

# Design of Experiments & Hypothesis Testing

Felipe José Bravo Márquez

August 31, 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 **confounding variable** is responsible for both.



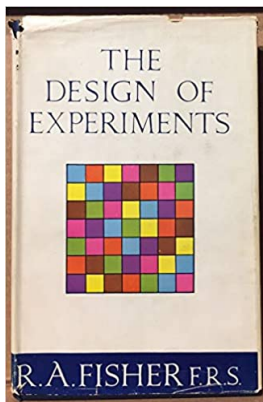
- A confounding variable is a variable that may or may not be apparent at the outset but, once identified, could explain the pattern between the variables.
- We know that bigger children have bigger feet, and they drink more milk because they eat and drink more of everything than do smaller children.

# Experiments and Inference About Cause

- The right explanation is the third one: the child's **overall size** is the confounding variable.
- However, suppose we 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 standard practice in statistics to establish cause and effect relationships.

# The Design of Experiments

- The main ideas of the design of experiments were proposed in 1936 in the book "Design of Experiments" by the English statistician Ronald Fisher [Fisher, 1936].



# Main Concepts of Experimental Design

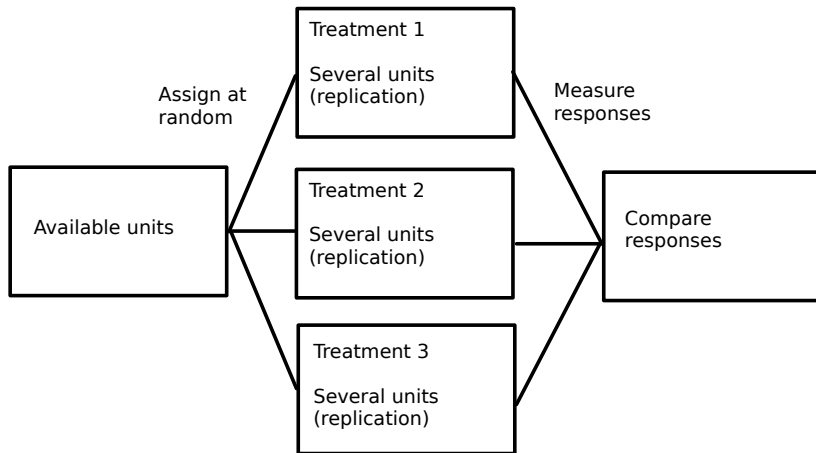
- **Experimental units:** the subjects on which we experiment (e.g, patients, users, laboratory animals). When the experiment units are people, we call them **subjects**.
- **Treatments:** the conditions on which we compare different unit groups. Examples: drinking milk vs. not drinking milk, smoking 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).
- 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

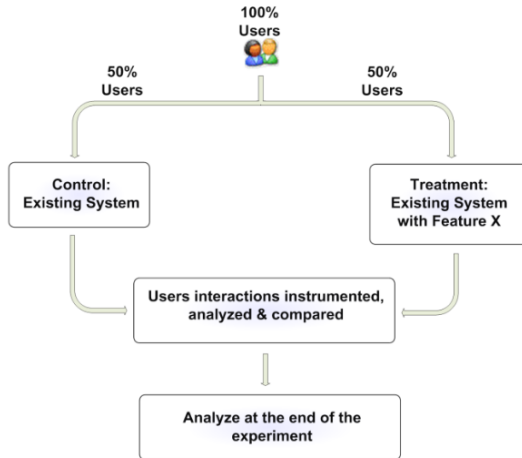


Image source: [Kohavi et al., 2012]

## 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 split between the variants in a persistent manner (a user receives the same experience in multiple visits) during the experiment period.

# Example: MSN Real Estate

Find a new home or apartment

☒ Existing Homes  
from REALTOR.com®
 ☐ New Homes  
from Move.com™
 ☐ Foreclosures  
from RealtyTrac.com™
 ☐ Rentals  
from Move.com™

Price Range: \$0 - No Maximum

Enter City Select a State

Or Enter ZIP Go

Senior Living Home Plans

Control

Existing Homes Foreclosures New Construction Rentals

Find Existing Homes for Sale


 Enter City State
 or  
 Enter Zip
 Find homes

Treatment 2

Find a new Home or Apartment

 Existing Homes
  New Construction
  Foreclosures
  Rentals

Enter Zip or Enter City State Search listings

Treatment 4

Existing Homes Foreclosures New Construction Rentals

Find Existing Homes for Sale


 Enter City State
 or  
 Enter Zip
 Find homes

Treatment1

What are you looking for?

☒ Existing Homes
 Enter City State
 ☐ New Construction
 Enter Zip
 ☐ Rentals
 \$0 to No Max
 ☒ Condos/Townhouse
 ☒ Single Family Home
 Find homes

☐ Foreclosures
 ☐ Senior Living
 ☐ Home Valuation
 ☐ Professional Services

Treatment 3

Find Your Dream Home or Apartment

City, State or ZIP

☒ Existing homes
 ☐ New construction
 ☐ Foreclosures
 ☐ Rentals
 Search listings

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.

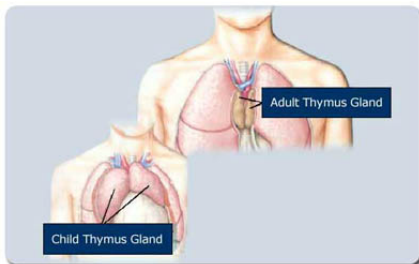
# Observational Studies and Confounding

- 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 in a way that makes it impossible to separate their effects on the responses [Watkins et al., 2010].



# A Confounded Observational Study

- The thymus, a gland located in our 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](http://esvc001414.wic005tu.server-web.com/tech_imm_bio_principle.htm)

# A 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.

# A Confounded Observational Study

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

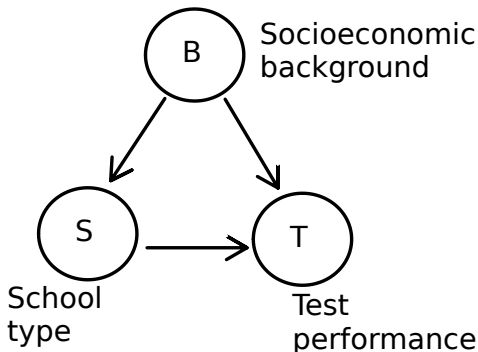
	Age	
	Child	Adult
Thymus size	Large Problems	No evidence
	Small No evidence	No problems

- 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.

## Another Example of Confounding



**Figure:** A possible causal explanation of socioeconomic background, school type, and test performance.

# 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 or **blocks**), 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 confounding 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 measure whether some assumed **property** about a population is contrasted with a statistical sample.
- By hypothesis we refer to a subset of values for our target population parameter  $\theta$ .
- In the context of experiments, NHST helps us to determine whether 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**  $\mu$ .



# Hypothesis Testing

- We will study the following types of parametric tests to the mean:
  - 1 **One sample test:** 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 first 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:

$$H_0: \mu = 3000g$$

$$H_1: \mu > 3000g$$

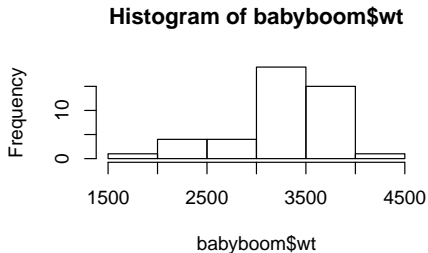
- The null hypothesis is written first followed by the research hypothesis.
- The research hypothesis is often called the **alternative hypothesis** even though it is often the first hypothesis we think of.

## 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  $\bar{x}$  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  $\bar{X}$  follows as Normal distribution when  $n$  is sufficiently large:  $\bar{X} \sim N(\mu, \sigma^2/n)$
- If we assume  $H_0$  is true, then  $\mu = 3000$ .
- The value of  $n$  is 44 and the value of  $\sigma$  is known in 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:

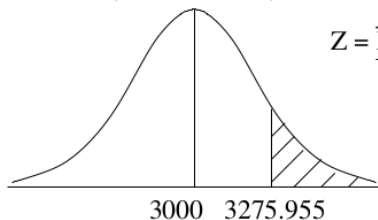
```
> #pvalue
> 1-pnorm(xbar, mean =mu0, sd =se)
[1] 0.0001256405
> #or
> Z.score<-(xbar-mu0)/se
> Z.score
[1] 3.660951
> p.value<-1-pnorm(Z.score)
> p.value
[1] 0.0001256405
```

## Example 2: Babies

$$\bar{X} \sim N(3000, 5681.818)$$

$$Z \sim N(0, 1)$$

$$P(\bar{X} > 3275.955)$$



$$Z = \frac{\bar{X} - 3000}{75.378}$$

$$P(Z > 3.66)$$

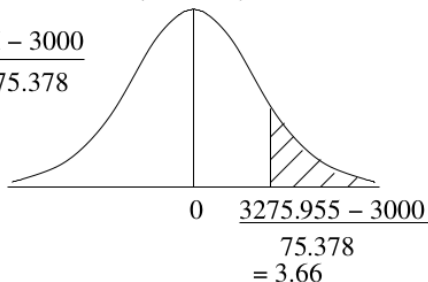


Figure: [Marchini, 2008]

## 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**  $\alpha$  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 significant 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”

# T-tests

- In the previous example, we assumed that  $\sigma$  was known.
- In many cases  $\sigma$  is unknown and we must estimate it using the unbiased estimator  $s$  that we saw in the previous class.
- In these cases we can calculate a  $T$  statistic  $\frac{\bar{X}_n - \mu_0}{\frac{s}{\sqrt{n}}}$

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

# T-tests

- From previous class we know that  $T$  follows a t-student distribution with  $n - 1$  degrees of freedom  $T \sim t_{n-1}$ .
- We can now perform a T-test using the t-student distribution instead of a Gaussian.
- The p-value can be calculated analogously 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 because the t-distribution has wider tails than the Normal distribution.
- The wide tails imply that there is more uncertainty because we had to estimate  $\sigma$ .

# T-tests

- We can perform t-tests straightforwardly 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
```



# Calculating a critical region

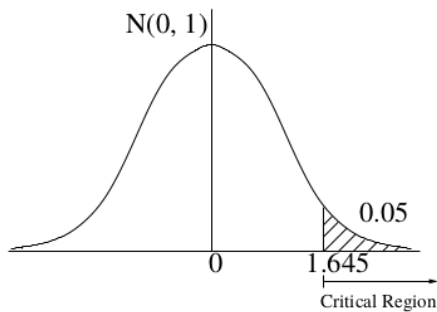
- Another way of thinking about this test is that there is some critical region of values such that if the test statistic lies in this region then we will reject  $H_0$ .
- If the test statistic lies outside this region we will not reject  $H_0$ .
- In the babies example, using a 5% level of significance this set of values will be the most extreme 5% of values in the right hand tail of the distribution.
- We can calculate that the boundary of this region, called the critical value:

```
> crit<-qnorm(1-alpha)
> crit
[1] 1.644854
```

- The value of our test statistic is 3.66 which lies in the critical region so we reject the null hypothesis at the 5% level.

```
> Z.score>=crit
[1] TRUE
```

# Calculating a critical region



- Alternatively, using a T-distribution:

```
> crit2<-qt(1-alpha, df = n-1)
> crit2
[1] 1.681071
> T.sta>=crit2
[1] TRUE
```

# Overview of NHST

## 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**  $H_0$ .

# Overview of NHST

## Methodology to Perform a Hypothesis Test

- Specify a null hypothesis  $H_0$  and alternative  $H_a$ .
- Set a test significance level  $\alpha$ .
- 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 (e.g., a Z-score).
- Assess the “statistical significance” of  $T$ .

The last part can be done with two approaches

- 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  $\alpha$ .
- Critical region: Calculate a region of values such that if  $T$  lies in this region then we will reject  $H_0$ .

# More on P-values

- 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.
- P-values allow us to quantify this.
- 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 (larger than the sample mean in previous example).
- We must consider all more extreme values because the probability of any particular value (such as the observed sample mean) is zero for continuous distributions.
- We must recall that we are trying to determine how weird our result would be if the null hypothesis were true.
- Hence, any result that is more extreme will be even more weird.
- So we want to count all of those weirder possibilities when we compute the probability of our result under the null hypothesis.

# Two-sided Tests

- In the previous example we wanted to test the research hypothesis that mean birth weight of Australian babies was greater than 3000g.
- This suggests that we had some prior information that the mean birth weight of Australian babies was definitely not lower than 3000g.
- If this were not the case then our research hypothesis would be that the mean birth weight of Australian babies was different from 3000g.
- This allows for the possibility that the mean birth weight could be less than or greater than 3000g.
- This is an example of a **two-sided** test as opposed to the previous example which was a **one-sided** test.
- In this two-sided case we would write our hypotheses as

$$H_0: \mu = 3000g$$

$$H_1: \mu \neq 3000g$$

# Two-sided Tests

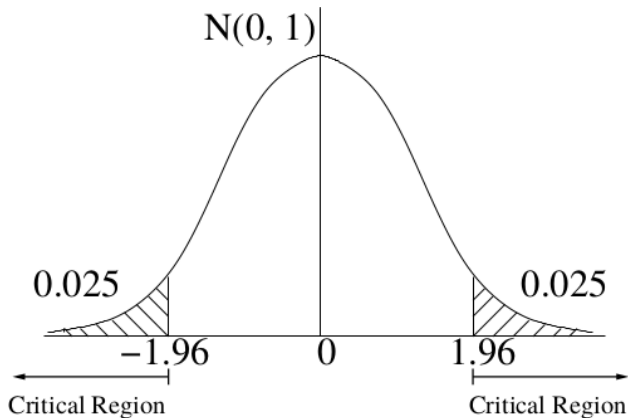
- As before we would calculate our test statistic as 3.66 for the Normal distribution and 3.47 for the T-student.
- In this case we allow for the possibility that the mean value is less than 3000g by setting our critical region to be lowest 2.5% and highest 2.5% of the distribution.
- In this way the total area of the critical region remains 0.05 and so the level of significance  $\alpha$  of our test remains 5%.
- The critical values for a Z-test are:

```
> crit.left<-qnorm(alpha/2)
> crit.left
[1] -1.959964
> crit.right<-qnorm(1-alpha/2)
> crit.right
[1] 1.959964
```

- Thus if our test statistic is less than -1.96 or greater than 1.96 we would reject the null hypothesis.
- In this example, the value of test statistic does lie in the critical region so we reject the null hypothesis at the 5% level.

```
> Z.score<=crit.left | Z.score >= crit.right
[1] TRUE
```

# Two-sided Tests





# Two-sided Tests

- For the case of the T-distribution our critical region is:

```
> crit2.left<-qt(alpha/2,df = n-1)
> crit2.left
[1] -2.016692
> crit2.right<-qt(1-alpha/2,df = n-1)
> crit2.right
[1] 2.016692
>
> T.sta<=crit2.left |T.sta >= crit2.right
[1] TRUE
```

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

# Two-sided Tests

- Alternatively, we could calculate a confidence interval for the sample mean with  $(1 - \alpha)\%$  confidence.
- The confidence interval becomes the **acceptance region** and we reject  $H_0$  if  $\mu_0 = 3000$  is not trapped by the interval.

```
> left.conf<-xbar-qt (p=1-alpha/2,n-1)*se.t  
> left.conf  
[1] 3115.418  
> right.conf<-xbar+qt (p=1-alpha/2,n-1)*se.t  
> right.conf  
[1] 3436.491  
> mu0 >= left.conf | mu0 <= right.conf  
[1] TRUE
```

- Since  $\mu_0 = 3000$  is not in my acceptance region, we reject the null hypothesis at the 0.05 significance level.
- Confidence intervals and p-values always agree on statistical significance [Editor, 2015].

# Two-sided Tests

- In order to calculate a p-value in a two-sided test we need to consider both left and right tails:

```
> pvalue<-pt (-T.sta,df=n-1) + (1-pt (T.sta,df=n-1))  
> pvalue  
[1] 0.001208524  
> # or more compactly  
> 2*pt (-abs(T.sta),df=n-1)  
[1] 0.001208524
```

- Notice that this p-value is larger than for the one-sided test.
- This reflects the fact that an extreme value is less surprising since it could have occurred in either direction.

# Two-sided Tests

- We can run a two-sided t-test in R with one single call:

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

One Sample t-test

```
data:  babyboom$wt  
t = 3.4666, df = 43, p-value = 0.001209  
alternative hypothesis: true mean is not equal to 3000  
95 percent confidence interval:  
 3115.418 3436.491  
sample estimates:  
mean of x  
 3275.955
```

# Unpaired Two Sample Tests

- The babyboom dataset has a column specifying the gender of each baby.

```
> summary(babyboom$gender)
girl  boy
  18   26
```

- Suppose our research hypothesis is that the mean birth weight of boys is different (two-sided) than mean birth weight of girls:

$$H_0: \mu_{boys} = \mu_{girls} \text{ or } \mu_{boys} - \mu_{girls} = 0$$

$$H_1: \mu_{boys} \neq \mu_{girls} \text{ or } \mu_{boys} - \mu_{girls} \neq 0$$

- We call this test a two sample tests (one sample with the births of boys and the other of girls).
- The two samples are independent or unpaired (we have different number of observations for boys and girls).
- This types of tests are very important for experimental data and observational studies (i.e., one sample is the control group and the other is the treatment).

# Unpaired Two Sample Tests

- Asymptotic theory tells us that the difference between two sample means (when the sample sizes are sufficiently large) has a Normal sampling distribution:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

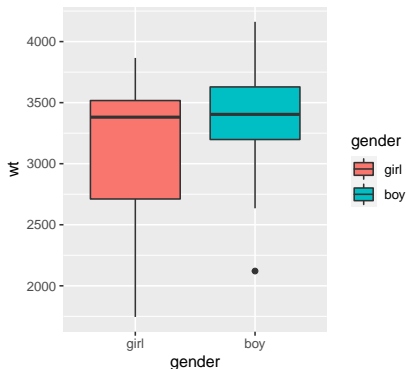
- When the standard deviations of each group are unknown ( $\sigma_1, \sigma_2$ ) we can estimate them as usual ( $s_1, s_2$ ) and build the following  $T$  statistics:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- The  $T$  statistic is distributed according to a t-student distribution.

# Unpaired Two Sample Tests

- When the groups are the same size and have equal variance, the degrees of freedom for the  $T$  test is  $n_1 + n_2 - 2$ .
- In this case the box plot shows that the “girl” group is more variable than the “boy” group.
- We also know that the number of observations in each group is different.



# Unpaired Two Sample Tests

- We need to use a more complex formula for the degrees of freedom, which is often referred to as a “Welch t-test”:

$$d.f. = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

- For this example  $d.f. = 27.631$  which is lower than what we would get by subtracting 2 from the sample size.
- Recall that the lower the d.f. the wider the tails in the t-student distribution.
- This is essentially, imposing a penalty on the test for differences in sample size or variance.



# Unpaired Two Sample Tests

- We can run a Welsh T-test in R as follows:

```
> t.test(babyboom$wt ~ babyboom$gender)
```

```
Welch Two Sample t-test
```

```
data: babyboom$wt by babyboom$gender
```

```
t = -1.4211, df = 27.631, p-value = 0.1665
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-593.1538 107.4273
```

```
sample estimates:
```

```
mean in group girl mean in group boy
```

```
3132.444
```

```
3375.308
```

- In this case, we fail to reject  $H_0$  at the 5% significance level.

# Paired Two Sample Tests

- Another very useful type of two sample test is the Paired T-Test.
- This test is used to compare the means between two **related groups** of samples.
- Here we have two values (i.e., pair of values) for the same samples [Kassambara, ].
- This type of data arises when we compare a response variable after and before a treatment for each subject.
- It also arises in matched pairs experiments.
- Paired t-test analysis is performed as follow:
  - 1 Calculate the difference  $d$  between each pair of value.
  - 2 Compute the mean  $\bar{d}$  and the standard deviation  $s_d$  of these differences.
  - 3 Compare the  $\bar{d}$  to 0 in the same way as previous tests.

# Paired Two Sample Tests

- The test statistics  $T$  is calculated as follows for the paired test:

$$T = \frac{\bar{d}}{s_d/\sqrt{n}}$$

- The the degrees of freedom (df) are simply  $n - 1$ .
- As an example of data, 20 mice received a treatment X during 3 months.
- We want to know whether the treatment X has an impact on the weight of the mice.
- The weight of the 20 mice has been measured before and after the treatment.
- Let's test the following hypotheses:

$$H_0: \bar{d} = 0$$

$$H_1: \bar{d} \neq 0$$

# Paired Two Sample Tests

- We can run a paired t-test in R as follows:

```
> before <-c(200.1, 190.9, 192.7, 213, 241.4, 196.9, 172.2,  
+           185.5, 205.2, 193.7)  
> # Weight of the mice after treatment  
> after <-c(392.9, 393.2, 345.1, 393, 434, 427.9, 422,  
+           383.9, 392.3, 352.2)  
> t.test(after, before, paired = TRUE, alternative = "two.sided")
```

Paired t-test

```
data: after and before  
t = 20.883, df = 9, p-value = 6.2e-09  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 173.4219 215.5581  
sample estimates:  
mean of the differences  
      194.49
```

# 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 was true (also called “false alarm”).
- Type II error: is when the null hypothesis is false but we do not have statistical evidence to reject it (also called a “miss”).

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

- Neyman and Pearson coined two terms to describe the probability of these two types of errors in the long run:
  - $P(\text{Type I error}) = \alpha$
  - $P(\text{Type II error}) = \beta$

# Errors

- If we set  $\alpha$  to .05, then in the long run we should make a Type I error 5% of the time.
- The standard value for an acceptable level of  $\beta$  is .2.
- That is, we are willing to accept that 20% of the time we will fail to detect a true effect when it truly exists.
- The concept of **statistical power** is the complement of Type II error:

$$power = 1 - \beta$$

- The power of a test is the likelihood of finding a positive result given that it exists.[Poldrack, 2019]
- To mitigate type I errors we generally use smaller values of  $\alpha$ .
- To mitigate type II errors (or increase the power of the test) we generally work with larger samples.
- There is a trade-off between type I and type II errors.
- There are tools to analyze the power of a test that go beyond the scope of this course.

# What does a significant result mean?

- There is a great deal of confusion about what p-values actually mean.
- Suppose we do an experiment comparing the means between conditions, and we find a difference with a p-value of .01.
- Does it mean that the probability of the null hypothesis being true is .01?
  - No. Remember that in NHST, the p-value is the probability of the data given the null hypothesis:  $P(data|H_0)$ .
  - It does not warrant conclusions about the probability of the null hypothesis given the data:  $P(H_0|data)$ .
- Does it mean that the probability that you are making the wrong decision is .01?
  - No. This would be  $P(H_0|data)$ , but remember as above that p-values are probabilities of data under  $H_0$ , not probabilities of hypotheses.

# What does a significant result mean?

- Does it mean that you have found a practically important effect?
  - No. There is an essential distinction between statistical significance and practical significance.
  - Suppose we performed an RCT to examine the effect of a particular diet on body weight, and we find a statistically significant effect at  $p < .05$ .
  - This doesn't tell us how much weight was actually lost.
  - The loss of one ounce (i.e. the weight of a few potato chips) can be statistically significant but not practically significant.
- Many scientist think that NHST is flawed and that is has been the cause of serious problems in science [Poldrack, 2019].
- For example, The American Statistical Association (ASA) released a "Statement on Statistical Significance and P-Values" indicating the proper use and interpretation of the p-value [Wasserstein and Lazar, 2016].



# There are many other tests

There are a plethora of other tests that we will not teach in this course

- Proportion tests.
- The Fisher's exact test.
- Analysis of Variance (ANOVA).
- The Chi-square tests of independence.
- The Wilcoxon signed-rank test.
- The Kolmogorov–Smirnov tests.

# Conclusions

- In this class we have introduced two important statistical concepts: design of experiments and NHST.
- Experiments are a powerful approach to determining cause-effect relationships.
- It is very important to identify and control confounding variables in the design of experiments.
- Hypothesis testing is a family of techniques for testing hypotheses using data.
- NHST must be used with care and we should always remind that p-values do not measure the probability of a given hypothesis.

# 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.



Editor, M. B. (2015).  
Understanding hypothesis tests: Confidence intervals and confidence levels.  
[https://blog.minitab.com/en/adventures-in-statistics-2/  
understanding-hypothesis-tests-confidence-intervals-and-confidence](https://blog.minitab.com/en/adventures-in-statistics-2/understanding-hypothesis-tests-confidence-intervals-and-confidence)



Fisher, R. A. (1936).  
Design of experiments.  
*Br Med J*, 1(3923):554–554.



Kassambara, A.  
Paired samples t-test in r.  
[http:  
//www.sthda.com/english/wiki/paired-samples-t-test-in-r.](http://www.sthda.com/english/wiki/paired-samples-t-test-in-r)



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.

# References II



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/>.



Wasserstein, R. L. and Lazar, N. A. (2016). The asa statement on p-values: context, process, and purpose.



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