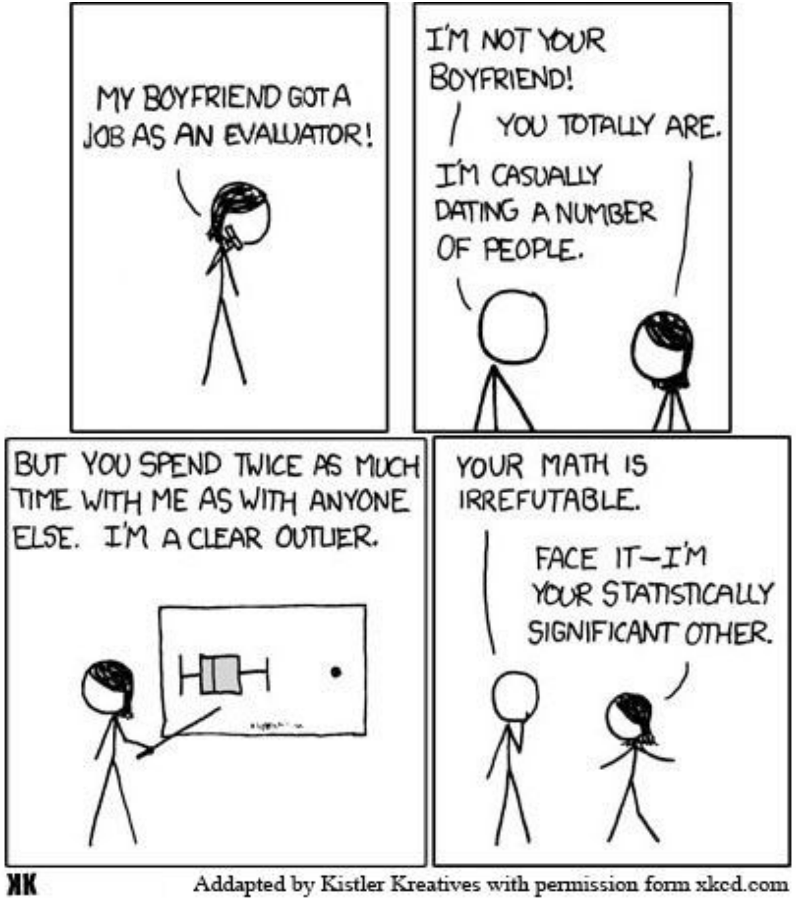
LESSON 20

Hypothesis Testing: Power & Error



Course Logistics: Your 9th Week At A Glance

Mon	Tues	Wed	Thurs	Fri
Attend & Participate in Class		Attend & Participate in Class		Attend & Participate in Class **Quiz 6: Scope: Lessons 15-16; HW 7**
			HW 8 due 11:59pm MT	HW 9 released 8am

Roadmap

Finish Lesson 18: Hypothesis Testing

Comparing a Sample to a Model

Lesson 19: A/B Testing

Video assignment for HW 8

Lesson 20: Hypothesis Tests:

- Significance level
- Power
- Errors

Lesson 20 Learning Objectives:

- **Define the significance level and explain what it is used for.**
- **State the mathematical definition of statistical power.**
- **State 3 factors that influence the power of a hypothesis test.**
- **Explain the difference between Type I and Type II errors and how to minimize them.**
- **Define p-hacking.**

Lesson 20:

- Hypothesis Tests:
  - Significance level
  - Power
  - Errors
- Supplemental Materials
  - More practice problems

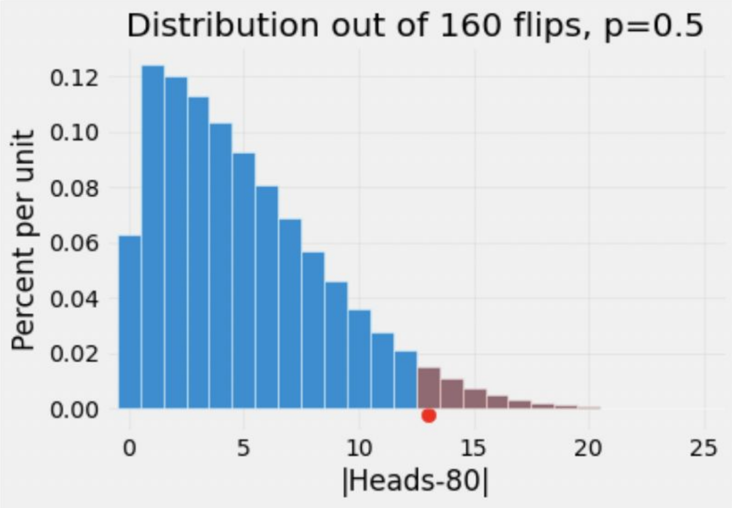
Errors in Hypothesis Testing

- Hypothesis Test Errors
- P-hacking

Recall: Significance Level as an Error Probability

- If:
  - your significance level (i.e. p-value cutoff) is 5%
  - and the null hypothesis happens to be true

- Then there is a 5% chance that the test will INCORRECTLY reject the null hypothesis.



Thus, the significance level is actually a conditional probability of making one type of error:

- **Significance level = P(reject null | null hypothesis is true)**

When null is true, 5% of the time you will get an observed test statistic in tail shaded pink even when the coin is fair JUST BY CHANCE!

Hypothesis Test Errors

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 fail to reject the null)?

Poll:

A). ~95%

B) ~50%

C). ~36%

D). ~5%

E). ~20%

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)?

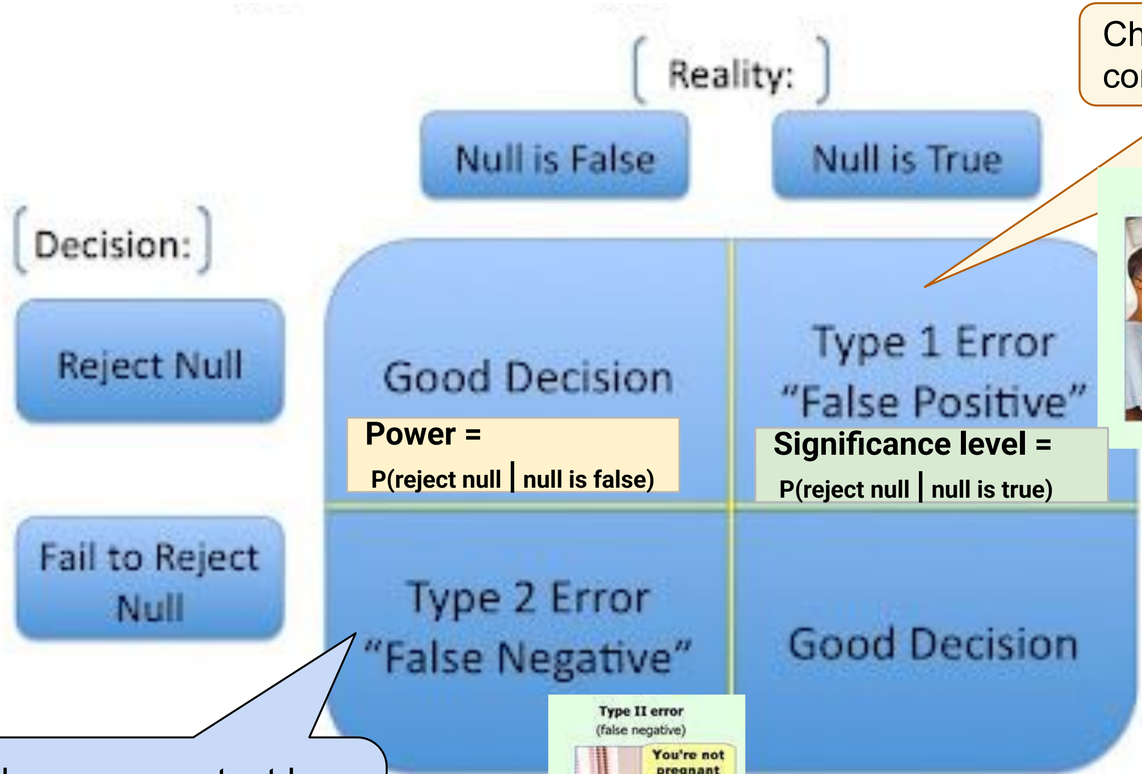
$$0.95^{20} = 0.3584859224$$

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.

"p-hacking," occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant

Can the Conclusion be Wrong? Yes.

Choose a **small significance level** to control this error

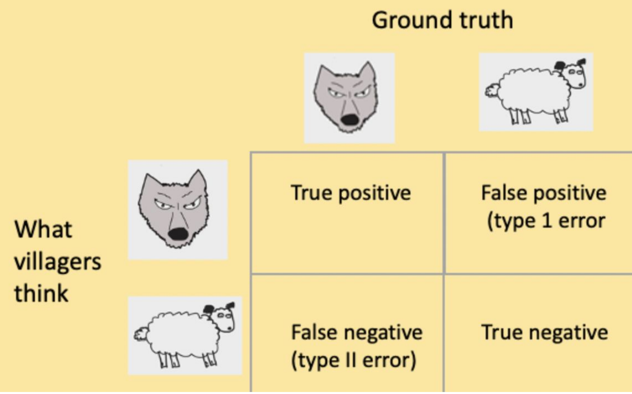


Power = P(reject null | null is false)

Significance level = P(reject null | null is true)

Make sure your test has a large enough **power** to control this error

Null: It's a sheep



		Ground truth: wolf	Ground truth: sheep
What villagers think	wolf	True positive	False positive (type 1 error
	sheep	False negative (type II error)	True negative

Learning Objectives:
- State the mathematical definition of statistical power.
- State 3 factors that influence the power of a hypothesis test.

Brief Intro to Statistical Power

- Statistical Power

Back to our example

Suppose we select 5000 students. We give each student 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: | num of heads - 80 |
- Significance level (cutoff for the P-value): 5%

Suppose in reality all the coins are UNFAIR, with P(H) = 45%

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

A). 50 B). 250 C). 500 D). 1200 E). 1600

Demo

Power

Power

- **Definition:** The statistical **power** of a hypothesis test is the probability of correctly rejecting the null hypothesis when the null is false, that is:
    - **P(reject null hypothesis | null is false)**

For calculating power or required sample size, there are four moving parts:

1). Sample Size

2). Significance level (the p-value cutoff you chose)

3). Effect size (the minimal size of the effect you hope to be able to detect in a statistical test, such as a 5% difference in probability of heads or a 20% improvement in click rates on a website).

4). Power

Specify any 3 of the above and the 4th is completely determined.

Convention: We usually try to design hypothesis tests so the Power is at least 80%.

Most commonly, you would want to calculate sample size, so you must specify the other three.

Hypothesis Tests Caveats/Concerns

- Hypothesis Test Caveats
- Effect size vs significance

Hypothesis Test Concerns

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?
  - large as possible: empirical distribution → true distribution
- **The data you collected:**
  Did you happen to collect a sample that is similar to the population?
  - A larger sample will lead you to reject the null more reliably if the alternative is in fact true (higher "statistical power").
- **The truth:**
  If the alternative hypothesis is true, how extreme is the difference (i.e. what is the effect size)?
  - 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.

Can the Conclusion be Wrong?

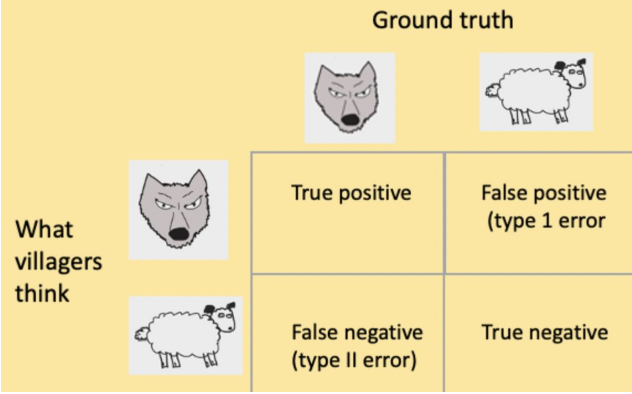
Yes.

Decision: \ Reality:	Null is False	Null is True
Reject Null	Good Decision **Power =** $P(\text{reject null} \mid \text{null is false})$	Type 1 Error “False Positive” **Significance level =** $P(\text{reject null} \mid \text{null is true})$
Fail to Reject Null	Type 2 Error “False Negative”	Good Decision

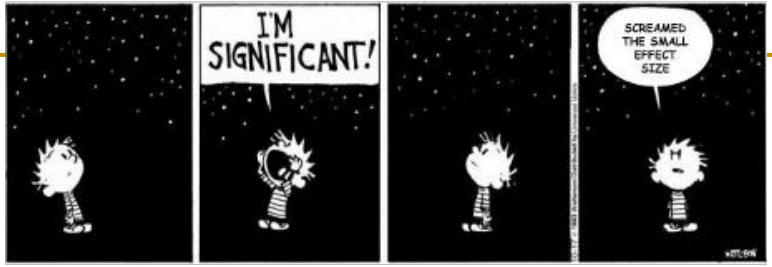
Choose a **small significance level** to control this error

Make sure your test has a large enough **power** to control this error

Null: It’s a sheep



Statistically Significant vs "Practically" Significant



Effect Size vs **Statistical Significance:**

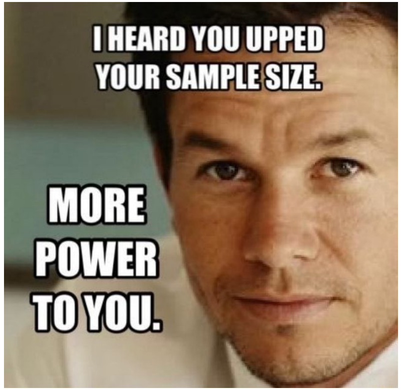
- **Statistical significance**: After accounting for random sampling error, your sample suggests that a non-zero effect exists in the population.

- **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?

  No statistical test can tell you whether the effect is large enough to be important in your field of study. Instead, you need to apply your subject area knowledge and expertise to determine whether the effect is big enough to be meaningful in the real world. In other words, is it large enough to care about?

Statistically Significant vs “Practically” Significant

Not all statistically significant differences are interesting!



- Here’s how small effect sizes can still produce tiny p-values:
    - 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!

Supplemental Materials: Practice Problems

Example:

- 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?

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?

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?

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?

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 **What is the result?**

What types of errors might result from this hypothesis test and how can we minimize them? (Soln: See Juptyer Demo Lesson 21)