Chapter 8: Hypothesis Testing

MATH 361: Probability & Statistics I

Prof. Katie Fitzgerald, PhD

Spring 2022

Contents

Treating Chronic Fatigue Syndrome . . . 2
"Stochastic proof by contradiction" . . . 2
Making a decision: Critical regions . . . 4
Decision making trade-offs . . . 4
Example: law . . . 5
Example: medicine . . . 5
Hypothesis tests: decision-making in statistics . . . 6
Significance level α . . . 6
Statistical power . . . 7
CI vs HT . . . 7
Case study: Gender discrimination . . . 8

Treating Chronic Fatigue Syndrome

In many scientific pursuits, the goal is not simply to estimate a population parameter. Instead, the goal is often to understand if there is a difference between two groups in the population or if there is a relationship between two (or more) variables in the population, and to make a decision based off of the evidence.

Recall the results of a clinical trial of a new treatment for Chronic Fatique Syndrome we discussed in Chapter 6:

—	Improved	Did not improve	Total
Treatment	19	8	27
Control	5	21	26
Total	24	29	53

$\hat{p}_t = 19/27 = .7037$

$\hat{p}_c = 5/26 = .1923$

$\hat{p}_t - \hat{p}_c = .51$

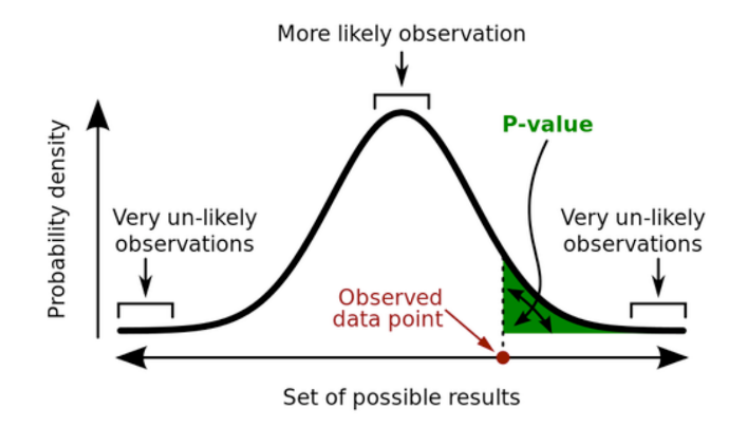
Do these data show a "real" difference between the groups? Is there sufficient evidence to claim the treatment was effective, or was this perhaps due to chance?

"Stochastic proof by contradiction"

There are three steps in a Proof by Contradiction. In order to illustrate these, assume we wish to prove that there is a relationship between X and Y.

1. Negate the conclusion: Begin by assuming the opposite – that there is no relationship between X and Y. (establish a null hypothesis)
2. Analyze the consequences of this premise: If there is no relationship between X and Y in the population, what would the sampling distribution of the estimate of the relationship between X and Y look like?
3. Look for a contradiction: Compare the relationship between X and Y observed in your sample to this sampling distribution. How (un)likely is this observed relationship?

If likelihood of the observed relationship is small (given your assumption of no relationship), then this is evidence that there is in fact a relationship between X and Y in the population.



Once the distribution of the sample statistic under the null hypothesis is determined, to complete the stochastic proof by contradiction, you simply need to ask: Given this distribution, how likely is it that I would have drawn a random sample in which the estimated value is this extreme or more extreme?

This is called the **p-value**: The probability of your observing an estimate as extreme as the one you observed if the null hypothesis is true. If this p-value is small, it means that this data is unlikely to occur under the null hypothesis, and thus the null hypothesis is unlikely to be true. (See, proof by contradiction!)

CFS Example (cont'd)

We want to determine if there is a relationship between the treatment and patient outcomes. Does the treatment affect the proportion of patients who have good outcomes? Mathematically, we want to determine if $p_T \neq p_C$, where p_T is the proportion of patients in the treatment group who had good outcomes, and p_C is the proportion of patients in the control group who had good outcomes. What are the three steps for the proof by contradiction for this problem?

1. Negate the conclusion: $p_T = p_C$

2. Analyze the consequences of this premise:

$$p_T = p_C \Rightarrow p_T - p_C = 0$$

$$\Rightarrow \hat{p}_T - \hat{p}_C \approx N\left(p_T - p_C, \underbrace{\frac{p_T(1-p_T)}{n_T} + \frac{p_C(1-p_C)}{n_C}}_{V(\hat{p}_T - \hat{p}_C)}\right)$$

$$\Rightarrow \frac{\hat{p}_T - \hat{p}_C - 0}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_C}}} \approx N(0,1)$$

3. Look for a contradiction: Compare what's observed in our sample (i.e. compute $\hat{p}_T - \hat{p}_C$) to what would be expected if the null hypothesis is true. How (un)likely is this observed value? *Hint: compute a p-value in two ways*

(1) Via simulations: https://www.rossmanchance.com/applets/2021/twopopprop/twopopprop.html

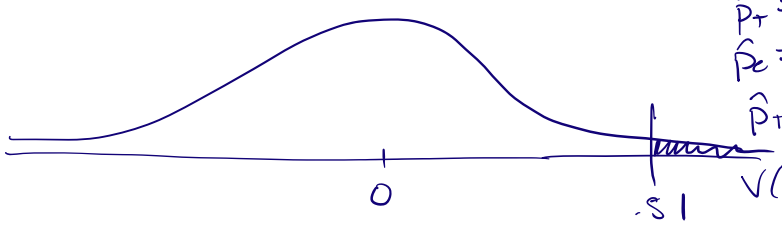
(2) Mathematically, by computing $P\left((\hat{p}_T - \hat{p}_C) > 0.51\right)$, assuming H_0 is true.

$\hat{p}_T = .7037$

$\hat{p}_C = .1923$

$\hat{p}_T - \hat{p}_C = .51$

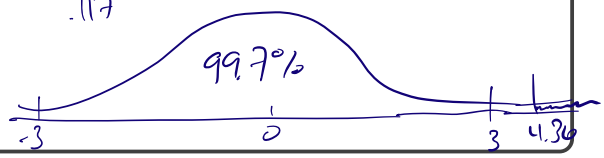
$V(\hat{p}_T - \hat{p}_C) =$



1 - pnorm(.51, 0, sqrt(V($\hat{p}_T - \hat{p}_C$))) ≈

Note: converting to a Z-score can also be useful in determining how unlikely an observed value is.

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} \qquad Z = \frac{\hat{p}_T - \hat{p}_C - 0}{\sqrt{V(\hat{p}_T - \hat{p}_C)}} = \frac{.51}{.117} = 4.36$$



Making a decision: Critical regions

We start with some null hypothesis, H_0 about the parameter of interest, θ.

We observe a point estimate $\hat{\theta}$ from our sample, and we know the sampling distribution of $\hat{\theta}$

We determine some **critical region**, C, such that if $\hat{\theta} \in C$, we reject H_0, and if $\hat{\theta} \in C'$ we do not reject H_0.

How do we determine the critical region? There are tradeoffs...

Decision making trade-offs

Imagine that you've been invited to a party, and you are trying to decide if you should go. On the one hand, the party might be a good time, and you'd be happy that you went. On the other hand, it might not be that much fun and you'd be unhappy that you went. In advance, you don't know which kind of party it will be. How do you decide?

TABLE 12.1: Party decision making

•	**Party is fun**	**Party is not fun**
Go to party	Great decision!	Type I error (wasted time)
Stay home	Type II error (missed out)	Great decision!

If you're someone that has FOMO – which type of error are you more concerned about?

Type II

What could you do to minimize that error?

go to all parties

What's the trade-off of minimizing that error?

increase Type I (waste time at not-fun parties)

If you're someone who hates wasting time – which type of error are you more concerned about?

Type I

What could you do to minimize that error?

Stay home always

What's the trade-off of minimizing that error?

increase Type II - miss out on fun parties

When making a decision, you cannot know in advance what the actual outcome will be.

Sometimes your decision will be the right one. Ideally, you'd like this to be most of the time. But, sometimes your decision will be the wrong one.

Importantly, you cannot minimize both Type I and II errors at the same time. One will be minimized at the expense of the other.

Depending upon the context, you may decide that minimizing Type I or II errors is more important to you.

Example: law

Imagine that you are on the jury of a criminal trial. You are presented with evidence that a crime has been committed and must make a decision regarding the guilt of the defendant. But you were not there when the crime was committed, so it is impossible to know with 100% accuracy that your decision is correct. Instead, you again encounter this 2x2 table:

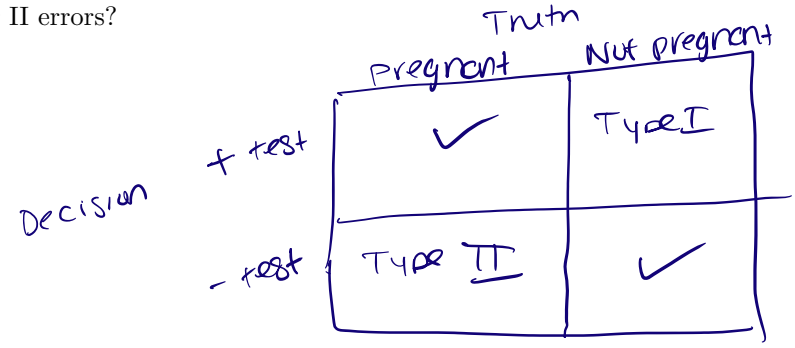
TABLE 12.3: Criminal trial decision making

•	**Guilty**	**Innocent**
"Guilty" verdict	Correct	Type I error: Wrongly Convicted
"Not Guilty" verdict	Type II error: Insufficient Evidence	Correct

Example: medicine

Imagine that you might be pregnant and take a pregnancy test. This test is based upon levels of HcG in your urine, and when these levels are "high enough" (determined by the pregnancy test maker), the test will tell you that you are pregnant (+). If the levels are not "high enough", the test will tell you that you are not pregnant (-). Depending upon how the test determines "high enough" levels of HcG, however, the test might be wrong.

1. Set up a 2x2 table for the 4 possible outcomes
2. Describe in words what a Type I and Type II error mean in this context
3. In developing these tests, which do you think test manufacturers focus on minimizing: Type I or II errors?



Type I = false positive Type II = false negative

Hypothesis tests: decision-making in statistics

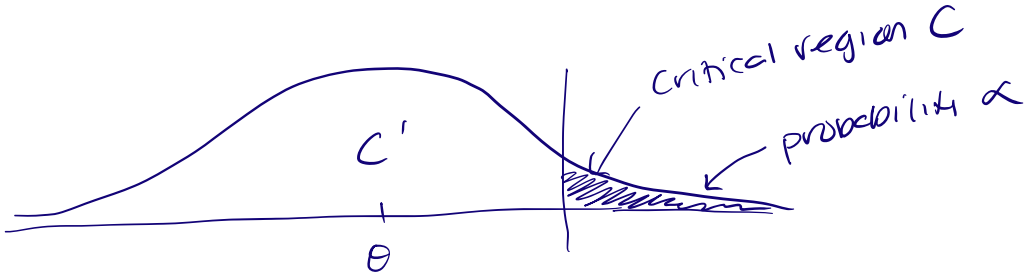
The same sort of decision making problems face statistics as well: based on some p-value criterion (or equivalently, some critical region), we could either reject the null hypothesis or not. And either the null hypothesis is true, or it is not.

.	H_0 is NOT true	H_0 is true
Reject H_0	✓	TYPE I
Do not reject H_0	TYPE II	✓

We always start by assuming the null hypothesis is true, and then see whether there is enough evidence to overturn it. So, the null hypothesis is typically a statement that there is "no relationship" or "no difference" (e.g. $p_1 = p_2$ or $\mu_1 = \mu_2$ or $\rho = 0$), and the alternative hypothesis we're trying to investigate is that there IS a relationship or difference.

Significance level α

The significance level α provides the cutoff for the p-value which will lead to a decision of "reject the null hypothesis." Mathematically, $P_\theta(\hat{\theta} \in C) = \alpha$



- Traditional level is $\alpha = 0.05$ (that is, we reject H_0 if p-value < 0.05).
- Note, this means, we will make a Type I error $\alpha * 100\%$ of the time when H_0 is true.
- However, **this is an arbitrary threshold!!**

Choosing the right α should be context-specific. It is sometimes helpful to adjust the significance level depending on the consequences of any conclusions reached from the test.

- If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g., 0.01 or 0.001).
- If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g., 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

Good, ethical statistical/scientific practice is to set an α value ahead of time. Why?

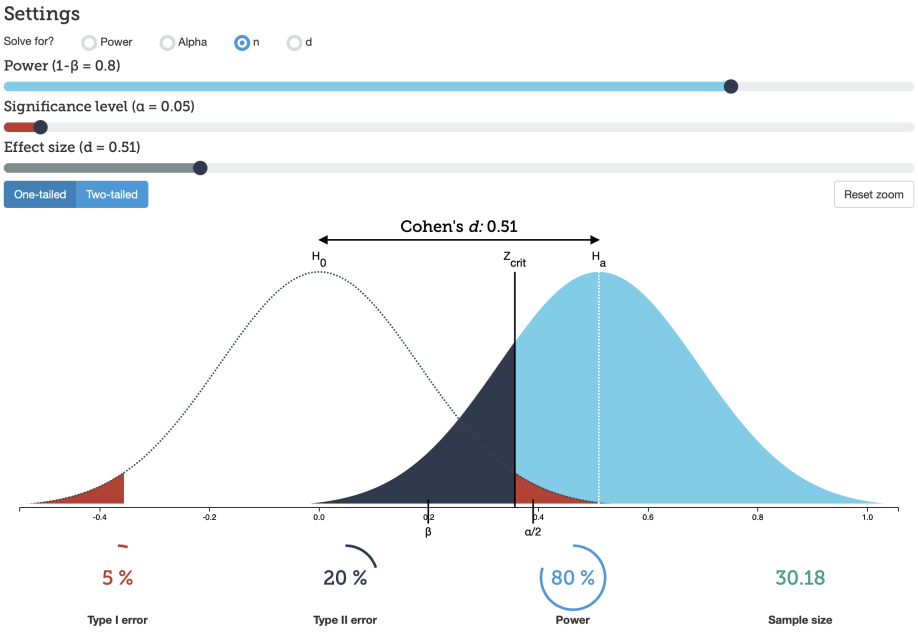
- To avoid "p-hacking" and manipulating the data

Statistical power

The probability of (correctly) rejecting the null hypothesis when it is false, is called the statistical **power** of the test. That is, the probability of NOT making a Type II error when H_A is true. In other words, the probability of being able to claim a treatment works when it does.

There is a trade-off between Type I and Type II errors, mediated by sample size. And, sample size = time, financial & logistical costs.

https://rpsychologist.com/d3/nhst/



CI vs HT

Asking "does the parameter fall outside the 95% confidence interval" is mathematically equivalent to asking "Is there sufficient evidence to reject the null hypothesis at the level $\alpha = 0.05$?"

- If the CI contains the parameter, FAIL to reject H_0
- If the CI does NOT contain the parameter, REJECT H_0

90% confidence corresponds to $\alpha = 0.10$

99% confidence corresponds to $\alpha = 0.01$

Etc. . . .

Case study: Gender discrimination

In 1972, as a part of a study on gender discrimination, 48 male bank supervisors were each given the same personnel file and asked to judge whether the person should be promoted to a branch manager job that was described as "routine".
The files were identical except that half of the supervisors had files showing the person was male while the other half had files showing the person was female.
It was randomly determined which supervisors got "male" applications and which got "female" applications.
Of the 48 files reviewed, 35 were promoted.
The study is testing whether females are unfairly discriminated against.

.	Promoted	Not promoted	Total
Male	21	3	24
Female	14	10	24
Total	35	13	48

At first glance, does there appear to be a relationship between promotion and gender? Compute the proportion of males promoted and proportion of females promoted.

$\hat{p}_M = .875$
$\hat{p}_F = .583$ $\Rightarrow \hat{p}_M - \hat{p}_F = .292$

Set up a proof by contradiction. State the null hypothesis (the negation of what the researchers are trying to prove) in terms of the parameters p_f and p_m.

$H_0: p_M = p_F \Rightarrow p_M - p_F = 0$

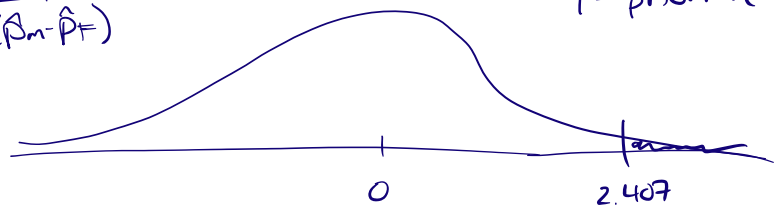
Assuming H_0 is true, how is $\hat{p}_m - \hat{p}_f$ distributed?

$$\hat{p}_M - \hat{p}_F \approx N\left(0, \frac{p_M(1-p_M)}{n_M} + \frac{p_F(1-p_F)}{n_F}\right)$$

Look for a contradiction. Compare the observed $\hat{p}_m - \hat{p}_f$ to the distribution you determined above. Is there enough to overturn (reject) H_0? The researchers used $\alpha = 0.05$. What did they conclude?
https://www.rossmanchance.com/applets/2021/twopopprop/twopopprop.html

$$Z = \frac{\hat{p}_M - \hat{p}_F - 0}{\sqrt{V(\hat{p}_M - \hat{p}_F)}} = 2.407$$

1 - pnorm(2.407) ≈ 0.008



p-value < α $\Rightarrow$ Reject H_0
Sufficient evidence to claim gender discrimination.