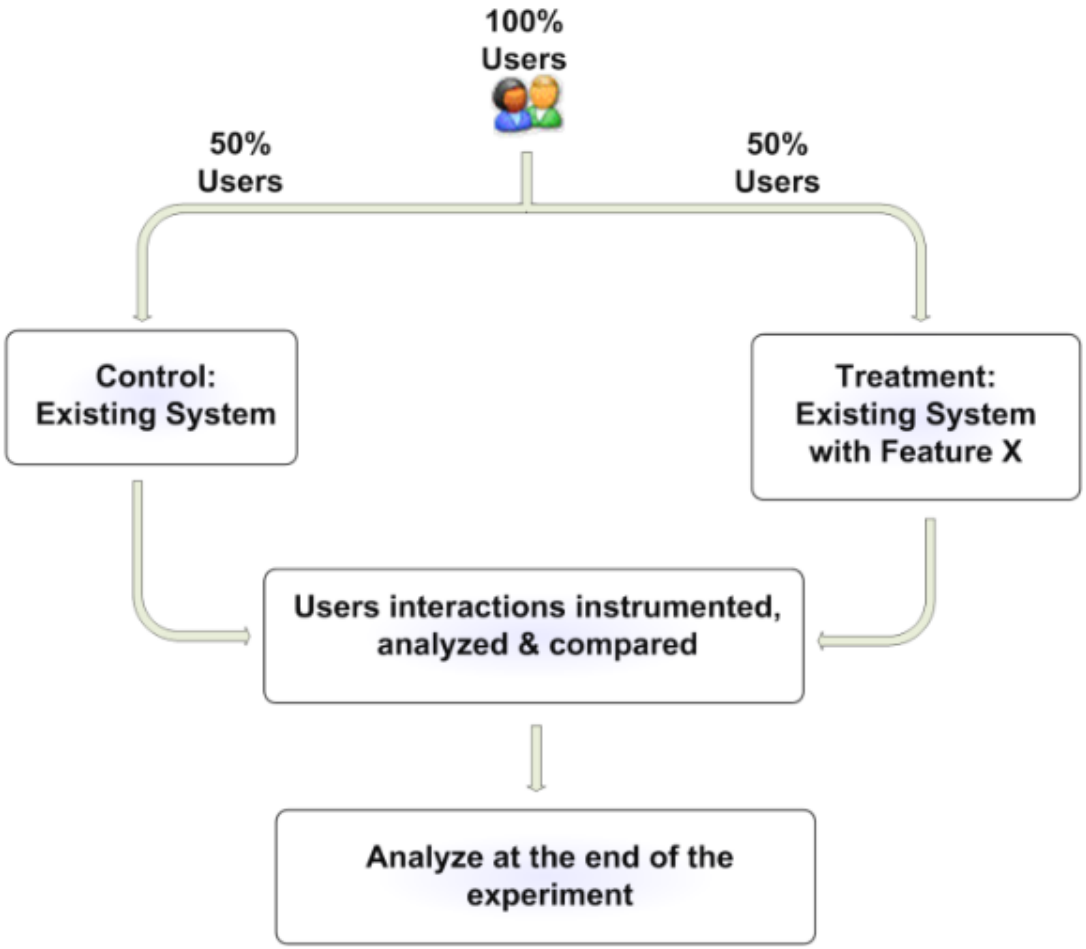


Image source: [Kohavi et al., 2012]

Example: MSN Real Estate

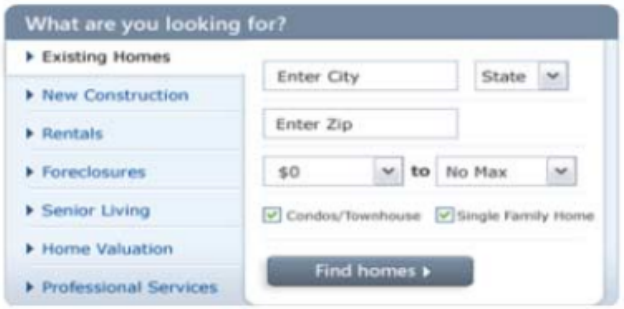
- The team running the MSN Real Estate site wanted to test different designs for the “Find a home” widget [Kohavi et al., 2009].
- Visitors who click on this widget are sent to partner sites, and Microsoft receives a referral fee.
- Six different designs of this widget, including the incumbent (control), 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

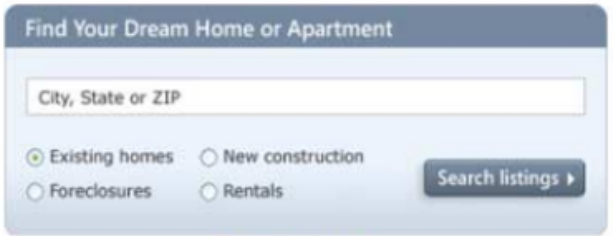
Treatment1

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 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 **observational study** the conditions of interest are already 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 phenomenon 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.

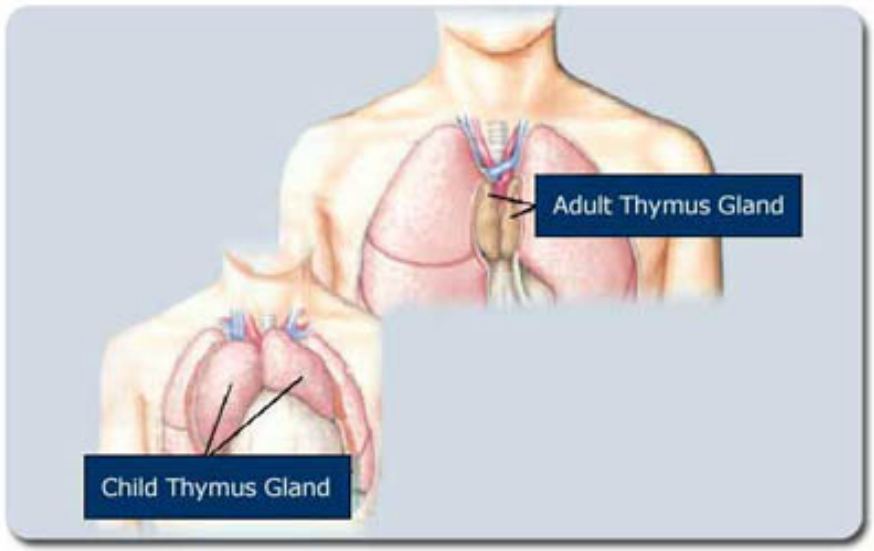


Figure: source: `http://esvc001414.wic005tu.server-web.com/tech_imm_bio_principle.htm`

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 against autopsies 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 a standardized tests (e.g., SIMCE, PSU) between 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

- Now that we understand what experimental data looks like we are in place to introduce Null Hypothesis Statistical Testing (NHST).
- A **hypothesis test** allows us to to measure whether some assumed **property** about a population is contrasted with a statistical sample.
- In the context of experiments, NHST helps us to determine weather observed differences between treatment and control groups are unlikely to have occurred by chance.
- Hypothesis testing can be applied to all kinds of population parameters (e.g., mean, variance, median).
- In the class we will focus on testing the **population mean** μ.

Hypothesis Testing

- We will study the following types of parametric tests to the mean:
  1. **One sample tests**: we contrast the sample mean to a pre-specified value.
  2. **Unpaired two sample test**: we compare the sample means of two independent groups (control vs. treatment).
  3. **Paired two sample test**: here we compare the means of two dependent groups where we have two values for the same samples. For example: in matched pairs experiments.
- All these tests can be one-sided or two-sided.
- In the same way as for confidence intervals we will use Normal and T-student distributions for modeling the sampling distribution of sample means.
- Warning: there are many counterintuitive concepts around NHST (e.g., null hypothesis, p-values).
- Thus, we will fist introduce these concepts with two examples taken from [Poldrack, 2019] and [Marchini, 2008].
- Then we will formalize them in more detail.

Example 1: Body-worn Cameras

- Body-worn cameras are thought to reduce the use of force and improve behavior of police officers.
- An RCT of the effectiveness of body-worn cameras was performed by the Washington, DC government and DC Metropolitan Police Department in 2015/2016.
- Officers were randomly assigned to wear a body-worn camera or not.
- Their behavior was then tracked over time to determine whether the cameras resulted in less use of force and fewer civilian complaints about officer behavior.

Figure: source: `https://www.nytimes.com`

Example 1: Body-worn Cameras

- Let's say we want to specifically test the hypothesis of whether the use of force is decreased by the wearing of cameras.
- The RCT provides us with the data to test the hypothesis – namely, the rates of use of force by officers assigned to either the camera or control groups.
- The next obvious step is to look at the data and determine whether they provide convincing evidence for or against this hypothesis.
- That is: What is the likelihood that body-worn cameras reduce the use of force, given the data and everything else we know?
- It turns out that this is **not** how null hypothesis testing works.

Example 1: Body-worn Cameras

- Instead, we first take our hypothesis of interest (i.e. that body-worn cameras reduce use of force), and flip it on its head, creating a **null hypothesis**.
- In this case, the null hypothesis would be that cameras do not reduce use of force.
- Importantly, we then assume that the null hypothesis is true.
- We then look at the data, and determine how likely the data would be if the null hypothesis were true.
- If the data are sufficiently unlikely under the null hypothesis that we can reject the null in favor of the **alternative hypothesis** which is our hypothesis of interest.
- If there is not sufficient evidence to reject the null, then we say that we retain (or "fail to reject") the null.
- Then we stick with our initial assumption that the null is true.

Example 2: Babies

- From previous experience we know that the birth weights of babies in England have a mean of 3000g and a standard deviation of 500g.
- We think that maybe babies in Australia have a mean birth weight greater than 3000g and we would like to test this hypothesis.
- We take a sample of babies from Australia, measure their birth weights and see if the sample mean is significantly larger than 3000g.
- The main hypothesis that we are most interested in is the **research hypothesis**, denoted H_1 , that the mean birth weight of Australian babies is greater than 3000g.

Example 2: Babies

- The other hypothesis is the null hypothesis, denoted H_0, that the mean birth weight is equal to 3000g.
- We can write this compactly as:

$$H_0: \mu = 3000g$$
$$H_1: \mu > 3000g$$

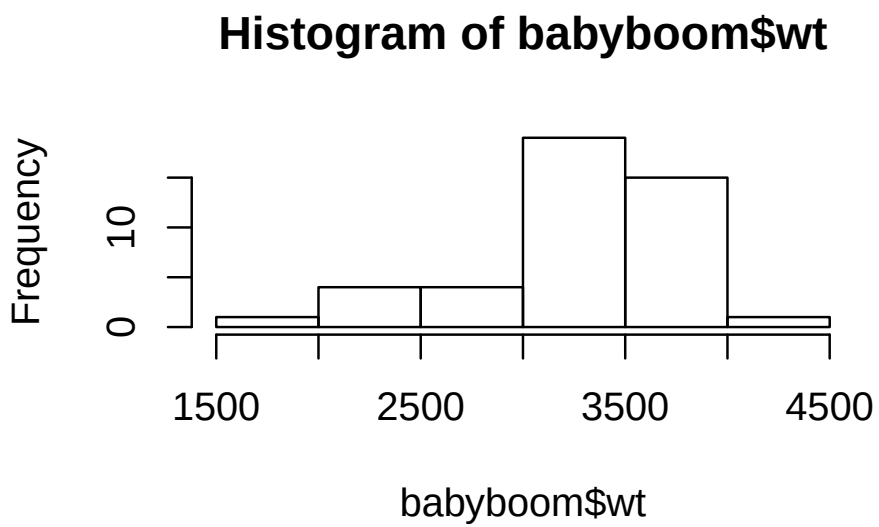
- The null hypothesis is written first followed by the research hypothesis.
- The research hypothesis is often called the **alternative hypothesis** even though it is often the first hypothesis we think of.

Example 2: Babies

- Normally, we start with the research hypothesis and “set up” the null hypothesis to be directly counter to what we hope to show.
- We then try to show that, in the light of our collected data, that the null hypothesis is false.
- We do this by calculating the probability of the data if the null hypothesis is true.
- If this probability is very small it suggests that the null hypothesis is false.
- Once we have set up our null and alternative hypothesis we can collect a sample of data.
- For example, we can imagine we collected the birth weights of the 44 babies in the Babyboom dataset.

```
>library(UsingR)
>data(babyboom)
>hist(babyboom$wt)
```

Example 2: Babies



- The sample mean of the dataset is $\overline{x}$ is:

```
> xbar<-mean(babyboom$wt)
> xbar
[1] 3275.955
```

Example 2: Babies

- We now want to calculate the probability of obtaining a sample with a mean as large as 3275.955 under the assumption of the null hypothesis H_0.
- From the CLT we know that the sampling distribution of $\overline{X}$ follows as Normal distribution when n is sufficiently large: $\overline{X} \sim N(\mu, \sigma^2/n)$
- If we assume H_0 is true, then $\mu = 3000$.
- The value of n is 44 and the value of σ is known is this case and is equal to 500.
- Let's calculate the standard error $\frac{\sigma}{\sqrt{n}}$:

```
> mu0<-3000
> sd<-500
> n<-nrow(babyboom)
> se<-sd/sqrt(n)
> se
[1] 75.37784
> se^2
[1] 5681.818
```

Example 2: Babies

- Now we can calculate the probability of obtaining a sample with a mean as large as 3275.955:

```
> #pvalue
> 1-pnorm(xbar, mean =mu0, sd =se)
[1] 0.0001256405
> #or
> Z.score<-(xbar-mu0)/se
> Z.score
[1] 3.660951
> p.value<-1-pnorm(Z.score)
> p.value
[1] 0.0001256405
```

Example 2: Babies

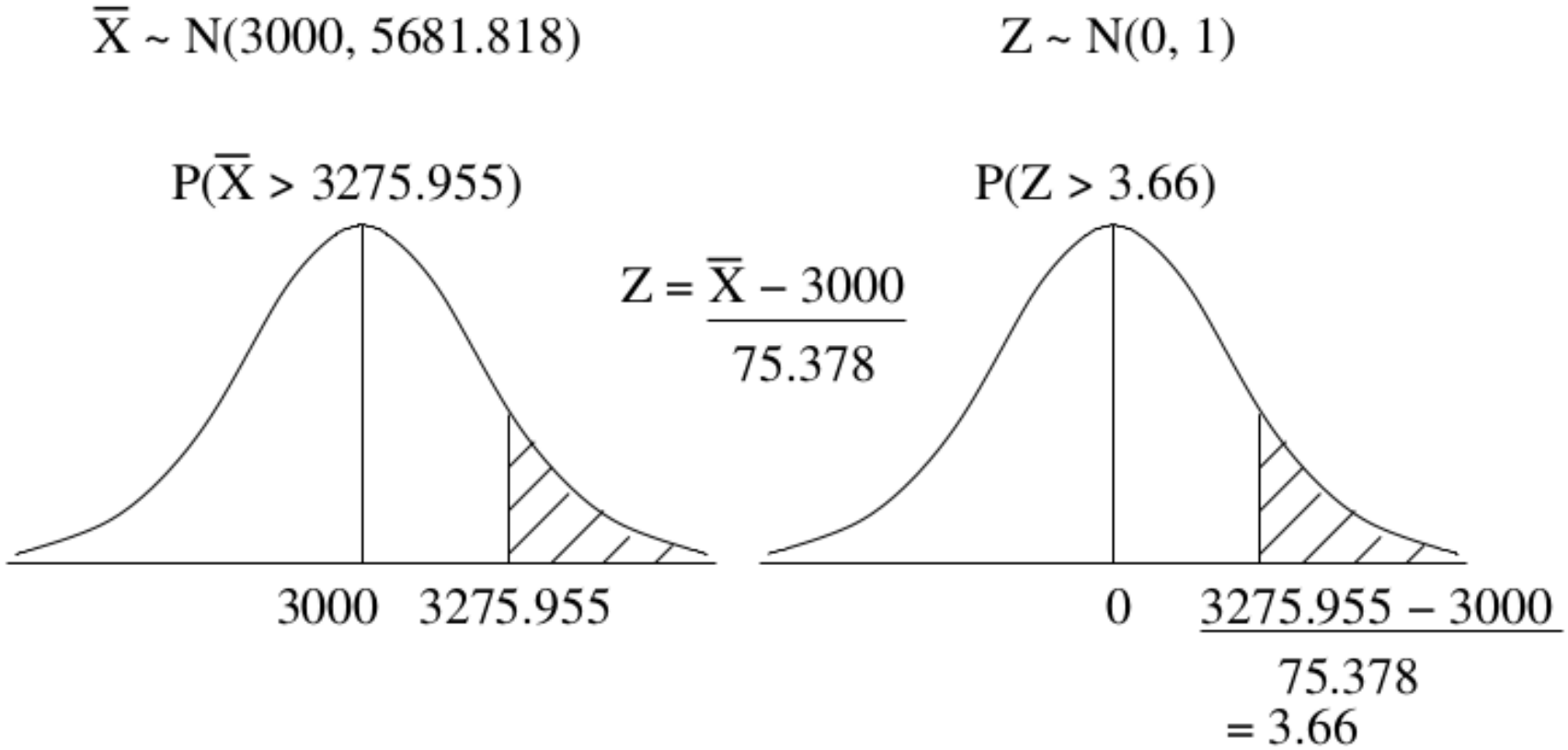


Figure: [Marchini, 2008]

Example 2: Babies

- The probability we calculate is called the **p-value** of the test.
- In this case the p-value is very low.
- This says that the probability of the data is very low if we assume the null hypothesis is true.
- But how low does this probability have to be before we can conclude that the null hypothesis is false.
- The convention within statistics is to choose a **level of significance** α before the experiment that dictates how low the p-value should be before we reject the null hypothesis.
- In practice, many people use a significance level of 5% and conclude that there is significant evidence against the null hypothesis if the p-value is less than or equal to 0.05.
- A more conservative approach uses a 1% significance level and conclude that there is significant evidence against the null hypothesis if the p-value is less than 0.01.

Example 2: Babies

- In our current example, the p-value is 0.00013 which is lower than $\alpha = 0.05$.

```
> alpha<-0.05
> p.value<=alpha
[1] TRUE
```

- In this case, we would conclude that:
  "there is signicant evidence against the null hypothesis at the 5% level".
- Another way of saying this is that:
  "we reject the null hypothesis at the 5% level"
- If the p-value for the test much larger, say 0.23, then we would conclude that:
  "the evidence against the null hypothesis is not significant at the 5% level"
- Another way of saying this is that:
  "we cannot reject the null hypothesis at the 5% level"

Example 2: Babies

- In the previous example, we assumed that σ was known.
- In many cases σ is unknown and we must estimate it using the unbiased estimator s that we saw in previous class.
- In these cases we can calculate a T statistic $T = \frac{\overline{X_n} - \mu_o}{\frac{s}{\sqrt{n}}}$

```
> s<-sd(babyboom$wt)
> s
[1] 528.0325
> se<-s/sqrt(n)
> se
[1] 79.60389
>
> T.sta<-(xbar-mu0)/se
> T.sta
[1] 3.466596
```

Example 2: Babies

- From previous class we know that follows a t-student distribution with $n-1$ degrees of freedom $T \sim t_{n-1}$.
- When we perform the test using t-student distribution we are using a T-test.
- The p-value can be calculated analougusly to the previous case now using the t-student distribution.

```
> p.value<-1-pt(T.sta,df = n-1)
> p.value
[1] 0.0006042622
```

- We also reject the null hypothesis in this case with $\alpha = 0.05$.
- But the p-value is larger than before.
- This is becasue the t-distribution has wider tails than the Normal distribution.

Example 2: Babies

- We can perform t-tests directly in R as follows:

```
> t.test(x = babyboom$wt,mu = 3000,
alternative = "greater",conf.level = 1-alpha)

One Sample t-test

data:  babyboom$wt
t = 3.4666, df = 43, p-value = 0.0006043
alternative hypothesis: true mean is greater than 3000
95 percent confidence interval:
 3142.135      Inf
sample estimates:
mean of x
 3275.955
```

Hypothesis Testing

The two hypotheses in NHST

- **Null Hypothesis** H_0: what has been considered real up to the present or what would we expect the data to look like if there is no effect.
  - The null hypothesis always involves some kind of equality ($=$).
- **Alternative Hypothesis** H_a: it is the alternative model that we want to consider or what we expect if there actually is an effect.
  - The alternative hypothesis always involves some kind of inequality ($\neq$, $>$, or $<$).

- Importantly, null hypothesis testing operates under the assumption that the null hypothesis is true unless the evidence shows otherwise.
- 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

- Specify a null hypothesis H_0 and alternative H_a.
- Set a test significance level α.
- Collect some data relevant to the hypothesis.
- Fit a model to the data and compute a test statistic T.
  - In parametric tests, T is a standardized value that we can check in a distribution table (e.g., a Z-score).
- Assess the "statistical significance" of T.

The last part can be done with two approaches

- Critical region: Calculate a region of values such that if T lies in this region then we will reject H_0
- P-value approach: compute the probability of the observed value (or more extreme values) of that statistic assuming that the null hypothesis is true and compare it with α.

Single-sample Tests

- 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

There are many other tests

- Propotion tests
- Analysis of Variance (ANOVA)
- Chi-square tests of idependence
- Kolmogorov–Smirnov test

Conclusions

- Bla bla bla

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.


Marchini, J. (2008).
Introduction to probability and statistics.
https://jmarchini.org/teaching/#introduction-to-probability-and-statistics.


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

References II

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