Hypothesis Testing

Presented by David John Baker
August 2019

FLATIRON SCHOOL

CI Review

- What is a confidence interval?
- Why would we use it?
- What parameters will affect your a confidence interval?

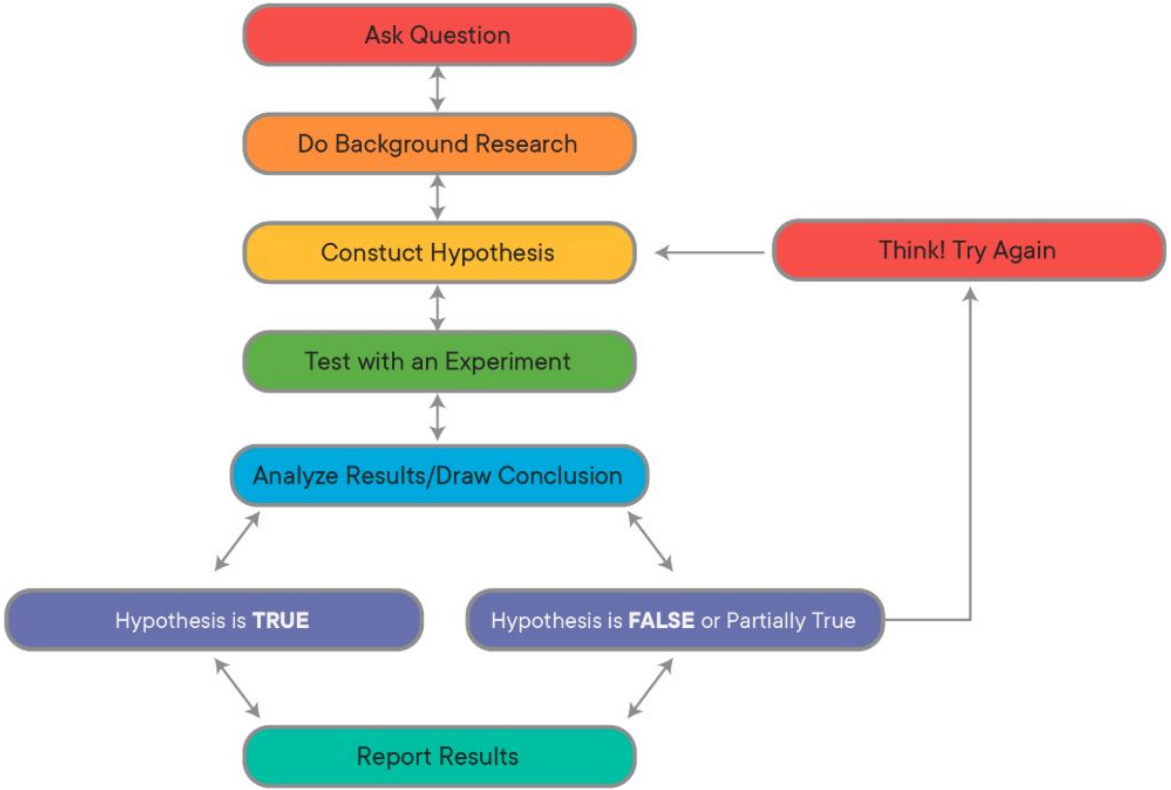
Hypothesis Testing

- Knowing stuff about the world is hard
- The Scientific Method is there to help us out
- Today we make our first pass at understanding one way of thinking about how we can do science
- Dave's Opinion: Time invested in thinking about how we know what we know / philosophy of science will develop your critical thinking skills better than any other investment.

Lesson goals

- Scientific Method (Theory vs Practice)
- The Problem of Induction
- Popper, Falsifiability, Demarcation
- Logic of Null Hypothesis Significance Testing
- Four Types of Outcomes in NHST
- Introduction to p values
- Run through a single statistical test

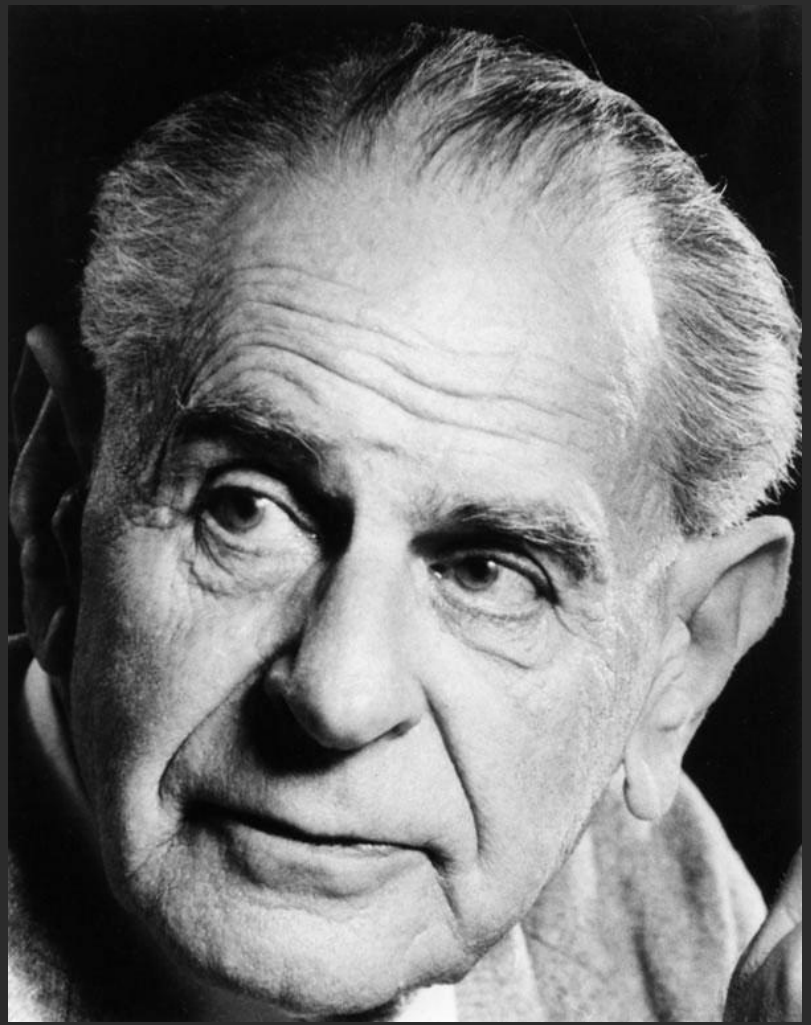
The Scientific Method



Karl Popper

- Problem of demarcation
- Falsifiability
- Problem of Induction

What does it mean for a theory to be falsifiable? In your groups, come up with both an example of a falsifiable and non falsifiable claim.



Deduction vs Induction

All swans are white.
Roger is a swam.

Roger is white.

——————————————————

Melvin the swan in white.
Gary the swan is white.
Mary the swan is white.
Terry the swan is white.
Cherry the swan is white.

All swans are white (?)

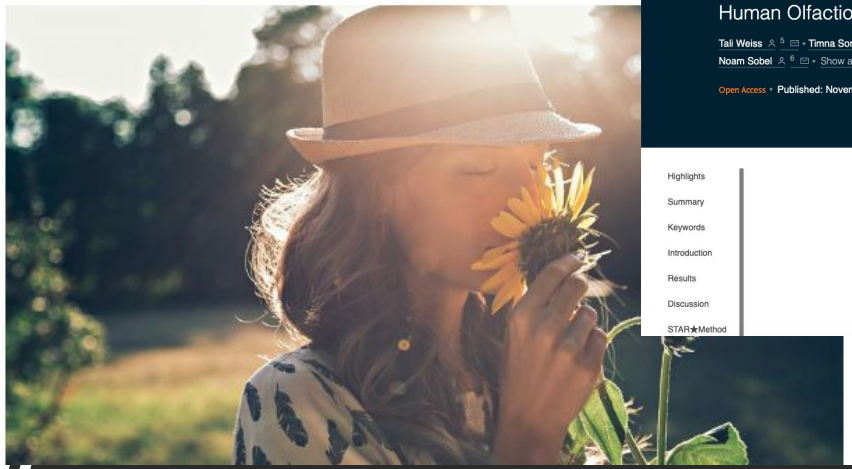
//

Problem of Induction

Women Missing Brain's Olfactory Bulb Can Still Smell, Puzzling Scientists

By Yasemin Saplakoglu - Staff Writer 13 hours ago Health

Researchers have discovered a small group of people that seem to defy medical science.



Neuron

Log in Register Subscribe Claim

CASE STUDY | ONLINE NOW

Human Olfaction without Apparent Olfactory Bulbs

Tali Weiss • Timna Soroka • Lior Gorodisky • ... Edna Furman-Haran • Thijs Dhollander • Noam Sobel • Show all authors • Show footnotes

Open Access • Published: November 06, 2019 • DOI: https://doi.org/10.1016/j.neuron.2019.10.006

PDF Figures Save Share Reprints Request

PlumX Metrics

Highlights
Summary
Keywords
Introduction
Results
Discussion
STAR★Method

Highlights

- Humans can have normal olfaction without apparent olfactory bulbs
- Olfaction without apparent bulbs is seen in 0.6% of women, but not in men
- Olfaction without apparent bulbs is associated with left-handedness

Summary

Problem of Induction

Formulation of the problem [edit]

In inductive reasoning, one makes a series of observations and infers a new claim based on them. For instance, from a series of observations that a woman walks her dog by the market at 8 am on Monday, it seems valid to infer that next Monday she will do the same, or that, in general, the woman walks her dog by the market every Monday. That next Monday the woman walks by the market merely adds to the series of observations, it does not prove she will walk by the market every Monday. First of all, it is not certain, regardless of the number of observations, that the woman always walks by the market at 8 am on Monday. In fact, David Hume would even argue that we cannot claim it is "more probable", since this still requires the assumption that the past predicts the future.

Second, the observations themselves do not establish the validity of inductive reasoning, except inductively. Bertrand Russell illustrated this point in *The Problems of Philosophy*:

> Domestic animals expect food when they see the person who usually feeds them. We know that all these rather crude expectations of uniformity are liable to be misleading. The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken.

In several publications it is presented as a story about a turkey, fed every morning without fail, who following the laws of induction concludes this will continue, but then his throat is cut on Thanksgiving Day.[3]

Usually inferred from repeated observations: "The sun always rises in the east."

Usually not inferred from repeated observations: "If someone dies, it's never me."

We can’t accumulate evidence for a theory by just adding more data since the addition of one piece of contrary evidence (the appearance of a black swan) has the potential to destroy our theory.

This has happened over and over again, see the history of science.

So how do we get around this problem?

Exploit the asymmetry between getting data to establish a theory and finding data to be critical of it. Enter NHST.

//

Null Hypothesis Significance Testing

- Instead of accumulating evidence FOR a theory. We instead set up TWO competing hypotheses.
- H0: Null Hypothesis: Assumes nothing is happening.
- H1: Alternative Hypothesis: Assumes something is happening.

//

Null Hypothesis Significance Testing Examples

- I am interested in theory that people who have plant based diet will have lower cholesterol than those who eat a mixed diet.
- How do I begin to build support for this theory?
- Can't just go around asking people, "Well I know one guy who eats just meat and HIS cholesterol is just fine" (Induction problem)
- Need to be able to generalize this theory…

//

Null Hypothesis Significance Testing Examples

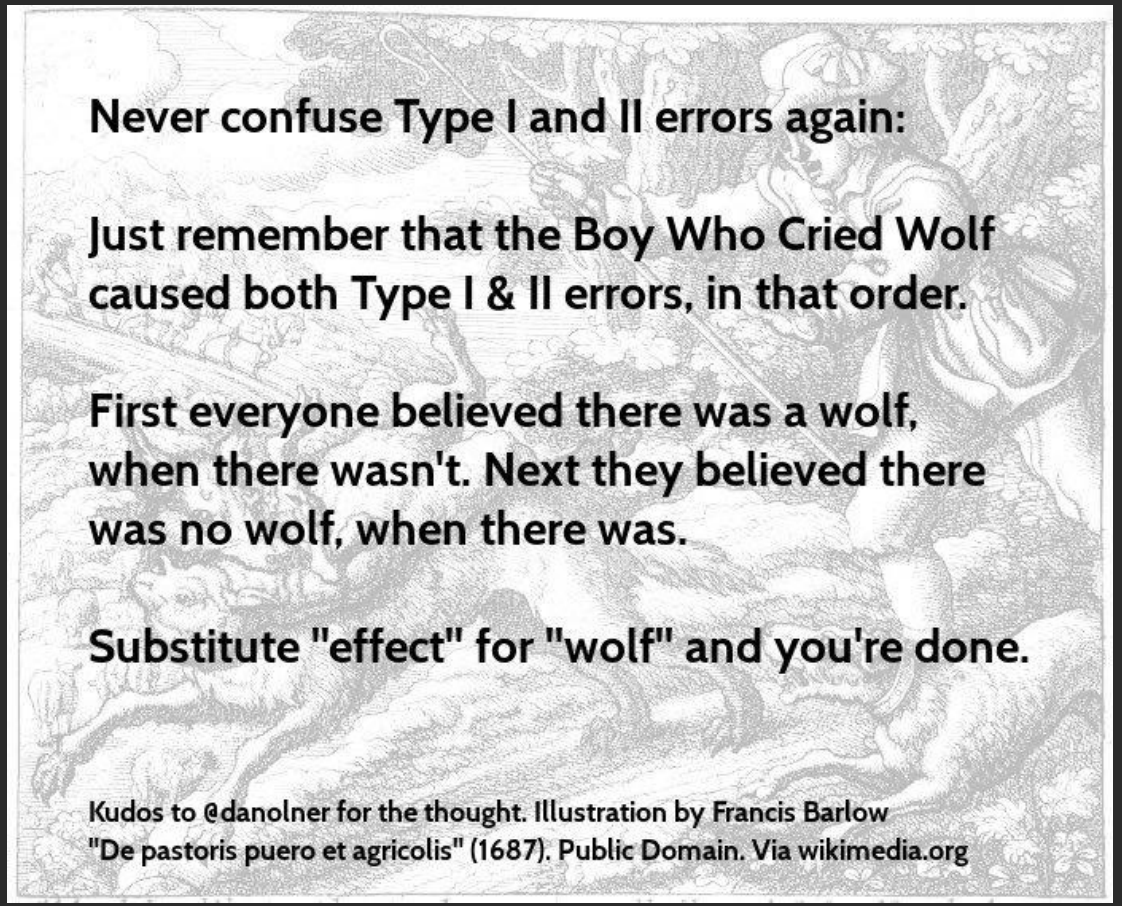
- So instead of building FOR theory, we instead say "I have a **(null) hypothesis** that there is no difference in cholesterol levels between plant only eaters and those who eat a mixed diet."
- If this hypothesis were to be proven wrong, what are we left with?
- A competing **(alternative)** hypothesis noting there IS a difference between these two groups (hopefully in the direction we thought!)

//

Types of Errors

	H0 True	H1 True
Significant Finding	False Positive	True Positive
Non-Significant Finding	True Negative	False Negative

How to Remember Types of Errors



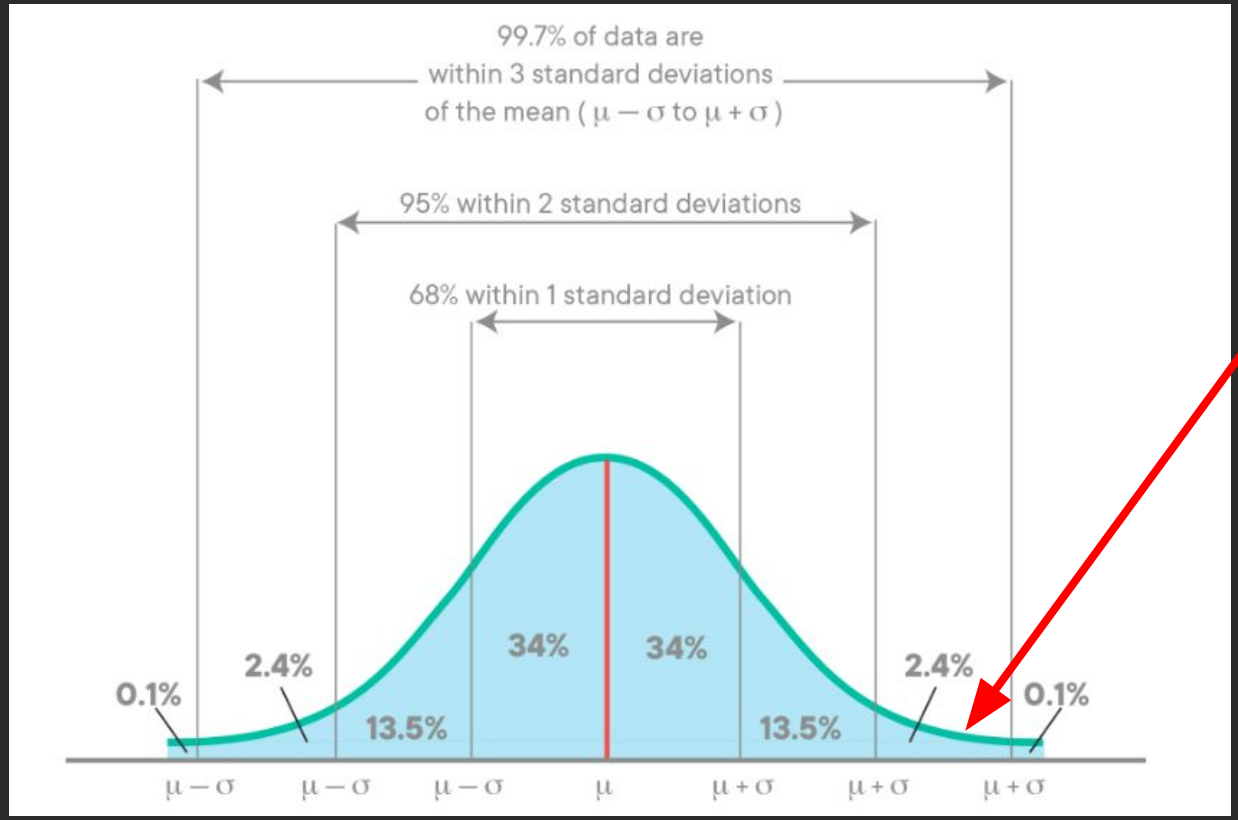
Types of Errors

	H0 True	H1 True
Significant Finding	False Positive	True Positive
Non-Significant Finding	True Negative	False Negative

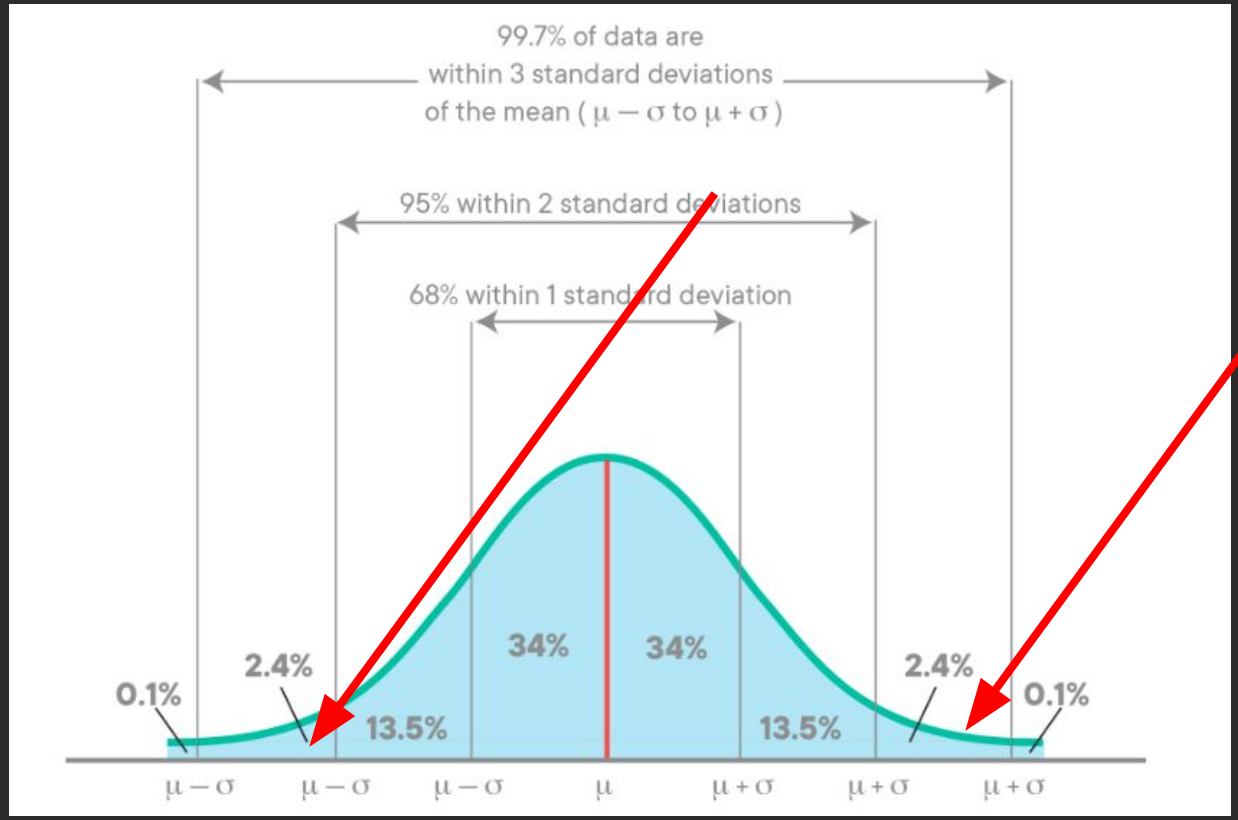
Types of Errors

	H0 True (50%)	H1 True (50%)
Significant Finding $\alpha = 5\%$, $1-\beta=80\%$	False Positive 5%*50%=2.5%	True Positive 80%*50%=40%
Non-Significant Finding $1-\alpha = 95\%$, $\beta=20\%$	True Negative 95%*50%=47.5%	False Negative 20%*50%=10%

Rejection Region (One Tailed)



Rejection Region (Two Tailed)



Calculating a Test Statistic

- The general formula for a test statistic is …

$$\frac{\text{Statistic - Parameter}}{\text{Standard error}}$$

If we know standard deviation, we use z test

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$$\frac{\text{Statistic - Parameter}}{\text{Standard error}}$$

If we don't know standard deviation, we used t test

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} \qquad s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Steps of Hypothesis Testing(ish)

- State Null and Alternative Hypotheses
- Set Your Alpha Level (and pick direction of your test)
- Select Sample and Collect
- Locate Region of Rejection / Critical Values
- Compute Your Test Statistic
- Decide if you reject null hypothesis!

//

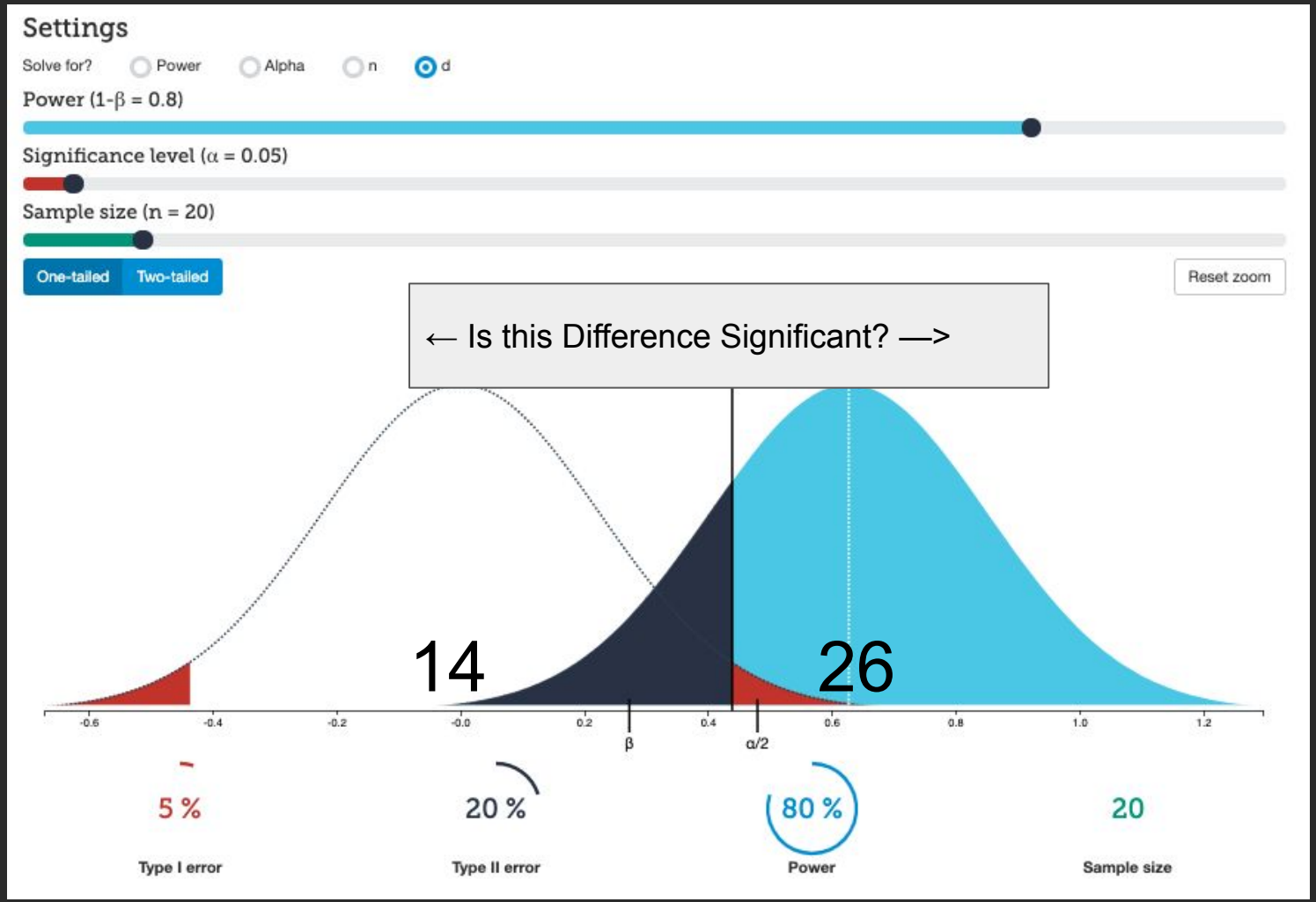
Step One

Scenario: On a standardized anagram task, $\mu = 26$ anagrams solved with a $\sigma = 4$. A researcher tests whether the arousal from anxiety is distracting and will decrease performance. A sample of $n = 14$ anxiety patients is tested on the task. There average performance is 23.36 anagrams.

a. **Step one**: State the null and alternative hypotheses.

$$H_0 : \mu = 26 \qquad H_A : \mu < 26$$

Always consider directionality in this step!!!


Settings
Solve for? Power Alpha n d
Power (1-β = 0.8)
Significance level (α = 0.05)
Sample size (n = 20)
One-tailed Two-tailed
Reset zoom
← Is this Difference Significant? —>
14
26
-0.6 -0.4 -0.2 -0.0 0.2 0.4 0.6 0.8 1.0 1.2
β
α/2
5 %
Type I error
20 %
Type II error
80 %
Power
20
Sample size

Step Two

Scenario: On a standardized anagram task, $\mu = 26$ anagrams solved with a $\sigma = 4$. A researcher tests whether the arousal from anxiety is distracting and will decrease performance. A sample of $n = 14$ anxiety patients is tested on the task. There average performance is 23.36 anagrams.

b. **Step two**: Set the criterion for rejecting H_0. Alpha is usually set to .05, but could be other values depending on the research context. Again, directionality is important to consider.

Step Three and Four

Scenario: On a standardized anagram task, $\mu = 26$ anagrams solved with a $\sigma = 4$. A researcher tests whether the arousal from anxiety is distracting and will decrease performance. A sample of $n = 14$ anxiety patients is tested on the task. There average performance is 23.36 anagrams.

c. **Step three**: Select the sample and collect your data.

d. **Step four**: Locate the region of rejection and the critical value(s) of your test statistic. Again, directionality is important to consider.

Step Five

Scenario: On a standardized anagram task, $\mu = 26$ anagrams solved with a $\sigma = 4$. A researcher tests whether the arousal from anxiety is distracting and will decrease performance. A sample of $n = 14$ anxiety patients is tested on the task. There average performance is 23.36 anagrams.

e. **Step five**: Compute the appropriate test statistic. σ is known, so we use the z test.

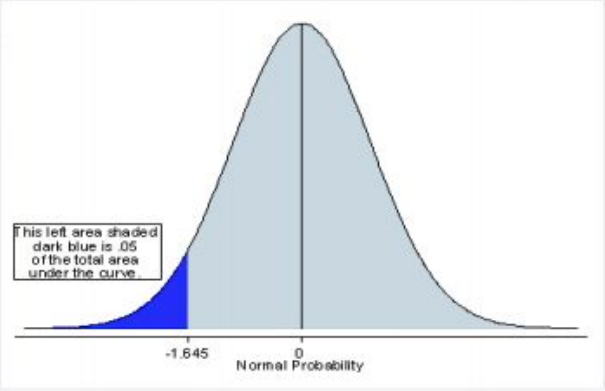
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{14}} = 1.07 \qquad z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{23.36 - 26}{4/\sqrt{14}}$$

$$z = \frac{-2.64}{1.07} = -2.47$$

Step Six

Scenario: On a standardized anagram task, $\mu = 26$ anagrams solved with a $\sigma = 4$. A researcher tests whether the arousal from anxiety is distracting and will decrease performance. A sample of $n = 14$ anxiety patients is tested on the task. There average performance is 23.36 anagrams.

f. **Step six**: Decide whether to reject H_0. Is -2.47 more extreme than the critical value?



NHST Visualization

Settings

Solve for? ○ Power ○ Alpha ○ n ◉ d

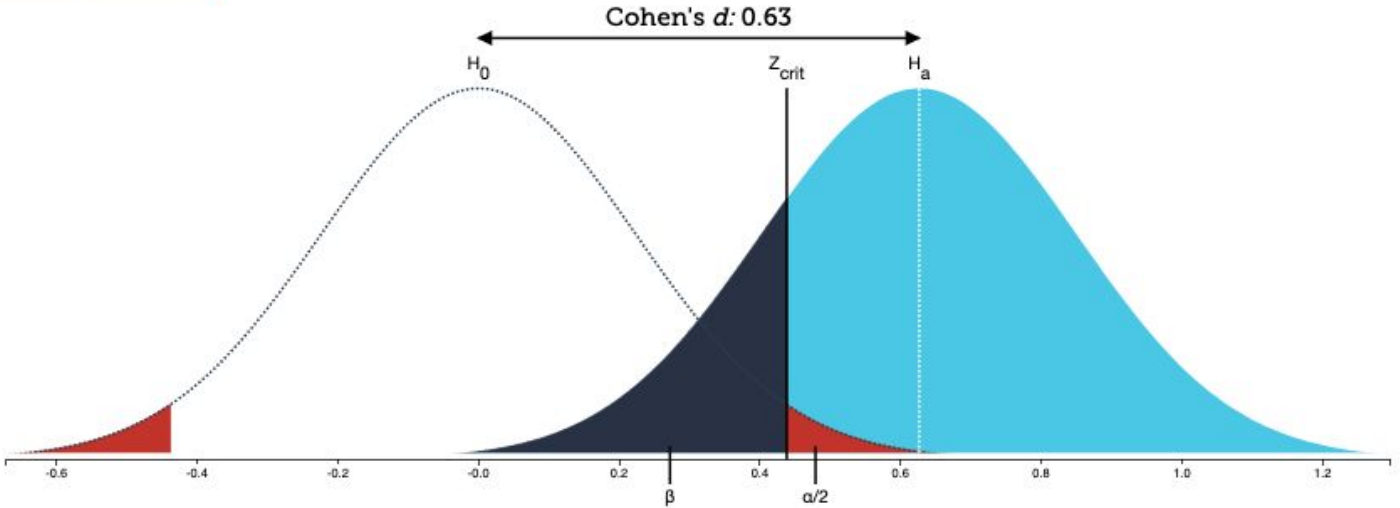
Power (1-β = 0.8)

Significance level (α = 0.05)

Sample size (n = 20)

One-tailed | Two-tailed

Reset zoom



5 %	20 %	80 %	20
Type I error	Type II error	Power	Sample size

https://rpsychologist.com/d3/NHST/