PSY 503: Foundations of Statistical Methods in Psychological Science

Introduction to Statistical Inference (hypothesis testing, p-values, sampling distributions)

Suyog Chandramouli

311 PSH (Princeton University)

20th October, 2025

Descriptive Statistics vs Inferential Statistics

- Descriptive statistics:
  - Tools for summarizing and describing the data we have
    - the shape, center, and variance of sample data
    - Means, standard deviations, correlations..

Examples:
- Mean height of students in class: 170cm
- Median income of survey respondents: $45,000
- Standard deviation of test scores: 15 points

- Inferential statistics:
  - Tools for making conclusions about the populations from limited sample data
    - Account for uncertainty in our conclusions
    - Helps with decision-making

Examples:
- Estimating average height of all students from sample
- Testing if new treatment is better than existing one
- Determining if two groups differ significantly

Key differences (Descriptive & Inferential stats)

- Sample Data: Test scores

- Descriptive
  - Purpose
  - Mean score = 75
  - Standard deviation = 10
  - Frequency Distribution is normal

- Inferential
  - Population mean likely between 72-78
  - Scores this year are higher than last year
  - Can generalize to similar students

Why Do We Need Inferential Statistics?

- **The Population Problem**
  - Usually impossible or impractical to measure entire population *(Population parameters: μ, σ, ρ)*
  - Must rely on samples to draw conclusions about the population *(Sample statistics: $\bar{x}$, s, r)*
    - However,
      - Different samples from the same population give different results
      - We need to account for this uncertainty
- **Example: Political Polling**
  - **Population**: All registered voters (millions)
  - **Sample**: 1,000 randomly selected voters
  - **Question**: How can we use this sample to predict election outcome?

Two Main Tasks in Inference

- **Estimation**
  - How large is the effect?
  - What's the population parameter?
  - How precise is our estimate?

- **Hypothesis Testing**
  - Is there an effect?
  - Are groups different?
  - Did intervention work?

Two Main Tasks in Inference

- **Estimation**
  - How large is the effect?
  - What's the population parameter?
  - How precise is our estimate?

- **Hypothesis Testing**
  - Is there an effect?
  - Are groups different?
  - Did intervention work?

Hyptothesis Testing

What are hypotheses?

Hypotheses

- **Testable** claims about the world (or a population)
  - Must be possible to gather evidence for/against the claim
  - These claims are often about a population, but **inferred** based on a sample from the population
    - Inferential statistics
  - Hypotheses aren't true by definition
- Hypotheses guide investigation
  - Experimental design
  - Analysis approach
- Hypotheses can be evaluated
  - Through "converging lines evidence"
  - Through statistical tests..

- Desiridata
  - Falsifiable
  - Precise hypotheses > Vague hypotheses
  - Meaningful

Two types / levels of hypotheses

Research Hypotheses

- About theoretical constructs and relationships
- Express scientific claims about phenomena
- Guide research design

Examples:

- "Social media use affects student productivity"
- "Anxiety influences memory performance"
- "Practice improves skill acquisition"

Statistical Hypotheses

- About data-generating processes
- Express mathematical relationships between parameters
- Guide statistical analysis

Examples:

- H_0: $\mu_1 = \mu_2$ (null hypothesis)
- H_1: $\mu_1 \neq \mu_2$ (alternative hypothesis)
- H_0: $\rho = 0$ (no correlation)

These levels are connected

- **Research Hypothesis**
  - A testable scientific claim
  - About psychological/real-world constructs
  - Example: "ESP exists”
- **Statistical Hypothesis**
  - Mathematical statement about parameters
  - About data-generating mechanisms
  - Example: "$\theta = 0.5$" (probability of correct guesses)

Discuss: Research vs Statistical Hypotheses

- Let’s take the following research hypothesis:
  - "Social media use decreases student productivity"

- Convert this to statistical hypotheses

Discuss: Research vs Statistical Hypotheses

- Let’s take the following research hypothesis:
  - "Social media use decreases student productivity"

- Convert this to statistical hypotheses
  Also, think about methods learned in class already

Hypotheses in experiments

- Let's say you run an experiment: You measure two samples in each of two experimental conditions, and you find a difference between the sample means.

- Questions answered by the hypothesis test:
  - Did the experimental manipulation cause the difference between the sample means?
  - Could random chance (sampling error) have caused the difference?

Statistical Hyptothesis Testing :

Null hypotheses,

Alternate Hypotheses

What is Hypothesis Testing?

- Usage context: “The researcher has some theory about the world and wants to determine whether or not the data actually support that theory.”

- A method to evaluate claims about populations
- Involves comparing two competing hypotheses:
  - Null hypothesis (H_0)
  - Alternative hypothesis (H_1)
- Based on sample data

Ingredients of a hypothesis test

- **Ingredients of a hypothesis test**
  - Research hypothesis mapped onto a statistical hypothesis
  - A second statistical hypothesis that is the oppositie of what we want to believe (Null hypotheses)!
  - Data
  - A decision-making criterion

- **Key idea 1:** Data observed produced can be produced by either hypotheses:
  a) Due to the research hypothesis
  b) Due to the null hypothesis + sampling error + randomness/noise in the world

  In this binary setting, we want are trying to find if there is more evidence for (a) or for (b).

Null hypotheses & alternative hypotheses

- Discuss:
  - For the earlier statistical hypotheses you developed, what is the equivalent null?

Null hypotheses & alternative hypotheses

- **Key idea 1:** Data can be produced even if the true state of the world is one of the other
  a) Due to the research hypothesis
  b) Due to the null hypothesis + sampling error + randomness/noise in the world

  In this binary world, we want are trying to find if there is more evidence for (a) or for (b)

- **Key idea 2:** "Null hypothesis" significance testing is centered around the null hypothesis, not your research hypothesis

  In this binary world, rejection of the null is considered "evidence" for the alternate hypothesis!

Null hypotheses & alternative hypotheses

- **Key idea 1:** Data can be produced even if the true state of the world is one of the other
  a) Due to the research hypothesis
  b) Due to the null hypothesis + sampling error + randomness/noise in the world

  In this binary world, we want are trying to find if there is more evidence for (a) or for (b)

- **Key idea 2:** "Null hypothesis" significance testing is centered around the null hypothesis, not your research hypothesis

  In this binary world, rejection of the null is considered "evidence" for the alternate hypothesis!

- **Key idea 3:** There is always the possibility of an error in inference (given Key idea 1)

Errors are Inevitable

- Why?
  - All data contains noise
  - Samples are not a perfect representation of population
  - The world is complex and messy. All models are wrong.
  - Our experiments rarely control for all confounds

- Philosophical Point
  - Cannot achieve absolute certainty
  - Must work with probabilities
  - Need systematic approach to errors

When Decisions Go Wrong

- **Scenario: Testing New Medicine**
  - **Truth**: Medicine is ineffective (H_0 true)
  - **Our Data**: Shows improvement
  - **Decision**: Approve medicine
  - **Result**: Type I Error

- **Scenario: Testing New Medicine**
  - **Truth**: Medicine works (H_0 false)
  - **Our Data**: Shows no effect
  - **Decision**: Reject medicine
  - **Result**: Type II Error

Types of errors

- **Type I Error (α)**
  - Rejecting true null hypothesis
  - "False Positive"
  - Like convicting innocent person

- **Type II Error (β)**
  - Failing to reject false null hypothesis
  - "False Negative"
  - Like acquitting guilty person

Types of Errors

Assume H_0 is a hypothesis

	Retain H_0	**Reject H_0**
H_0 is true	Correct decision	Type I Error
H_0 is false	Type II Error	Correct decision

Types of Errors

False positive =
Assuming H_0 is false when it is true.

Assume H_0 is the “Null hypothesis”

	Retain H_0	**Reject H_0**
H_0 is true	Correct decision	Type I Error
H_0 is false	Type II Error	Correct decision

False negative =
Assuming H_0 is true when it is not.

Types of Errors

False **positive** =
Assuming H_0 is false when it is true.

Assume H_0 is the "Null hypothesis"

	Retain H_0	**Reject H_0**
H_0 is true	Correct decision	Type I Error
H_0 is false	Type II Error	Correct decision

False **negative** =
Assuming H_0 is true when it is not.

Positive/negative are in relation to the research hypothesis/ alternate hypothesis

Types of Errors

False **positive** =
Assuming H_0 is false when it is true.

Assume H_0 is the "Null hypothesis"

	Retain H_0	**Reject H_0**
H_0 is true	Correct decision	Type I Error (α)
H_0 is false	Type II Error (β)	Correct decision

False **negative** =
Assuming H_0 is true when it is not.

Positive/negative are in relation to the research hypothesis/ alternate hypothesis

Types of Errors

False **positive** =
Assuming H_0 is false when it is true.

Assume H_0 is the "Null hypothesis"

	Retain H_0	**Reject H_0**
H_0 is true	**1 - α** Probability of correct retention	**α** Type 1 error rate
H_0 is false	β Type 2 error rate	**1 - β** Type 2 error rate

False **negative** =
Assuming H_0 is true when it is not.

Positive/negative are in relation to the research hypothesis/ alternate hypothesis

Types of Errors

False **positive** =
Assuming H_0 is false when it is true.

Assume H_0 is the "Null hypothesis"

	Retain H_0	**Reject H_0**
H_0 is true	**1 - α** Probability of correct retention	**α** Type 1 error rate
H_0 is false	β Type 2 error rate	**1 - β** Type 2 error rate

False **negative** =
Assuming H_0 is true when it is not.

Positive/negative are in relation to the research hypothesis/ alternate hypothesis

Types of Errors

False **positive** =
Assuming H_0 is false when it is true.

Assume H_0 is the "Null hypothesis"

	Retain H_0	**Reject H_0**
H_0 is true	Correct decision	Type I Error (α)
H_0 is false	Type II Error (β)	Correct decision

False **negative** =
Assuming H_0 is true when it is not.

Positive/negative are in relation to the research hypothesis/ alternate hypothesis

Types of Errors

False **positive** =
Assuming H_0 is false when it is true.

Assume H_0 is the "Null hypothesis"

	Retain H_0	**Reject H_0**
H_0 is true	**1 - α** Probability of correct retention	**α** Type 1 error rate
H_0 is false	β Type 2 error rate	**1 - β** Type 2 error rate

False **negative** =
Assuming H_0 is true when it is not.

Positive/negative are in relation to the research hypothesis/ alternate hypothesis

Ideally, a powerful hypothesis test has

- Has fewer errors
  - i.e. fewer type-1 and type-2 errors
- Achieved by
  - Fixing the α to be sufficiently low (e.g. 0.05, 0.01, 0.001)
  - Increasing the sample size so that power increases and β decreases
    - ***Intuition:*** *Errors due to sampling noise is reduced if the samples are larger (or closer to the size of the population)*
- **Key idea:**
  - α is fixed (to a low level) by a researcher, is independent of sample size, and stays fixed
  - β is affected by sample size, and decreases with larger n
    - **Is managed rather than guaranteed**
    - We can *estimate* the desired n, based on our desired power and related parameters.

The Null hypothesis

- **Characteristics**
  - States "no effect" or "no difference"
  - What we assume is true until there is sufficient evidence to show otherwise

- **Examples of Null Hypotheses**
  1. New drug has no effect ($\mu_{drug} = \mu_{placebo}$)
  2. No gender wage gap ($\mu_{male} = \mu_{female}$)
  3. No correlation between variables ($\rho = 0$)

The alternative hypothesis

- **Types of Alternatives**
  1. **Two-sided**: Something is different

     $H_1: \mu_1 \neq \mu_2$
  2. **One-sided**: Direction specified

     $H_1: \mu_1 > \mu_2$

     $H_1: \mu_1 < \mu_2$

- **Example: Teaching Method**
  - **H_0**: New method = Traditional method
  - **H_1**: New method ≠ Traditional method
    - (Could be better OR worse)

Logic of hypothesis testing

- **Basic Framework:**
  1. Make claim about population
  2. Collect sample data
  3. Assess if data supports or contradicts claim
  4. Make decision based on evidence

- **Key Question:**
  - *Could chance alone explain what we observed?*

Analogy from LSR

- Statistical test is like a criminal trial
  - "Presumption of Innocence till proven guilty",
    Establish that the accused commited a crime "beyond reasonable doubt"
    Punishing the innocent considered much worse than failing to punish the guilty.

  ⇔

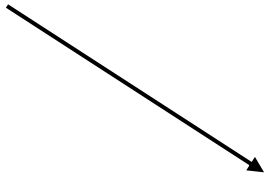
  - "Presumption of Null until shown otherwise"
    Establish that data was not generated by the NULL "beyond reasonable doubt"
    Falsely rejecting the Null much worse than falsely accepting the Null
    (Preventing false claims worse than missing new discoveries)
    - Significance level (α) controls when we accept or reject
      - α = 0.05 is commonly desired
      - for critical questions (e.g. efficacy of vaccines), α may be 0.001

Test statistics & Sampling Distributions

Logic of hypothesis testing

- **Basic Framework:**
  1. Make claim about population
  2. Collect sample data
  3. Assess if data supports or contradicts claim
  4. Make decision based on evidence

- **Key Question:**
  - *Could chance alone explain what we observed?*



Using sampling distributions of some sample statistic.

Sampling distribution

- The probability distribution of a given statistic (e.g., mean) taken from a random sample
  - Distribution of statistics (e.g., mean) that would be produced in an infinite repeated random sampling (with replacement) (in theory)

- **IMPORTANT:** Can be any statistic (proportion, mean, standard deviation, t-statistic, F)

Key points

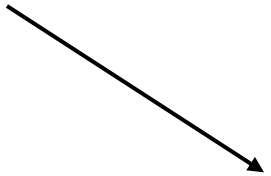
- Two samples from the same population will tend to have somewhat different means
  - Conversely, two different sample means does NOT mean that they come from different populations
  - The variance of the sampling distribution of means gets smaller as the sample size increases
  - Mores samples give better estimate of population mean

Null-hypothesis significance testing

Logic of hypothesis testing

- **Basic Framework:**
  1. Make claim about population
  2. Collect sample data
  3. Assess if data supports or contradicts claim
  4. Make decision based on evidence

- **Key Question:**
  - *Could chance alone explain what we observed?*



Using sampling distributions of some sample statistic.

Instantiating NHST

- Begin by assuming H_0, the opposite of your hypothesis
  - For instance, there is no relationship between X and Y

- 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?

- 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?

Instantiating NHST

- Null Hypothesis H0: There is no significant difference
  - 0 in population (does not have to be this)
- Alternative Hypothesis H1: There is a statistically significant difference
  - Some difference

Two-sided and One-sided Alternative Hypotheses

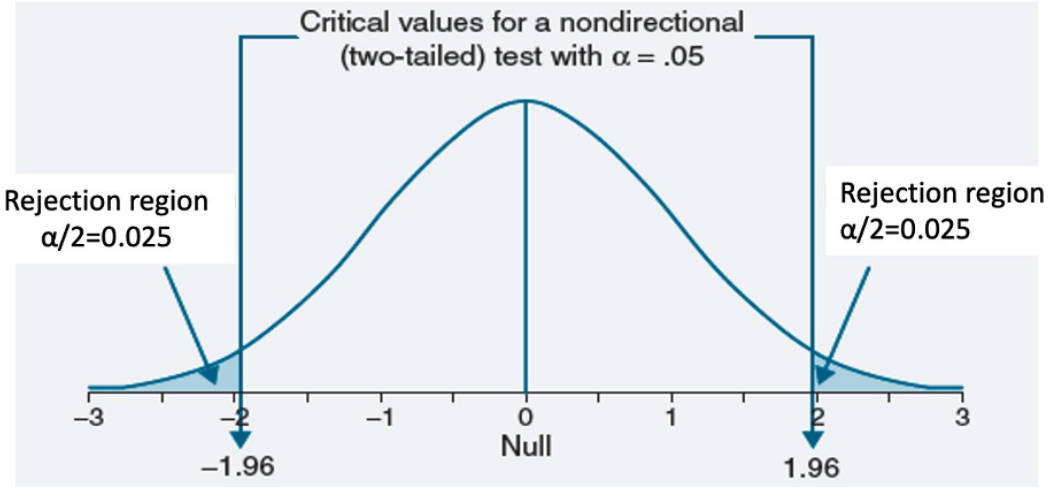
- Two-sided: H_o: $\mu f = \mu d$; $H_1 = \mu f \neq \mu d$
- One-sided: $H_o = \mu f = \mu d$; $H_1 = \mu f > \mu d$
- One-sided: $H_o = \mu f = \mu d$; $H_1 = \mu f < \mu d$
  - Only use a one-sided / directional hypothesis if you have a strong theoretical prediction (for example, from a model) or you preregister it
    - Can gain statistical power
- We can accommodate both two-tailed and one-tailed tests statistically

Define your level of significance (α)

- Level of significance (α): Probability of rejecting the H_o when H_o is true
  - $\alpha = 0.05$
  - Some use others (e.g. 0.01, 0.0000003 particle discovery in high energy physics)

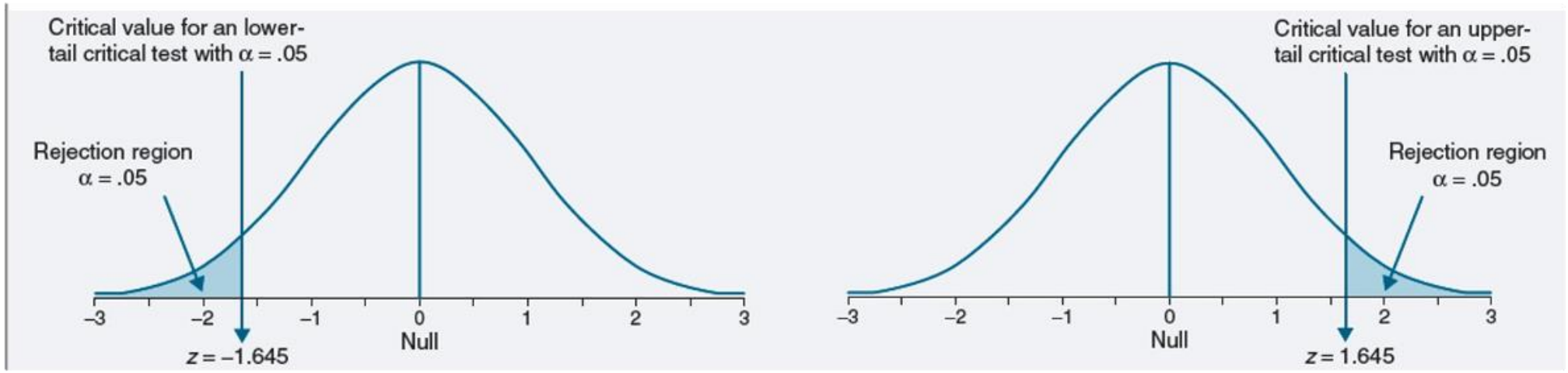
Two-tailed test

- The sum of the tails sums to α (0.025 in each tail for a two-tailed when α = 0.05)
- See where the statistic lies relative to a ‘critical score’ that depends on defined alpha (same procedure used to calculate confidence intervals)
- For the right and left tail, if the test statistic > 1.96 or less than -1.96 (critical value) reject null & accept alternative



2-tails: We are interested in any significant deviations from H_0

One-tailed test



- 0.05 in each tail
- If statistic within rejection region = reject null & accept alternative
- Do you see why you get power with a one-sided / directional hypothesis?

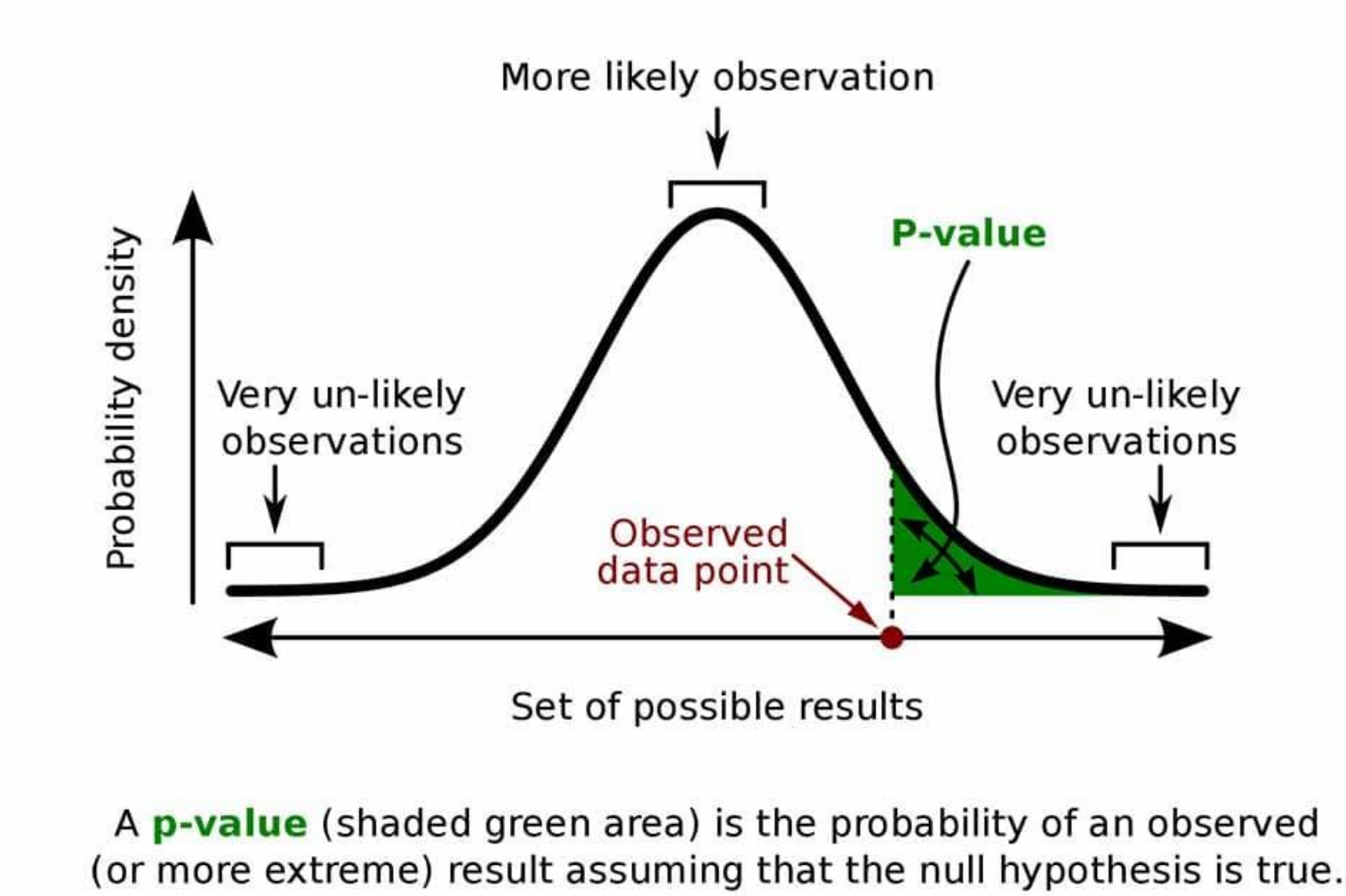
Should I use a one-tailed or two-tailed test?

- Always use two-tailed when there is no directional expectation
  - There are two competing predictions

- Can use one-tailed when there is strong justification for directional predictions

Caution: It's not a good practice to follow up with one-tailed, just because the two-tailed is not statistically significant

P-value



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

What is a p-value?

- The probability of observing the sample data, or more extreme data, *assuming the null hypothesis is true*

$$P(D|H0)$$

- How surprising / unexpected an observation is

p-value Conventions

- Conventions:
  - $p < 0.05$: significant evidence against H0
  - $p > 0.10$: non-significant evidence against H0
  - $0.05 < p < 0.10$: marginally significant evidence against H0

Null & Alternate Hypotheses (Sampling Distribution)

- Classic NHST cares only about sampling distribution of the Null Hypothesis (H0)

- H1 distribution can be useful to
  - See what to expect if the alternative were true
  - Calculate power

Steps for Hypothesis Testing

- State null hypothesis and alternative hypothesis
- Calculate the corresponding test statistic and compare the results against the “critical value”
- State your conclusion

Steps for Hypothesis Testing

- Step 1
  - Convert the research question to null and alternative hypotheses
    - The null hypothesis (H_0) is a claim of "no difference in the population"
      - No difference between one population parameter and another: $H_0 : \mu f = \mu d$
  - We usually want to reject this hypothesis
- The alternative hypothesis (H_1) claims H_0 is false: There is some difference
  - Difference between one population parameter and another: $H_0 : \mu f \neq \mu d$

Steps for Hypothesis Testing

- Step 2
  - Calculate the corresponding test statistic and compare the result against the “critical value”
    - t, z, F
      - Is the test statistic > or < than critical value?
- A value of the test statistic is interesting if it has only a small chance of occurring when the null hypothesis is true

Steps for Hypothesis Testing

- Step 3
  - State your conclusion
    - Reject the null $p < \alpha$
    - Fail to reject the null $p > \alpha$

Correctly reporting and interpreting p-values

- Exact *p*-values (3 decimal places)
- *p*-values reference the observed data and not a theory
- Report α
- Do not use *p*-values as a measure of evidence

Strange NHST language

The Peculiar Phrases

- We never "accept" the null hypothesis
- We only:
  - "Reject H_0"
  - "Fail to reject H_0"

Why "Never Accept"?

- **Philosophical Reason**:
  - Can't prove a hypothesis true
  - Can only gather evidence against it
  - Science progresses by disproving, not proving
- **Statistical Reason**:
  - Infinite possible alternative explanations
  - "Failing to reject" just means:
    - "This null is consistent with our data"
    - "But so might be many other explanations"

Historical Origins

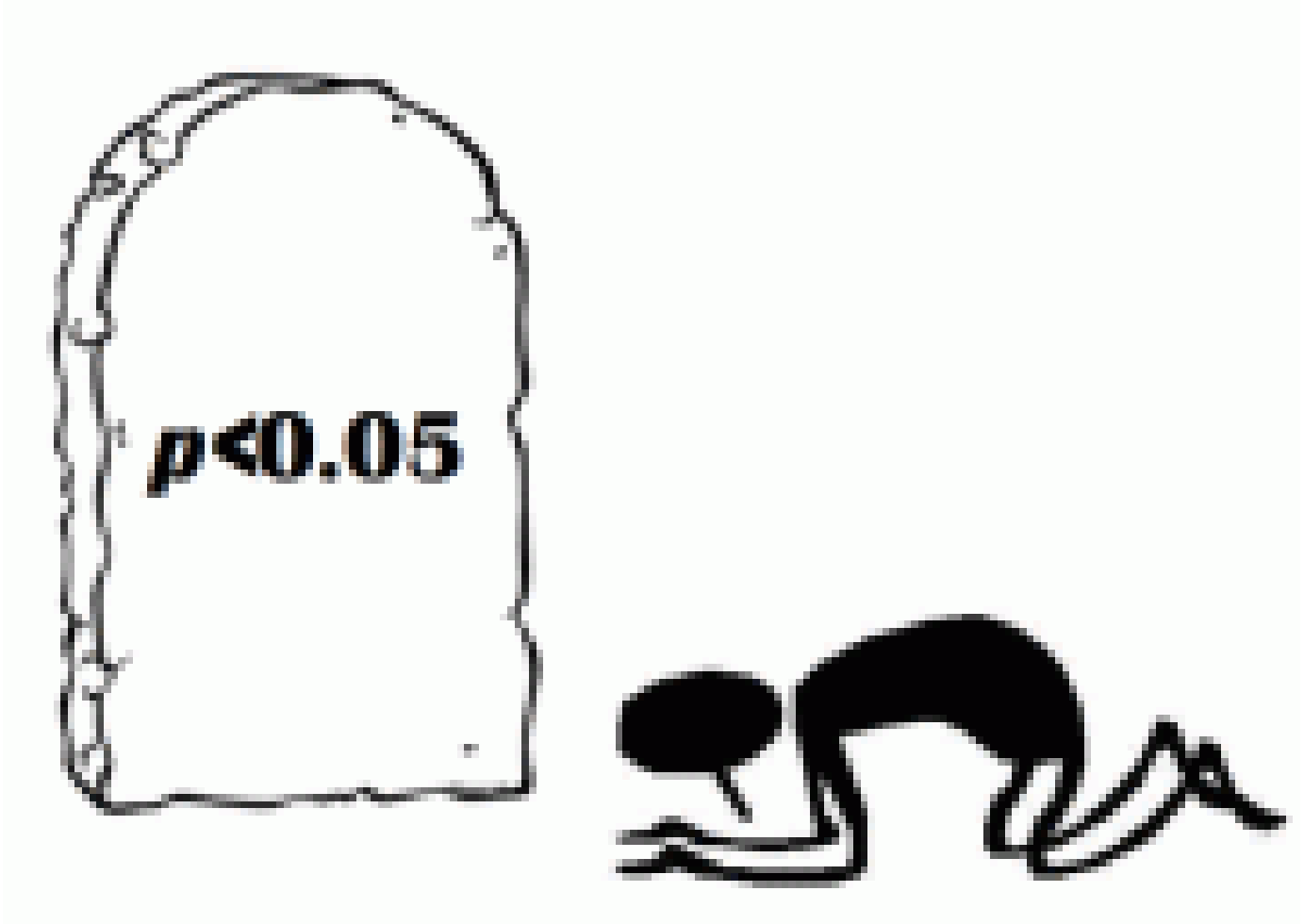
- **Neyman-Pearson Framework**:
  - Hypotheses aren't "true" or "false"
  - We make decisions with controlled error rates based on a treshold
- **Fisher's Influence**:
  - Quantifying evidence against null
  - No formal "acceptance"

Some issues with p-values

Practical issues

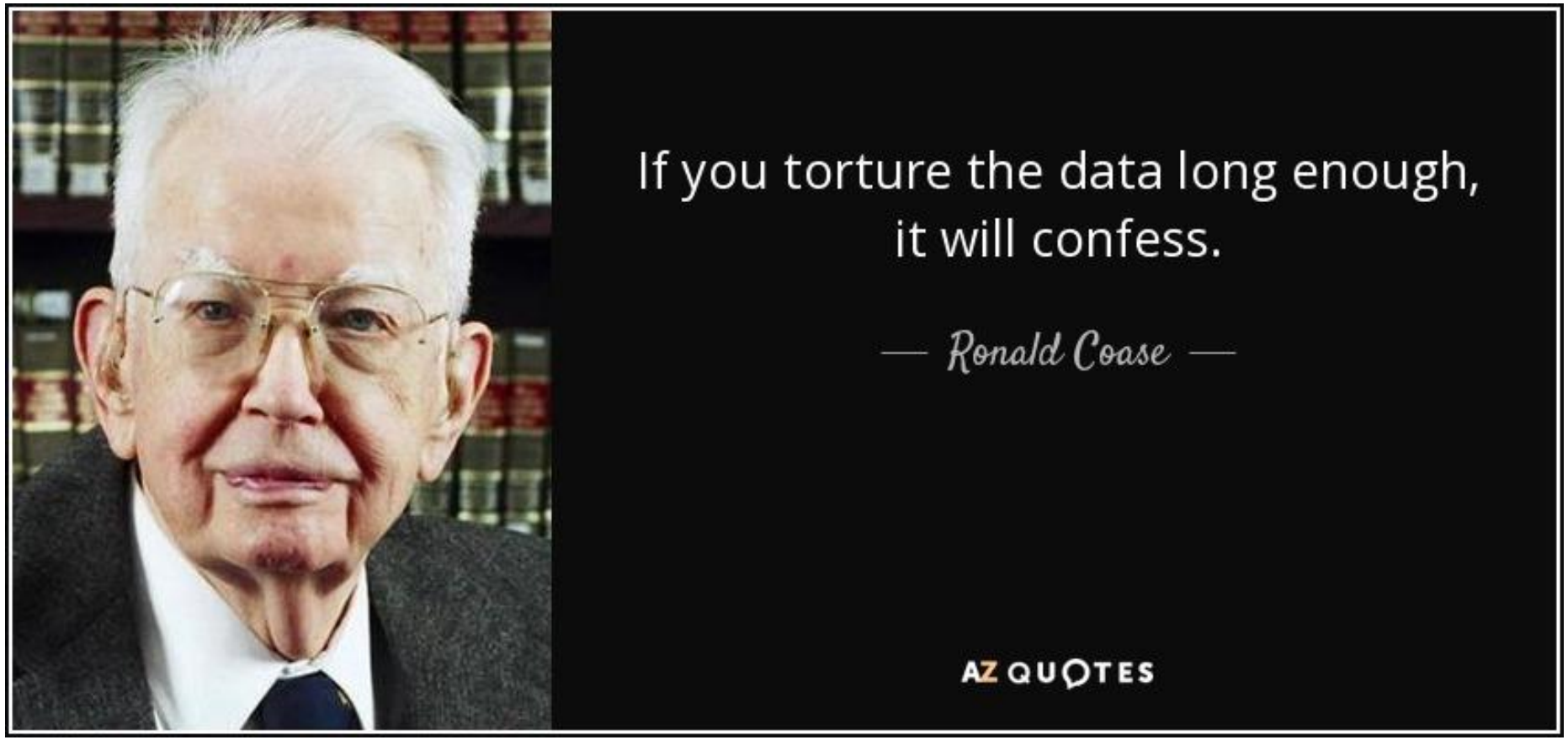
- **Asymmetric Error Treatment**
  - Over-emphasis on Type I errors, with management of Type II errors
  - May not match real-world priorities
  - Ignores relative costs of errors
- **Binary decision-making**
- **Scientific Impact**
  - Missed discoveries
    - Subtle effects, small sample studies, resource intensive research
  - Leads to Publication bias / file-drawer problems
    - Replication issues
  - Conservative science

Worship of p-values



P-hacking

- P-hacking: trying lots of analyses until you get desired outcome



Significant

- https://imgs.xkcd.com/comics/significant.png

Examples of p-hacking

- Stop collecting data once $p < .05$
- Analyze many measures, but report only those with $p < .05$
- Collect and analyze many conditions, but only report those with $p < .05$
- Use covars to get $p < .05$
- Exclude participants to get $p < .05$
- Transform the data to get $p < .05$