

Wrapping Up Hypothesis Tests: Guidelines, Errors and Caveats

LECTURE 22

CSCI 3022

Maribeth Oscamou

Content credit: [Acknowledgments](#)

Announcements

- Homework 9 released tonight
- Exam 2: Moved from Nov 11th to Friday Nov 17th

Today's Roadmap

CSCI 3022

- Wrapping Up Hypothesis Tests:
 - Guidelines
 - Errors
 - Caveats

Null Hypothesis:

- Null is meant to describe lack of an interesting pattern
 - Results are **due to chance**
- Need to be able to either calculate null distribution theoretically or simulate data under the null hyp.

Alternative Hypothesis:

- Should align with the question of interest

Null and **alternative** hypothesis can't be true at the **same time**. (Reject the Null → Alternative)

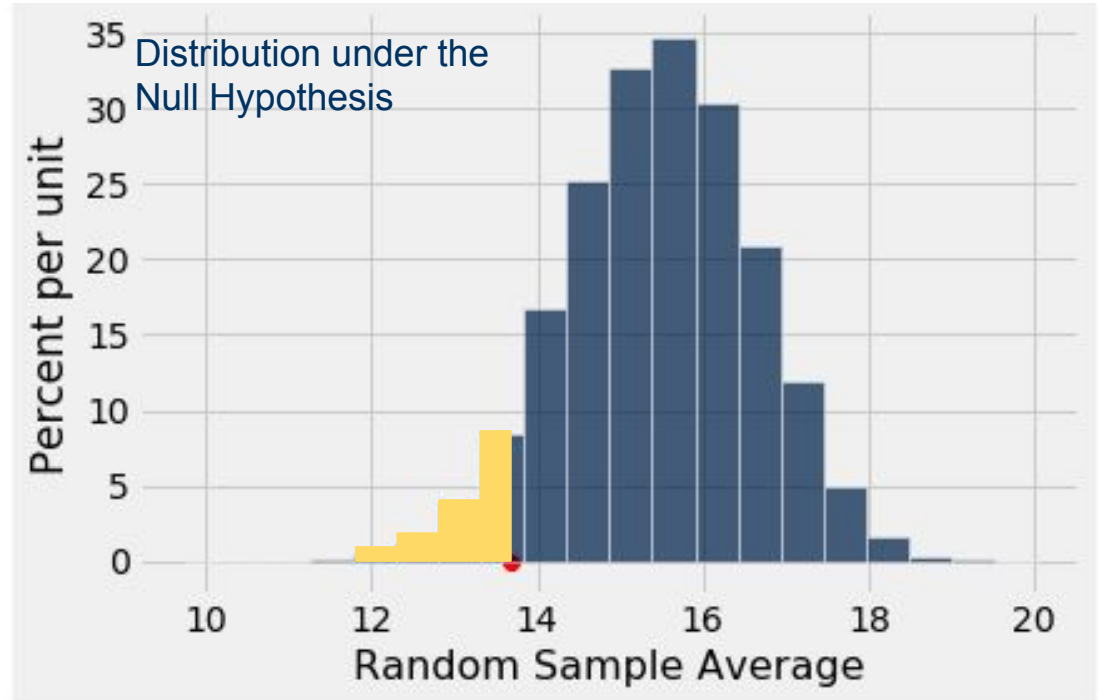
The outcome of a hypothesis test can be affected by:

- **The hypotheses you investigate:**
How do you define your null distribution?
- **The test statistic you choose:**
How do you measure a difference between samples?
- **The empirical distribution of the statistic under the null:**
How many times do you simulate under the null distribution?
- **The data you collected:**
Did you happen to collect a sample that is similar to the population?
- **The truth:**
If the alternative hypothesis is true, how extreme is the difference?

P-Values and Error Probabilities

Review: The p-Value as an Area

- Empirical distribution of the test statistic **under the null hypothesis**.
- Red dot denotes the observed statistic.
- Yellow area denotes the tail probability (p-value).



Formal name: **observed significance level**

The p -value is the chance (probability),

- under the null hypothesis,
- that the test statistic
- is equal to the value that was observed in the data
- or is even further in the direction of the alternative.

Discussion Question

Suppose there are 2000 CS Majors. We give each CS major a separate coin and have them toss it 160 times to test whether or not the coin is fair.

Null: The coin is fair

Alternative: The coin is unfair

- Test Statistic: ?
- Significance level (cutoff for the P-value): 5%

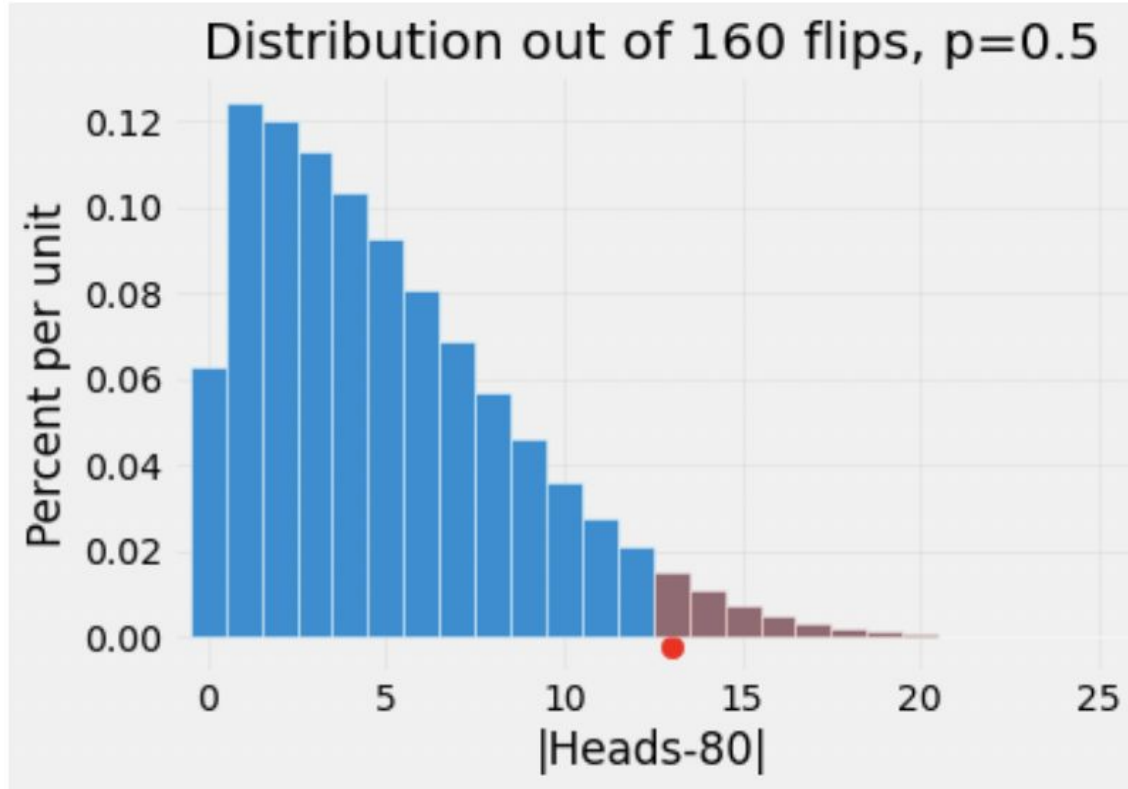
Suppose all coins are fair.

About how many students will conclude that their coins are unfair using this hypothesis test?

- A). 5 B). 25 C). 50 D). 100 E). 160**

(Demo)





Statistic Simulated Under the Null



About 5% of the area is shaded in pink

Can the Conclusion be Wrong?

Yes.

	Null is true	Null is False
Test favors null		
Test rejects null		

The significance level (i.e. **p-value cutoff**) is the probability of rejecting the null when it is actually true.
Choose this value before conducting your test (choose a small value to control this error)

Significance level (p-cutoff value): An Error Probability

- The significance level (i.e. cutoff value that you use for the P -value) is an error probability.
- If:
 - your cutoff is 5%
 - and the null hypothesis happens to be true
- then there is about a 5% chance that your test will INCORRECTLY reject the null hypothesis.

Choosing the P-value Cutoff

"It is convenient to take this point [5%] as a limit in judging whether a deviation is to be considered significant or not." [\[Fisher 1925\]](#)

- Decide on it **before** seeing the results
 - Don't change it!
- Common values at 5% and 1%
 - follow conventions in your area

"If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the author [Fisher] prefers to set a low standard of significance at the 5 percent point ..." [Fisher 1926]



Sir Ronald Aylmer Fisher [1890-1962]
Pioneer of Modern Statistics

- Significance level (i.e. P-value cutoff): You Pick It
 - Does not depend on observed data or simulation
 - “Acceptable” probability of rejecting the null hypothesis when it is true.
- P-value (You Compute It)
 - Depends on the observed data and simulation
 - Probability under the null hypothesis that the test statistic is the observed value or more extreme

Beware of P-Hacking

Ex: Suppose you do 20 different hypothesis tests (testing the relationship between jelly beans and acne) with a null hypothesis that there's no relationship. Assume you conduct each test at a significance level of 0.05.

If in reality *jelly beans aren't actually linked with acne*, what's the probability that NONE of our 20 tests are significant (i.e. that all of our tests correctly don't reject the null)?

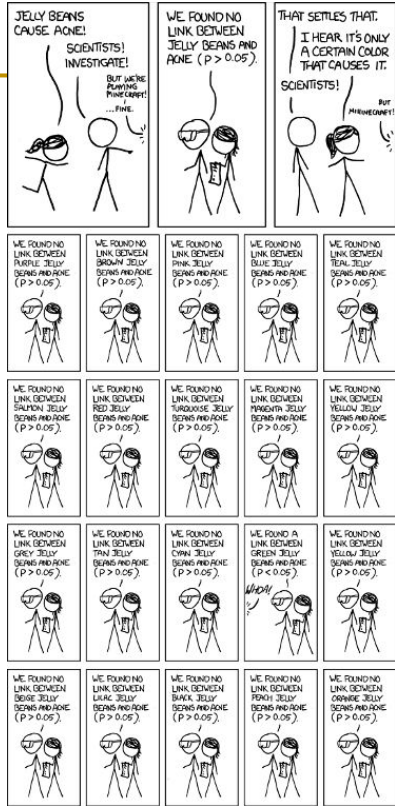
Beware of P-Hacking

Ex: Suppose you do 20 different hypothesis tests (testing the relationship between jelly beans and acne) with a null hypothesis that there's no relationship. Assume you conduct each test at a significance level of 0.05.

If in reality jellybeans aren't actually linked with acne, what's the probability that NONE of our 20 tests are significant (i.e. that all of our tests correctly don't reject the null)?

$$0.95^{20} = 0.3584859224$$

THAT MEANS THAT ABOUT 64% OF THE TIME, ONE OR MORE OF THESE TESTS WILL BE SIGNIFICANT, JUST BY CHANCE, EVEN THOUGH JELLY BEANS HAVE NO EFFECT ON ACNE.



Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

Question: Do fewer people prefer Super Soda than its rival, or is this just chance?

Null hypothesis:

Alternative hypothesis:

Test statistic:

p-value: Start at the observed statistic and look which way?

(Demo)

Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

Question: Do fewer people prefer Super Soda than its rival, or is this just chance?

Null hypothesis: The same proportion of people prefer Super as Rival

Alternative hypothesis:

Test statistic:

p-value: Start at the observed statistic and look which way?

(Demo)

Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

Question: Do fewer people prefer Super Soda than its rival, or is this just chance?

Null hypothesis: The same proportion of people prefer Super as Rival

Alternative hypothesis: A smaller proportion of people prefer Super

Test statistic:

p-value: Start at the observed statistic and look which way?

(Demo)

Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

Question: Do fewer people prefer Super Soda than its rival, or is this just chance?

Null hypothesis: The same proportion of people prefer Super as Rival

Alternative hypothesis: A smaller proportion of people prefer Super

Test statistic: Number of people (out of 200) who prefer Super

p-value: Start at the observed statistic and look which way?

(Demo)

Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

Question: Do fewer people prefer Super Soda than its rival, or is this just chance?

Null hypothesis: The same proportion of people prefer Super as Rival

Alternative hypothesis: A smaller proportion of people prefer Super

Test statistic: Number of people (out of 200) who prefer Super

p-value: Start at the observed statistic and look which way? LEFT

Conduct the test (DEMO)

(Demo)

Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

Question: Do fewer people prefer Super Soda than its rival, or is this just chance?

Null hypothesis: The same proportion of people prefer Super as Rival

Alternative hypothesis: A smaller proportion of people prefer Super

Test statistic: Number of people (out of 200) who prefer Super

p-value: Start at the observed statistic and look which way? LEFT




Conduct the test (DEMO)

What types of errors might result from this hypothesis test and how can we minimize them?

(Demo)

Hypothesis Test Errors: Can the Conclusion be Wrong?

Yes.

	Null is true	Null is False
Test favors null		
Test rejects null		

How do we minimize this type of error?

The significance level (i.e. **p-value cutoff**) is the probability of rejecting the null when it is actually true.
Choose this value before conducting your test (want a small value so that the error from this box is small)

Definition: The **Statistical Power** of a hypothesis test is the probability of correctly rejecting the null when it is false.

Goal: Typical conventions are to set-up your test such that this is at least 80%.

	Null is true	Null is False
Test favors null		
Test rejects null		

The Statistical Power of a test is the probability of correctly rejecting the null when it is false.

The Power of a hypothesis test is the ability for the test to detect a relationship or difference.

Power Depends On:

- 1). Sample Size
- 2). Significance level (the p-value cutoff you chose)
- 3). Effect size (the magnitude of the effect, i.e. the difference from the null).

(Demo)

If aren't using theoretical distribution, then need to decide number of simulations:

- large as possible: empirical distribution \rightarrow true distribution
- No new data needs to be collected (yay!)

Number of observations:

- A larger sample will lead you to reject the null more reliably if the alternative is in fact true (higher “statistical power”).

Difference from the null:

- If truth is similar to the null hypothesis (“small effect size”), then even a large sample may not provide enough evidence to reject the null.

Effect Size vs Statistical Significance:

- **Statistical significance:** After accounting for random sampling error, your sample suggests that a non-zero effect exists in the population.
- **Effect sizes:** The magnitude of the effect. It answers questions about how much or how well the treatment works. Are the relationships strong or weak?

Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude –not just, does a treatment affect people, but how much does it affect them.

-Gene V. Glass¹

The primary product of a research inquiry is one or more measures of effect size, not P values.

-Jacob Cohen²

Statistically Significant vs “Practically” Significant

Not all statistically significant differences are interesting!

Here's how small effect sizes can still produce tiny p-values (i.e. statistically significant results):

- **You have a very large sample size.** As the sample size increases, the hypothesis test gains greater statistical power to detect small effects. With a large enough sample size, the hypothesis test can detect an effect that is so minuscule that it is meaningless in a practical sense.
- **The sample variability is very low.** When your sample data have low variability, hypothesis tests can produce more precise estimates of the population's effect. This precision allows the test to detect tiny effects.

We need a method to determine whether the estimated effect (i.e. the difference between the treatment group and the control group) is still practically significant when you factor in the margin of error from sampling.

Solution: Up Next - Confidence Intervals!