CSCI 3022 Intro to Data Science
Testing

It's been a rough week. How are you doing?

Announcements and Reminders

- Only about a "half" lecture today: want to make sure this week is slow paced.
- Trying to get exam grades done... but if they're not and you're worried about a decision for the last day to withdraw...

I will be available in my office hours Zoom ID *all day today*. Every minute from the end of class until 6pm, with a small break from 1:45-2:45pm for my other class. I can look over your work and give you a more detailed estimate. Before you do though, check out the detailed rubrics for the first 3 homeworks and the exam solutions!

Rule of thumb: My rubrics *always* give you about 50% of a problem if you tried to work through it to its completion. If you've *tried* every assignment in this class (or all but 1-2, since we drop 2 HWs!), then you're on track to pass.

Now what?

Group 1: Sample proportion $\hat{p}_1 \approx 30\%$

Group 2: Sample proportion $\hat{p}_2 \approx 35\%$

Decomposing an interval like the interval from our two-sample proportion test

$$\hat{p_1} - \hat{p_2} \pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

into a yes or no decision is how we transition into statistical hypothesis testing. Based on our confidence interval on $p_1 - p_2$, we can try to answer whether $p_1 = p_2$, $p_1 < p_2$, etc.

Intervals and Decisions

$(0,1)$ $(0,1)$

Suppose we get a CI for $p_1 - p_2$:

$[-.3, -.12]$ w/ confidence 90%.

We can make a decision. 100% of the time we decide, with some risk of being wrong: we decide that $p_1 - p_2$ is $[-.3, -.12]$.

$p_1 - p_2 < -.12 \Rightarrow p_1 < p_2 - .12$ Conclude p_1 is smaller than p_2, "for sure!" 90%

Statistical Hypotheses

Definition: *Statistical Hypothesis*
A *Statistical Hypothesis* is a claim about the value of a parameter or population characteristic.

test "this coin is fair"

Examples:

ask: does data look like theoretical data of a fair coin.

Statistical Hypotheses

Definition: *Statistical Hypothesis*
A *Statistical Hypothesis* is a claim about the value of a parameter or population characteristic.

Examples:

test : $\mu_A > \mu_B$

1. Company A makes parts that last longer than company B.

Statistical Hypotheses

Definition: *Statistical Hypothesis*
A *Statistical Hypothesis* is a claim about the value of a parameter or population characteristic.

Examples:

1. Company A makes parts that last longer than company B.
2. In Boulder, it's usually a colder maximum daily temperature in February than June.

$T: \quad \mu_2 < \mu_6$

Statistical Hypotheses

Definition: *Statistical Hypothesis*
A *Statistical Hypothesis* is a claim about the value of a parameter or population characteristic.

Examples:

1. Company A makes parts that last longer than company B.
2. In Boulder, it's usually a colder maximum daily temperature in February than June.
3. Students in Zach's sections are generally much more dashing, resourceful, and socially meritorious than students in other sections.

Statistical Hypotheses

One example statisticians often revisit: is a coin fair?
This is a real world question!

`https://www.newscientist.com/article/dn1748-euro-coin-accused-of-unfair-flipping/`

As the Euro was introduced, Polish Mathematicians claimed that the Belgian 1 Euro coin was weighted so that it was more likely ot return a heads!
Suppose I handed you such a coin. How would you decide whether it ~~was fair~~?

2 questions: 1) "What does fair look like" ~ Probability
2) "Does our data fall into the "usual" spectrum of fair outcomes.

Logic of Hypothesis Testing

Analogy: Jury in a criminal trial.

When a defendant is accused of a crime, the jury (is supposed to) presumes that she is not guilty (not guilty; that's the "null hypothesis").

hypothesis that nothing matters — group A = group B, coin is fair, baseline

Then, we gather evidence. If the evidence is seems implausible under the assumption of non-guilt, we might reject non-guilt and claim that the defendant is (likely) guilty.

Logic of Hypothesis Testing

Important Question: Is there strong evidence for the alternative?

The burden of proof is placed on those who believe in the alternative claim.

The initially favored claim, the null hypothesis H_0, will not be rejected in favor of the alternative hypothesis, H_a or H_1, unless the sample evidence provides a lot of support for the alternative.

The two possible conclusions:

Logic of Hypothesis Testing

Important Question: Is there strong evidence for the alternative?

The burden of proof is placed on those who believe in the alternative claim.

The initially favored claim, the null hypothesis H_0, will not be rejected in favor of the alternative hypothesis, H_a or H_1, unless the sample evidence provides a lot of support for the alternative.

The two possible conclusions:
Fail to Reject the null hypothesis if there is insufficient statistical evidence to do so.

- coin doesn't look too unfair.

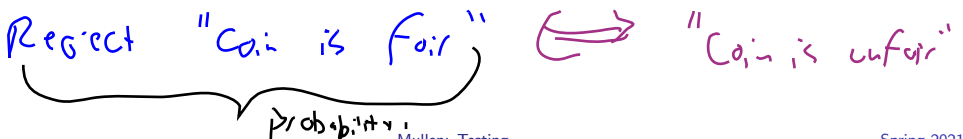
Logic of Hypothesis Testing

Important Question: Is there strong evidence for the alternative?

The burden of proof is placed on those who believe in the alternative claim.

The initially favored claim, the null hypothesis H_0, will not be rejected in favor of the alternative hypothesis, H_a or H_1, unless the sample evidence provides a lot of support for the alternative.

The two possible conclusions:
Fail to Reject the null hypothesis if there is insufficient statistical evidence to do so.
Reject the null hypothesis in favor of the alternative if there is statistically *significant* cause to do so.



Logic of Hypothesis Testing

Notation and general process:

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest θ satisfies $H_0 : \theta = \theta_0$.

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest θ satisfies $H_0 : \theta = \theta_0$.

2. State the alternative to be tested: H_a :
   $\theta > \theta_0$ **OR** $\theta < \theta_0$ **OR** $\theta \neq \theta_0$

3. Draw a decision based on how improbable or probable the actual data looks if the null hypothesis is true. If the observed data is very unlikely, it might be because our hypothesis was wrong!

Why *assume* the null hypothesis?

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest satisfies $H_0 : \theta = \theta_0$.

Why *assume* the null hypothesis?

1. Burden of proof
2. We know how to calculate probabilities when we *know* θ!

Logic of Hypothesis Testing

assume

The alternative to the null hypothesis $H_0 : \theta = \theta_0$ will look like one of the following three assertions:

Conclude:

1) reality is larger than baseline
2) reality is smaller than assumed/baseline
3) Either (1) or (2) is true

The equality sign is **always** with the null hypothesis.
The alternate hypothesis is the claim for which we are seeking statistical evidence.

Logic of Hypothesis Testing

The alternative to the null hypothesis $H_0 : \theta = \theta_0$ will look like one of the following three assertions:

1. $H_a : \quad \theta \neq \theta_0$

The equality sign is **always** with the null hypothesis.
The alternate hypothesis is the claim for which we are seeking statistical evidence.

Logic of Hypothesis Testing

The alternative to the null hypothesis $H_0 : \theta = \theta_0$ will look like one of the following three assertions:

1. $H_a : \quad \theta \neq \theta_0$

2. $H_a : \quad \theta > \theta_0$

3. $H_a : \quad \theta < \theta_0$

The equality sign is **always** with the null hypothesis.
The alternate hypothesis is the claim for which we are seeking statistical evidence.

Logic of Hypothesis Testing

Example: Suppose a company is considering putting a new type of coating on bearings that it produces.

The true average wear life with the current coating is known to be 1000 hours. With denoting the true average life for the new coating, the company would not want to make any (costly) changes unless evidence strongly suggested that exceeds 1000.

Logic of Hypothesis Testing

Example: An appropriate problem formulation would involve testing:

H_0:

H_a:

The conclusion that a change is justified is identified with H_a, and it would take conclusive evidence to justify rejecting H_0 and switching to the new coating.

Scientific research often involves trying to decide whether a current theory should be replaced, or "elaborated upon."

Logic of Hypothesis Testing

Example: An appropriate problem formulation would involve testing:

H_0: New company lifetime average is 1000

H_a: New company lifetime exceeds 1000

The conclusion that a change is justified is identified with H_a, and it would take conclusive evidence to justify rejecting H_0 and switching to the new coating.

Scientific research often involves trying to decide whether a current theory should be replaced, or "elaborated upon."

Test Statistics: The Evidence

$\bar{X}$? $\bar{X}$ – proposed baseline.

→ Sample Proportion

Definition: *Test Statistic*: Compute to compare to baseline

A test statistic is a a quantity derived based on sample data and calculated under the null hypothesis. It is used in a decision about whether to reject H_0.

We can think of a test statistic as our evidence. Next, we need to quantify whether we think our evidence is "rare" under the null hypothesis.

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$

assume

possible conclusion

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$
2. Get to flippin', collect some data

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$
2. Get to flippin', collect some data
3. Compute something from our data. Maybe a sample proportion of heads $\hat{p}$?

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$
2. Get to flippin', collect some data
3. Compute something from our data. Maybe a sample proportion of heads $\hat{p}$?
4. Decide whether $\hat{p}$ is **too far** from $p = .5$, and make a decision accordingly.

Test Statistics: The Evidence

Which test statistic is "best"?

There are an infinite number of possible tests that could be devised, so we have to limit this in some way or total statistical madness will ensue!

In the previous example, we might use $\hat{p}$.

Rejection Regions

baseline: what $\hat{p}$'s happen if coin is fair

what $\hat{p}$'s aren't in that region

How would we know when the test statistic is "sufficiently rare" under the null hypothesis such that we might regard the null as false?

We could define a **rejection region**: a range of values of the test statistic that leads a researcher to reject the null hypothesis.

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

IF FAIR:

- What would 10 heads mean? → outcome w/ prob = $\frac{1}{2^{10}}$

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

- What would 10 tails mean? $P(\text{all tails}) = 1/2^{10}$

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

- What would 6 heads mean? $P(\text{at most}) = \binom{10}{6}(.5)^6(.5)^4 = \binom{10}{8}\frac{1}{2^{10}}$

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

more probability on .4, .5, .6 for n=10

- Is there a difference between 60% heads if we flip 10 times and 60% heads if we flip 1000 times?

than for [.4, .600, ...8] for n=1000.

What is extreme: let's compute these!

Bring back α!

Definition: The **Significance level** α of a hypothesis test is the largest *probability* of a test statistic under the null hypothesis that would lead you to reject the null hypothesis.

Equivalently, it's the probability of the entire rejection region!

We thought of α last week during CIs as a term that widened or shrank as our tolerance for error grew, now it's very literally an *error rate*. Specifically, it's the probability of rejecting the null hypothesis when we were not supposed to do so.

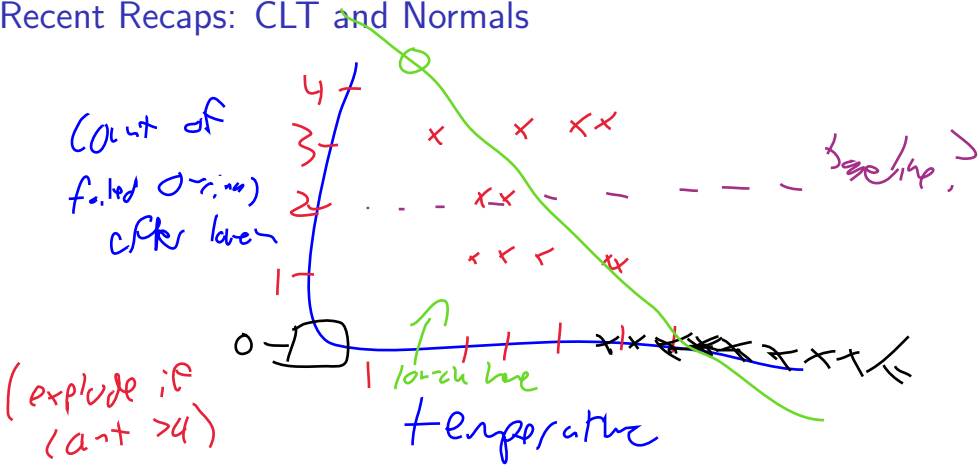
Where we at?

A summary of our process:

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$
2. Get to flippin', collect some data
3. Compute something from our data. Maybe a sample proportion of heads $\hat{p}$?
4. Decide whether $\hat{p}$ is **too far** from $p = .5$, and make a decision accordingly.
5. α is the value that describes the probability of rejecting a null hypotheses *given* that the hypothesis was true.

Recent Recaps: Expectation and Variance

Recent Recaps: CLT and Normals



Recent Recaps: One-Sample CIs

Recent Recaps: Two-Sample CIs

Daily Recap

Today we learned

1. Intro and Basics of Hypothesis Tests

Moving forward:

- nb day Friday over CIs

Next time in lecture:

- Hypotheses!