Module 3D: Hypothesis Testing

Applied Epistemology: A Framework for how we know things scientifically

Agenda:

1. H_O/H_A: Our model of the test universe (the distribution of the variable)
2. Test & assumptions: are the assumptions met? Is the test valid?
3. Quantitative evidence: **p-value**, or critical value.
   - False positive =Type I (α), False Negative = Type II (β), Type III errors
   - Sensitivity, Specificity, Power → confusion matrix, ROC/AUC curve
   - Positive Predictive Power, Negative Predictive Power
   - Confusion Matrix
   - **ROC/AUC curve**
4. Conclusion & uncertainty/estimation

What is "Statistical Thinking"?

- Understanding complexity via:
  - Understanding Distributions;
  - Models and their assumptions;
  - Quantification of uncertainty;
  - Thinking in probabilities;
  - Utilizing systematic criteria for decision making.
- Retraining our brains to not rely on heuristics/shortcuts and bias.

- Most of the work involved in statistics is clearly stating your hypothesis
    - What is your expectation? Can you quantify it? What is the sampling distribution?

- Hypothesis testing allows you to ask if a parameter ***significantly*** differs from the ***null*** expectation
    - It quantifies how unusual the data are *if you assume that the null hypothesis is true.*

- Hypotheses are about populations but are tested with data from samples
    - Assumes that the sampling is random.
    - (most common inferential statistics are parametric – they assume the sampling distribution follows a normal distribution)

Your pipeline for hypothesis testing in statistics

Step 1

Formulate your null **hypothesis**

- Null hypothesis is ***only*** *hypothesis that is tested*
- Falsification: *want* to reject your null

Step 2

Identify appropriate **test statistic**

- Assumptions of your test

Step 3

Quantify the results of your test

- **P value** or comparison to **critical values**
- How *unusual* is your data?

Step 4

Conclude: reject or fail to reject

- based on alpha value
- if appropriate, confidence interval of the parameter

<u>Hypothesis testing **automates binary** decision making:</u>

1. If p-value < α → Reject null hypothesis

2. If p-value > α → Fail to reject null hypothesis

- We can outline steps that help us make decisions

- **<u>Remember:</u> What is statistically significant is somewhat arbitrary:**

  p-value of 0.04999 is not so different from 0.050001

* α is also called "**significance level**". It is defined by the scientist before the experiment that quantifies acceptable levels of being wrong about the conclusion (usually the cut-off is 1 in 20 or 5% or 0.05).

Step 1: Making and using hypotheses:

The Null Hypothesis (H_0):

A specific statement about a population parameter made for the purpose of the argument. Usually carefully worded so that it can be rejected (falsified).

The Alternate Hypothesis (H_A):

Represents all other possible parameter values except that stated in H_0. It is often what the researcher hopes is true and remains after the Ho has been rejected.

H_0:

- The *only hypothesis actually tested by the data*

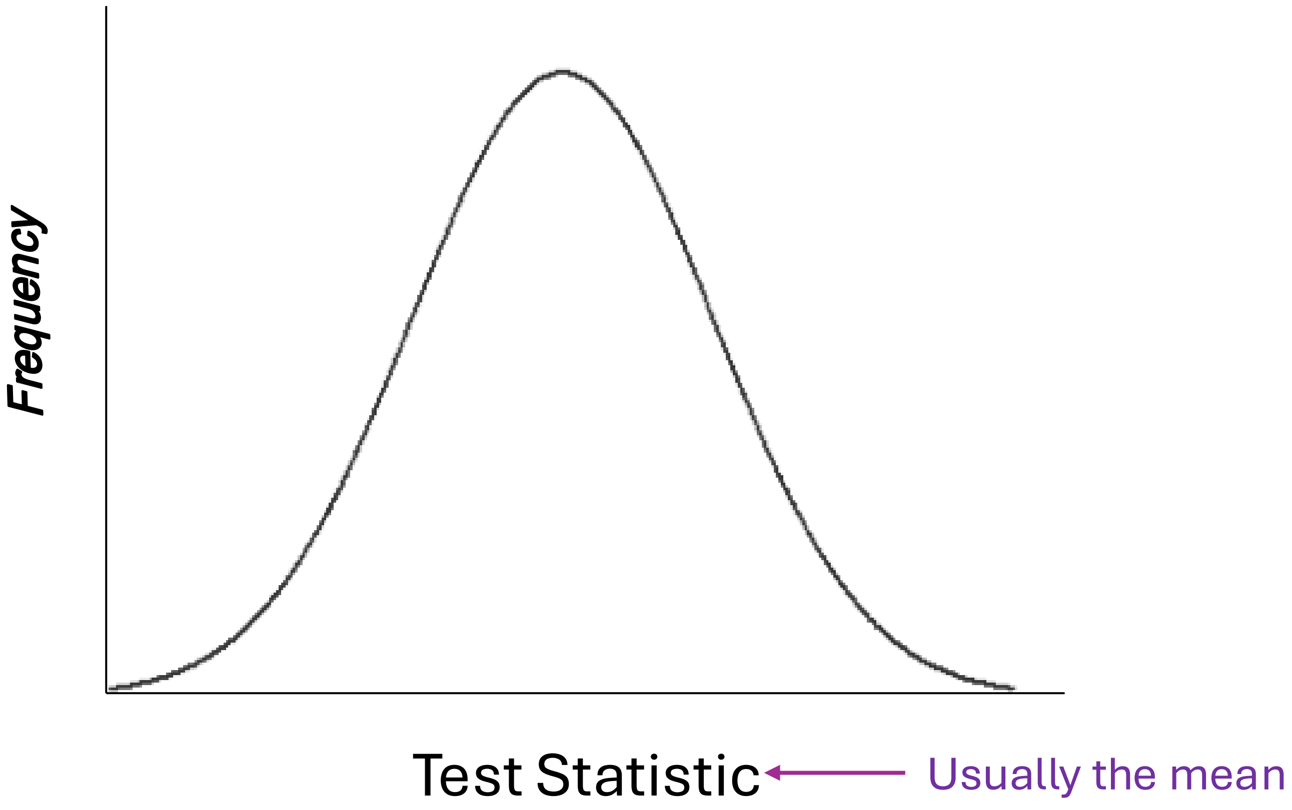
- *Usually, the skeptical POV*
    - *Claims* ***NO difference/effect***
    - *Observations are just due to chance*

- *Reject or Fail-To-Reject BUT* ***NEVER EVER*** *accept*

- *Rejecting* H_0 *reveals nothing about the magnitude of a parameter*

H_A:

- Usually, the statement that the researchers *hope* is true


Frequency
Test Statistic
Usually the mean

Step 2: Identify a Test Statistic:

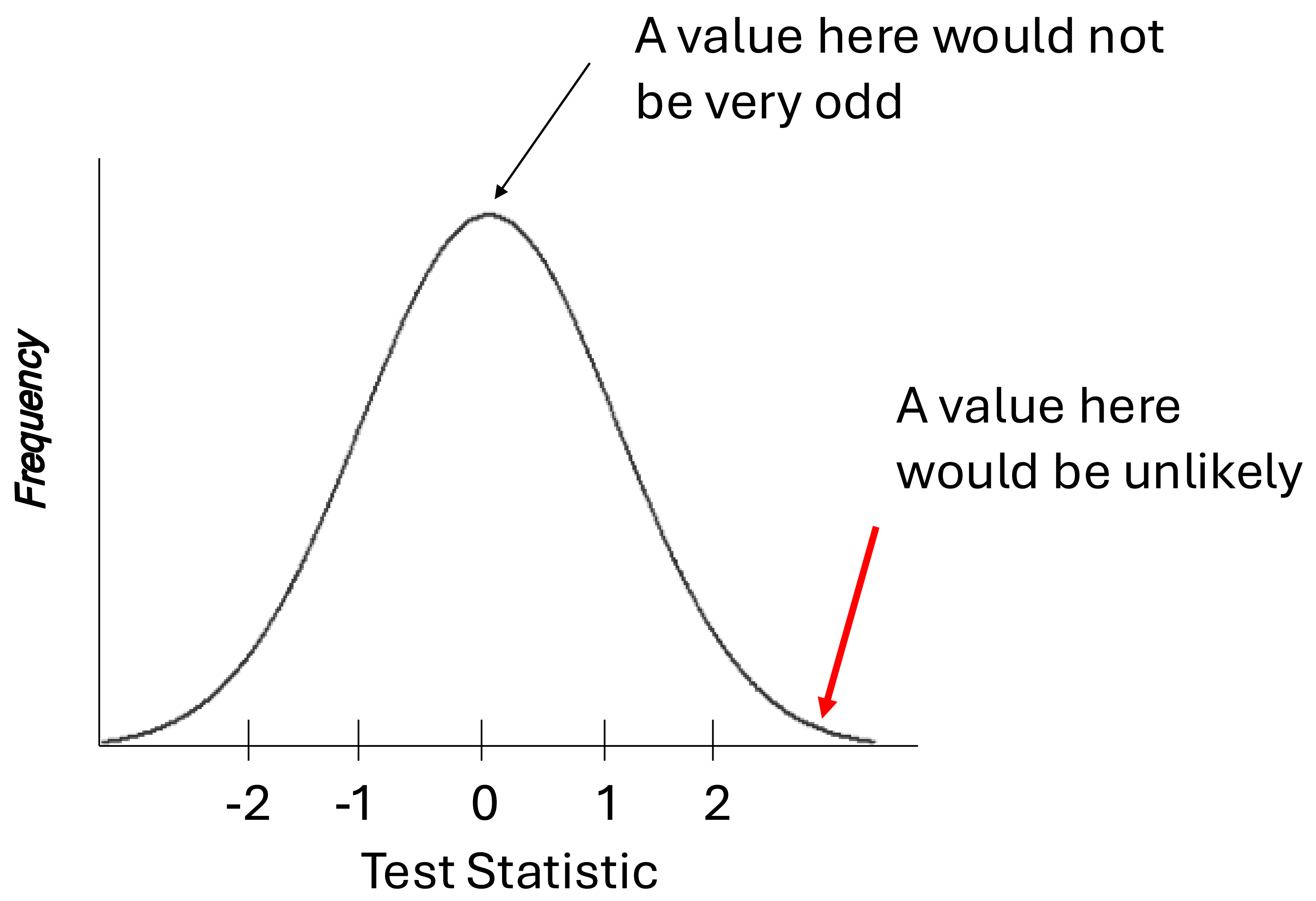
Quantity calculated from the data that is used to evaluate how compatible the results are with those expected the null hypothesis.

- How ‘weird’ are your results?
- Do your data support the assumptions of your test statistic?

Null Sampling Distribution:

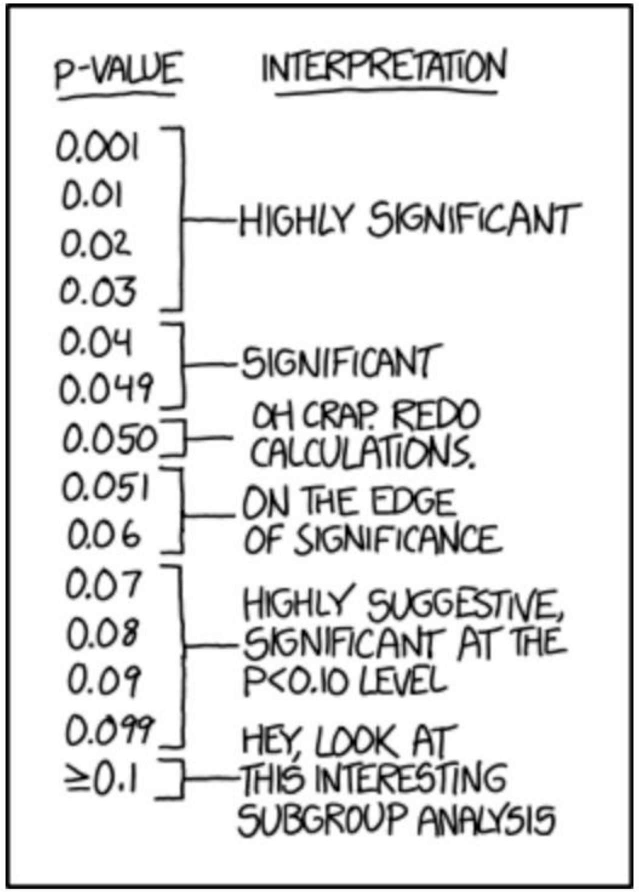
Probability of the test statistic assuming the null hypothesis

- Usually assume Normal Distribution (for means, we can usually rely on CTL!)
- Null distribution can be acquired via computer simulations/modeling


A value here would not be very odd
A value here would be unlikely
Frequency
-2
-1
0
1
2
Test Statistic

P-Value:

Probability of obtaining data that are <u>equal to or even more extreme</u> than the value assuming the null hypothesis is true



http://xkcd.com/1478/

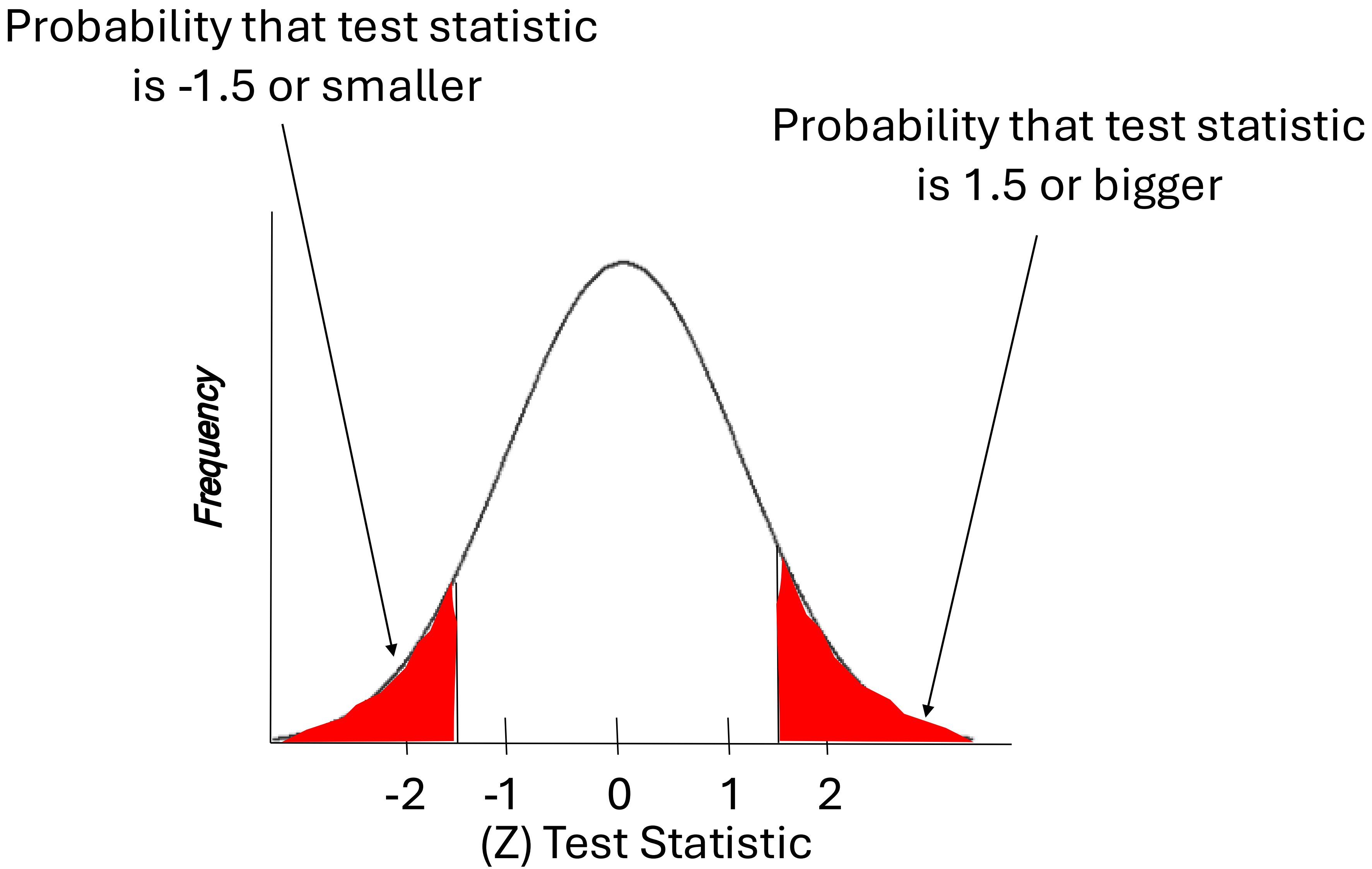
<u>How are P-values found?</u>

-Parametric tests: calculated in R or Python or use cut-off values in published tables.

-Re-sampling methods: permutation

-Simulation

P-value


Probability that test statistic
is -1.5 or smaller
Probability that test statistic
is 1.5 or bigger
Frequency
-2
-1
0
1
2
(Z) Test Statistic

How do you use a P-value?

In hypothesis testing you can do one of two things:

Reject or Fail-to-Reject H_0

Statistical Significance:

α is used as the basis for rejecting the null hypothesis (α is set by the experimenter; p-values are calculated from the sample)

If p-value $\leq \alpha$, H_0 Rejected

If p-value $> \alpha$, FTR H_0

* α is often 0.05

Hacking p-values: getting the p-value you need to publish your results

- Even well intentioned, honest researchers can accidentally "p-hack"
  - Stopping the study when p-value is significant (**n** individuals) but continuing other studies with more **n** when p-value isn't yet significant (so you end up with a bias towards studies that have greater **n** and so are more likely to pick up smaller differences)
  - Play with outliers (include or exclude) until a significant p-value is achieved.
  - https://www.nature.com/articles/d41586-025-01246-1

What are possible alternatives to P-values?

https://royalsocietypublishing.org/doi/10.1098/rsbl.2019.0174