Hypothesis Testing

Presented by David John Baker
August 2019

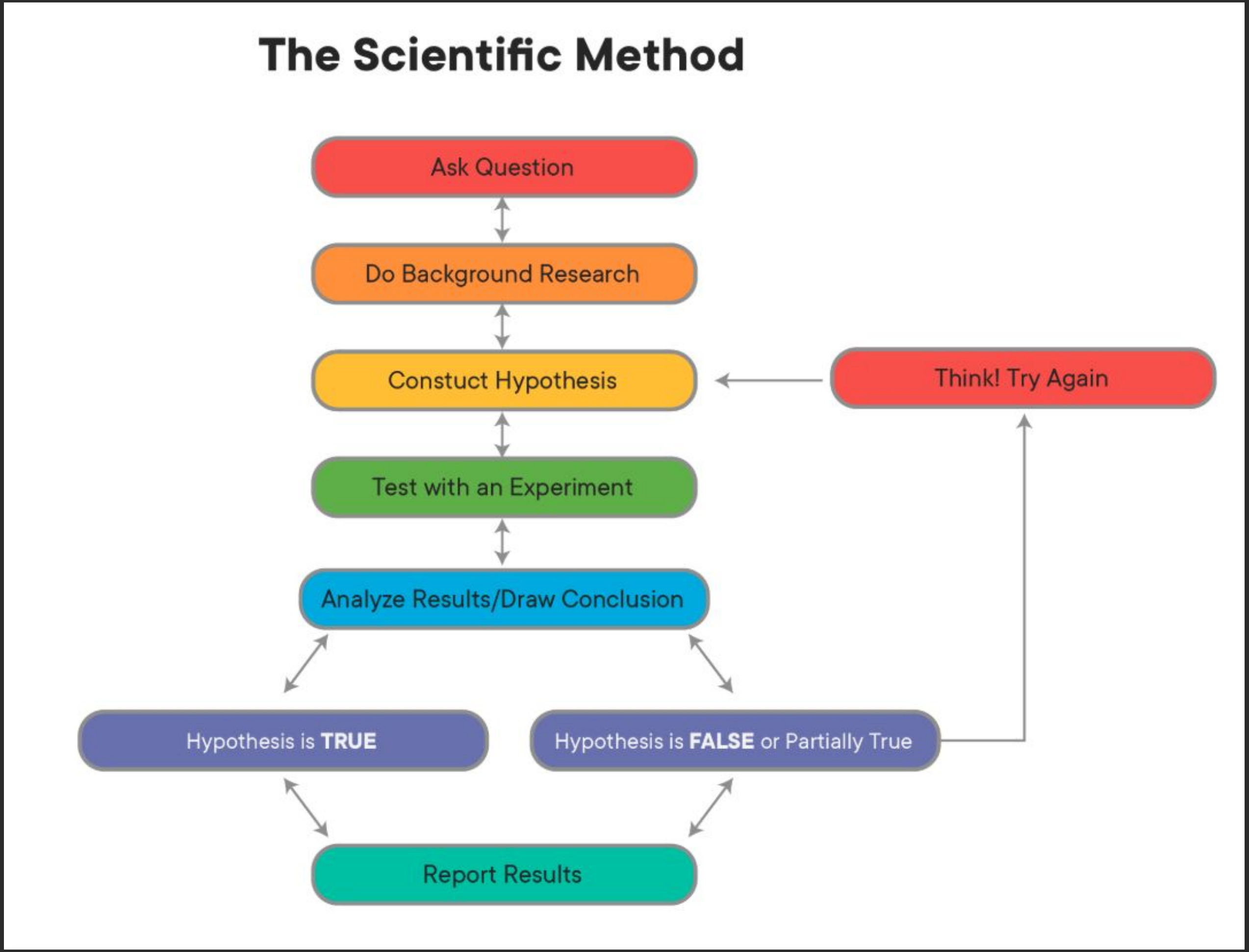
FLATIRON SCHOOL

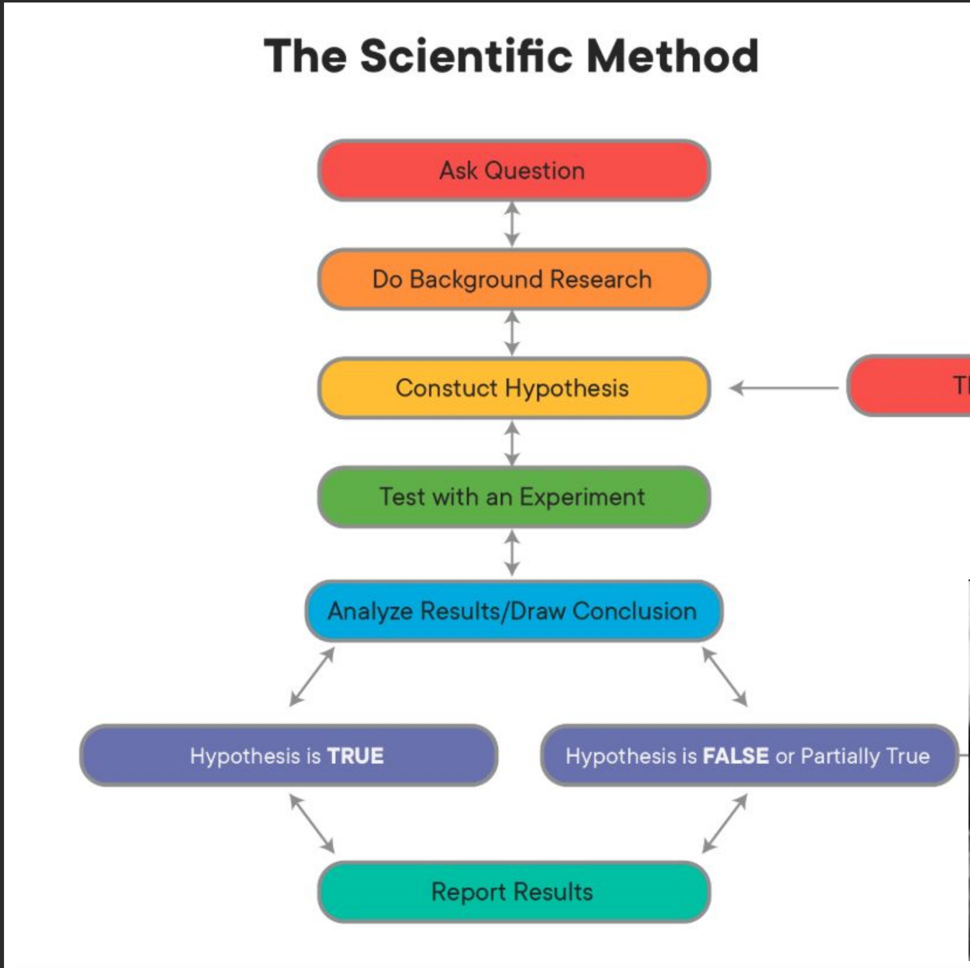
Hypothesis Testing

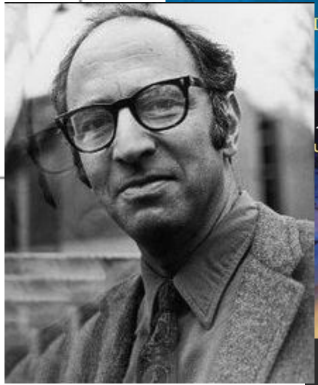
- Knowing stuff about the world is hard
- The Scientific Method is there to help us out
- Today we make our first pass at understanding one way of thinking about how we can do science
- Dave’s Opinion: Time invested in thinking about how we know what we know / philosophy of science will develop your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

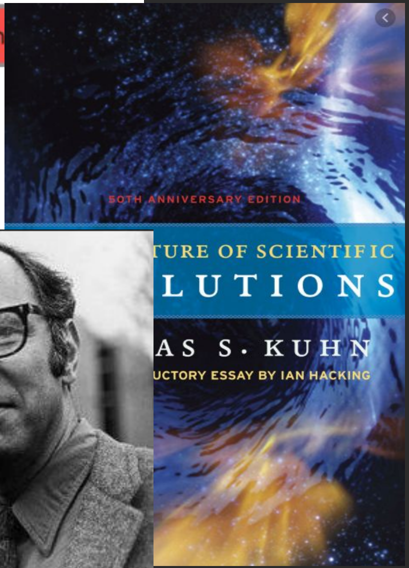
Lesson goals

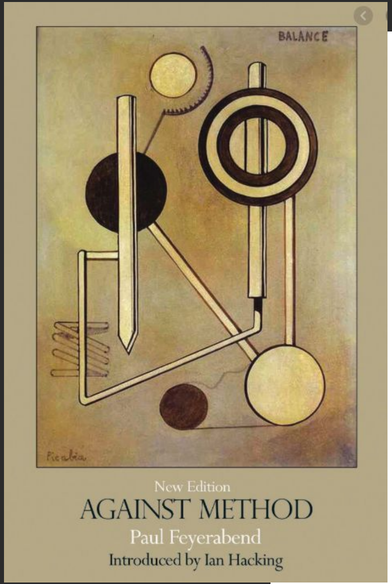
- Scientific Method (Theory vs Practice)
- The Problem of Induction
- Popper, Falsifiability, Demarcation
- Logic of Null Hypothesis Significance Testing
- Four Types of Outcomes in NHST
- Introduction to p values
- Run through a single statistical test

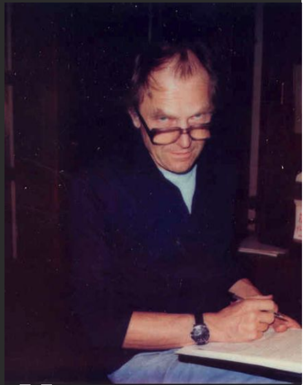

The Scientific Method
Ask Question
Do Background Research
Constuct Hypothesis
Think! Try Again
Test with an Experiment
Analyze Results/Draw Conclusion
Hypothesis is **TRUE**
Hypothesis is **FALSE** or Partially True
Report Results


The Scientific Method
Ask Question
Do Background Research
Constuct Hypothesis
Think! Try Again
Test with an Experiment
Analyze Results/Draw Conclusion
Hypothesis is **TRUE**
Hypothesis is **FALSE** or Partially True
Report Results

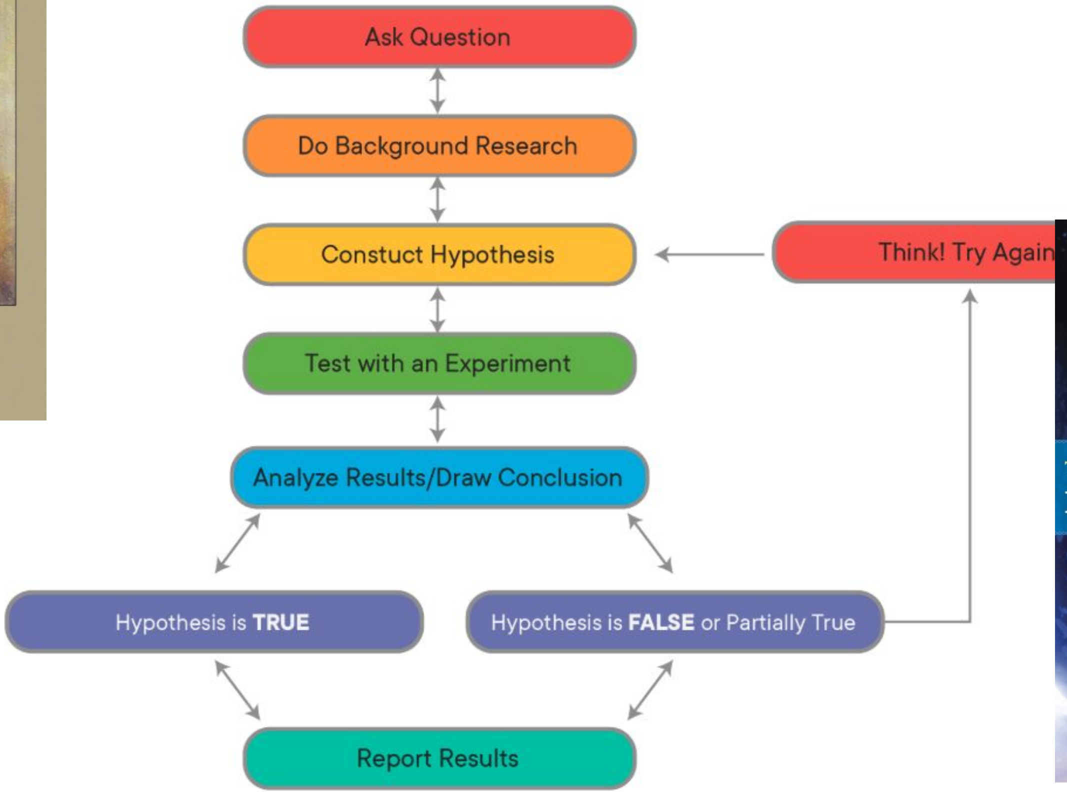


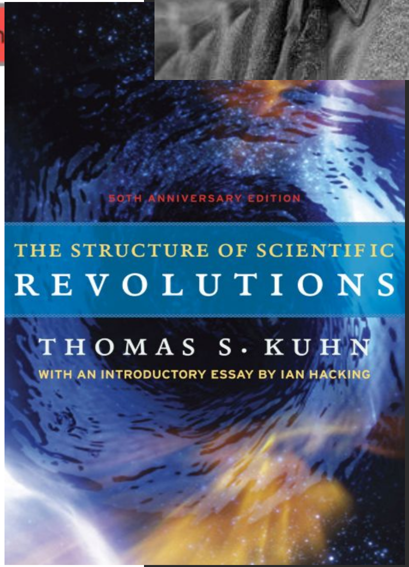

50TH ANNIVERSARY EDITION
TURE OF SCIENTIFIC
LUTIONS
AS S. KUHN
UCTORY ESSAY BY IAN HACKING





The Scientific Method





//

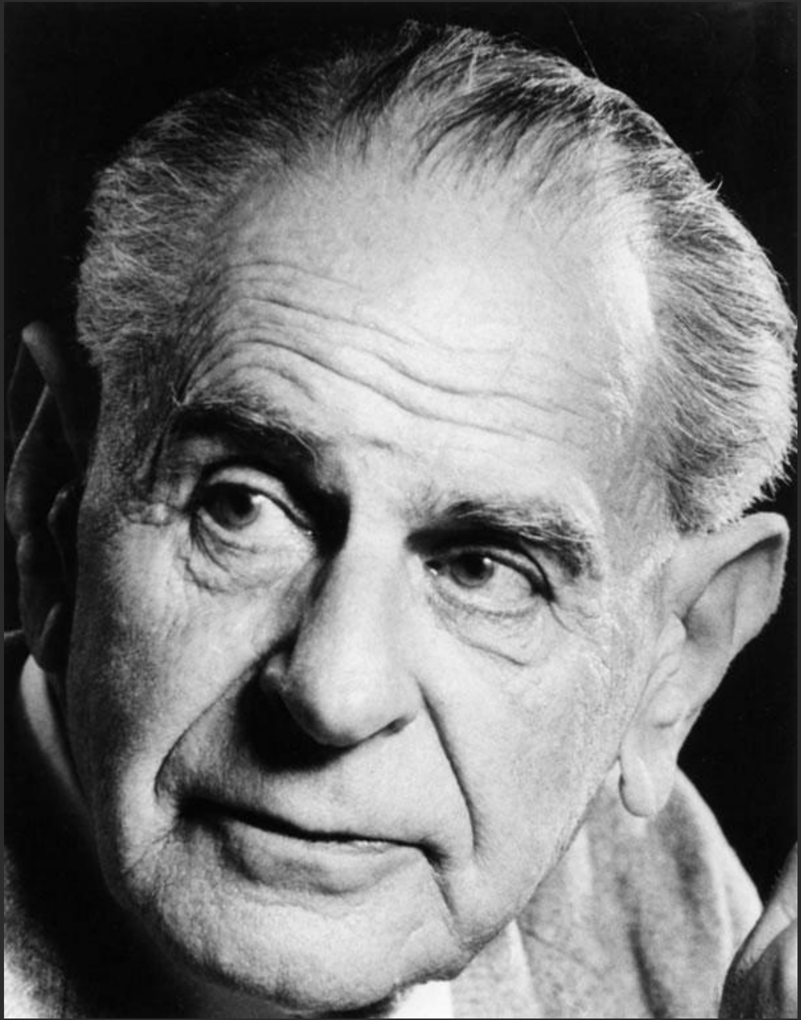
What makes a question scientific?

For Example…

Are questions about astrology (horoscopes) and astronomy (theory of relativity) equally scientific?

Karl Popper

- Problem of demarcation
- Falsifiability
- Problem of Induction

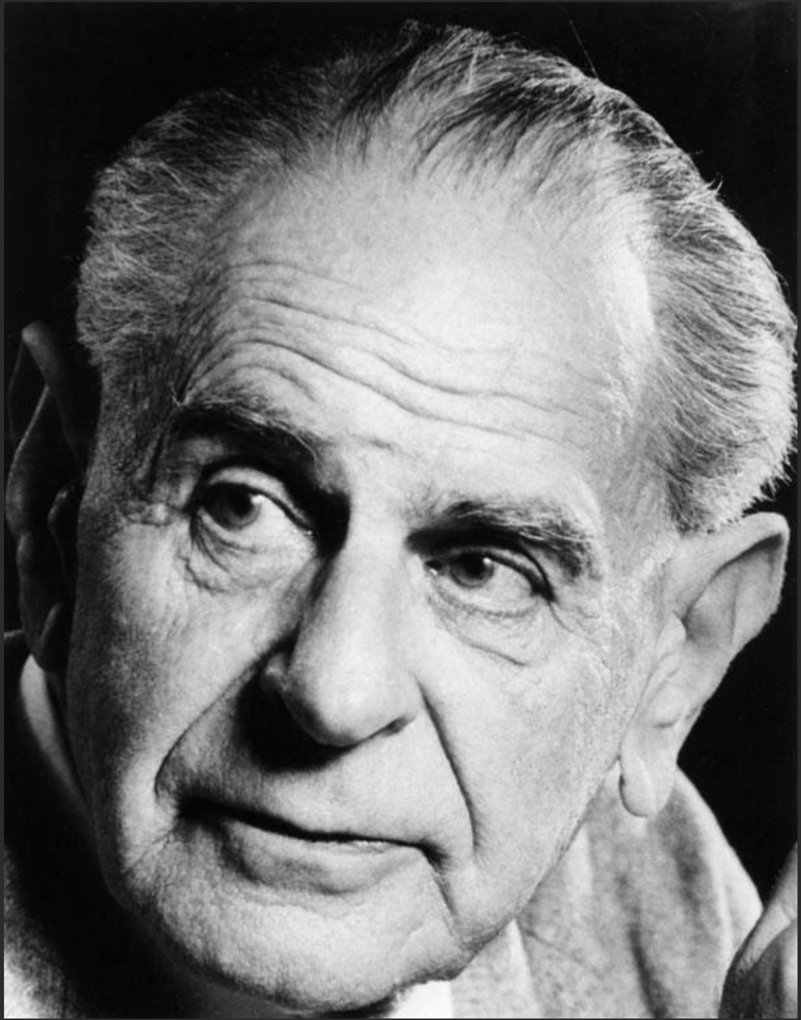


Karl Popper

- Problem of demarcation
- Falsifiability
- Problem of Induction

It’s harder to prove wrong that mercury being in retrograde will affect someone’s mood.

It’s easier to prove wrong that items, when dropped from similar heights will fall at different rates

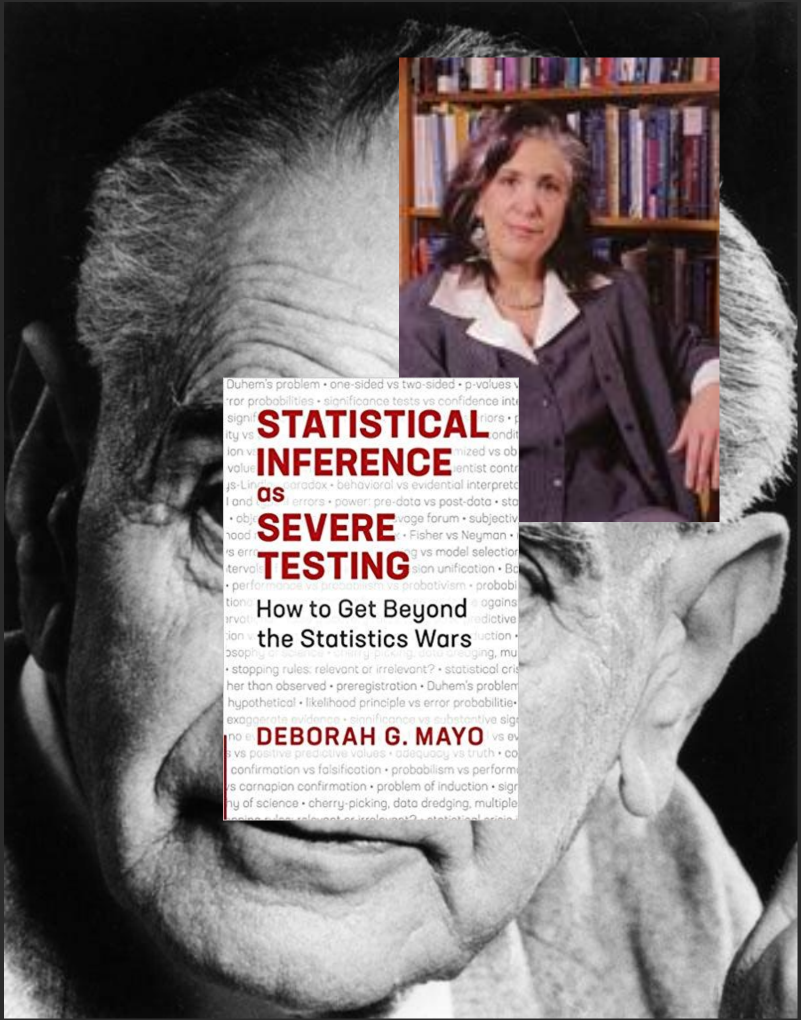


Karl Popper

- Problem of demarcation
- Falsifiability
- Problem of Induction

It’s harder to prove wrong that mercury being in retrograde will affect someone’s mood.

It’s easier to prove wrong that items, when dropped from similar heights will fall at different rates



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

PDF Figures Save Share Reprints Request

Human Olfaction without Apparent Olfactory Bulbs

Tali Weiss [5] • Timna Soroka [5] • Lior Gorodisky • ... Edna Furman-Haran • Thijs Dhollander • Noam Sobel [6] • Show all authors • Show footnotes

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

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

Feedback

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.

Is every day you live more evidence that you will continue to live?!?

Neuron

CASE STUDY | ONLINE NOW

PDF Figures Save Share Reprints Request

Human Olfaction without Apparent Olfactory Bulbs

Tali Weiss [5] • Timna Soroka [5] • Lior Gorodisky • ... Edna Furman-Haran • Thijs Dhollander • Noam Sobel [6] • Show all authors • Show footnotes

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

PlumX Metrics

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…

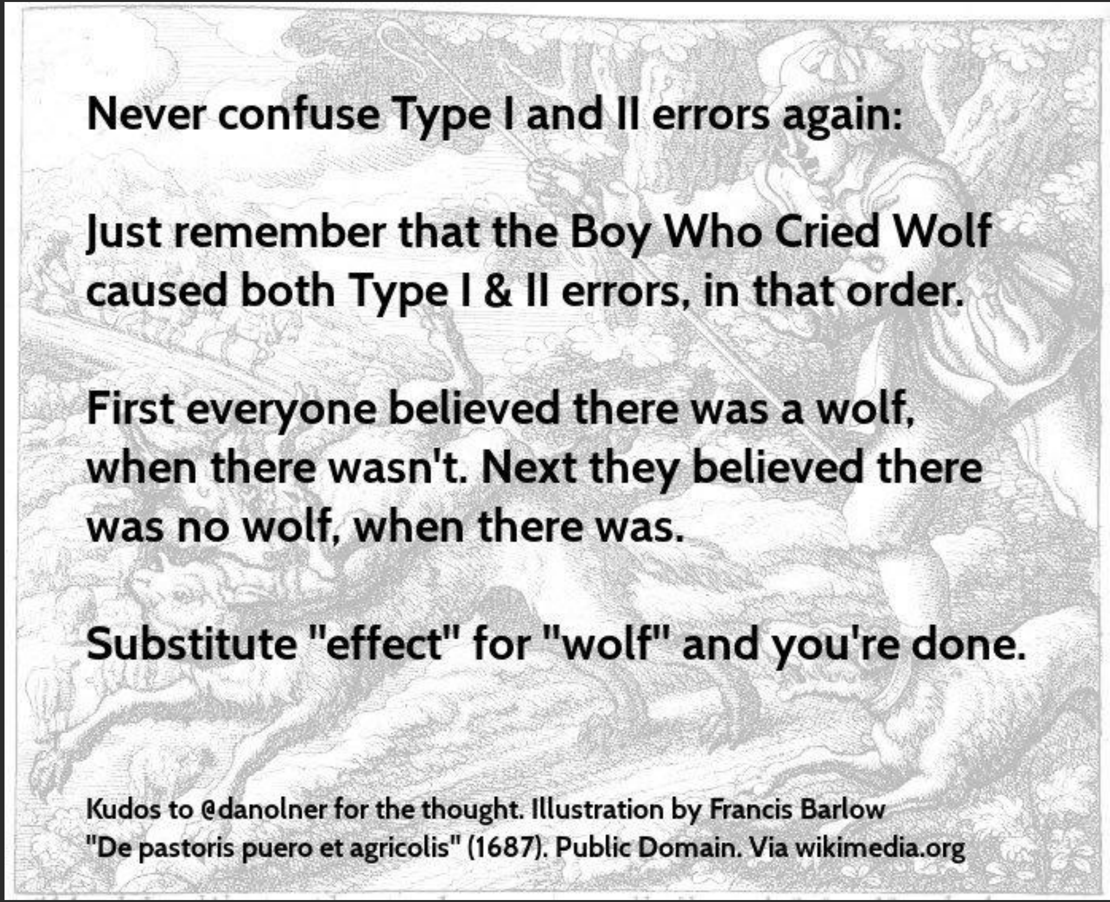
Null Hypothesis Significance Testing Examples

- 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



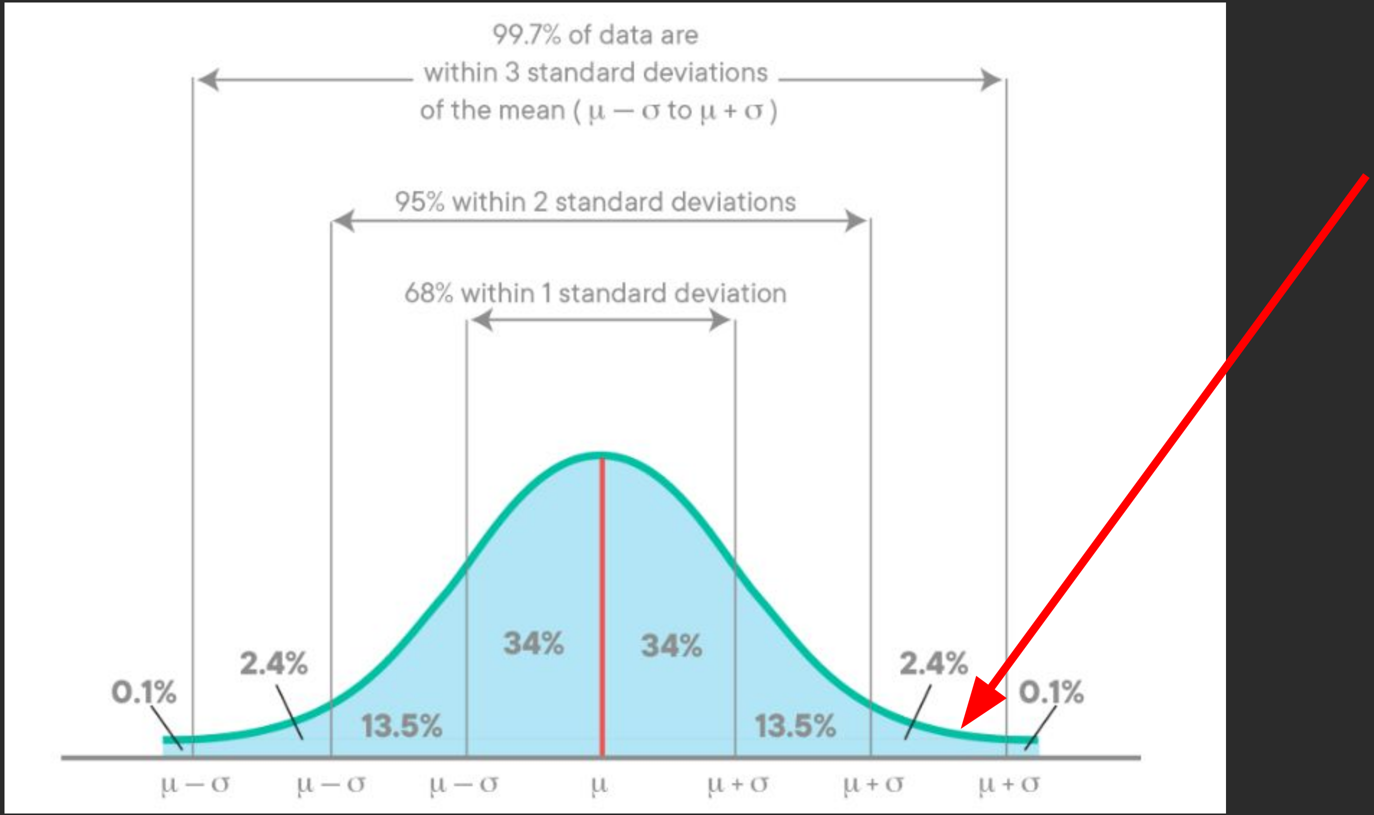
Types of Errors

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

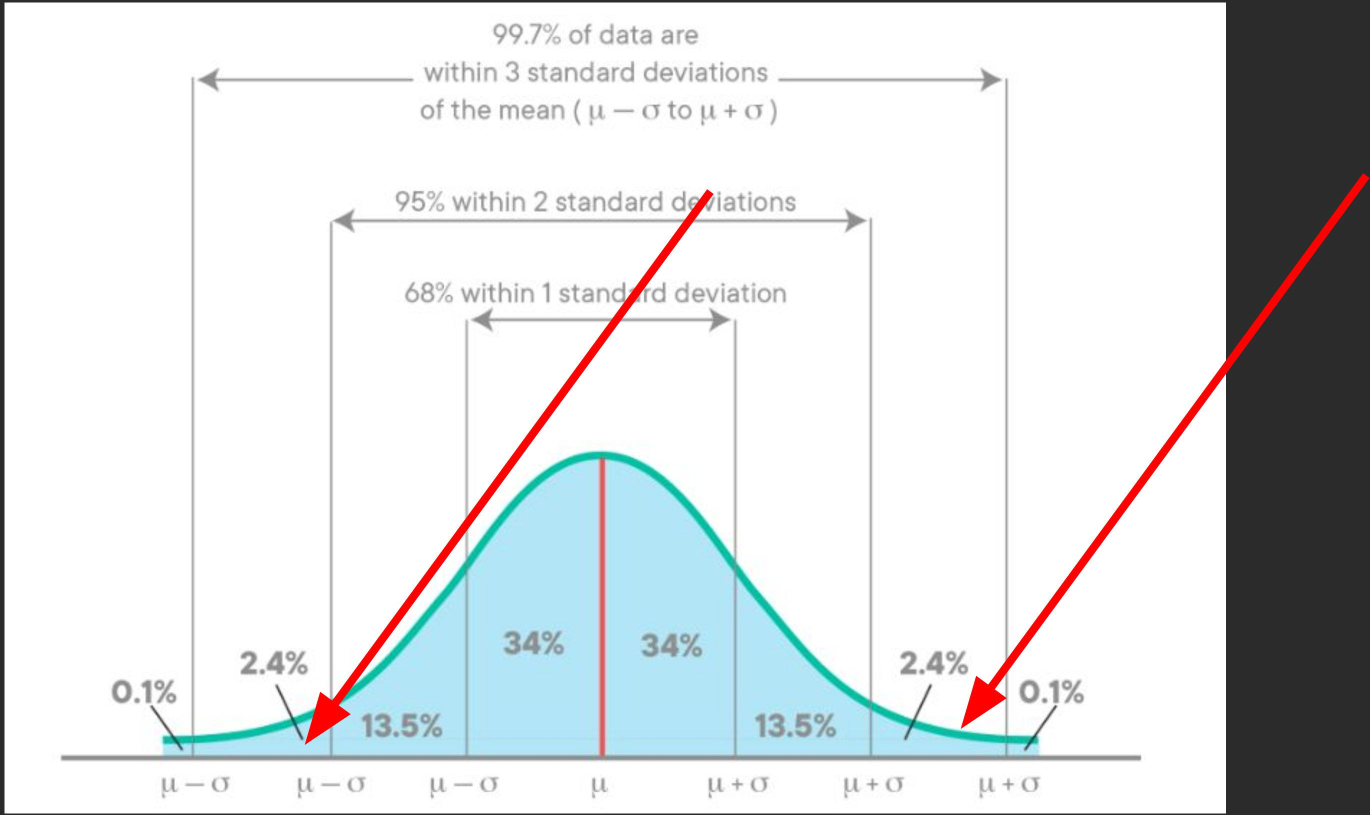
Types of Errors

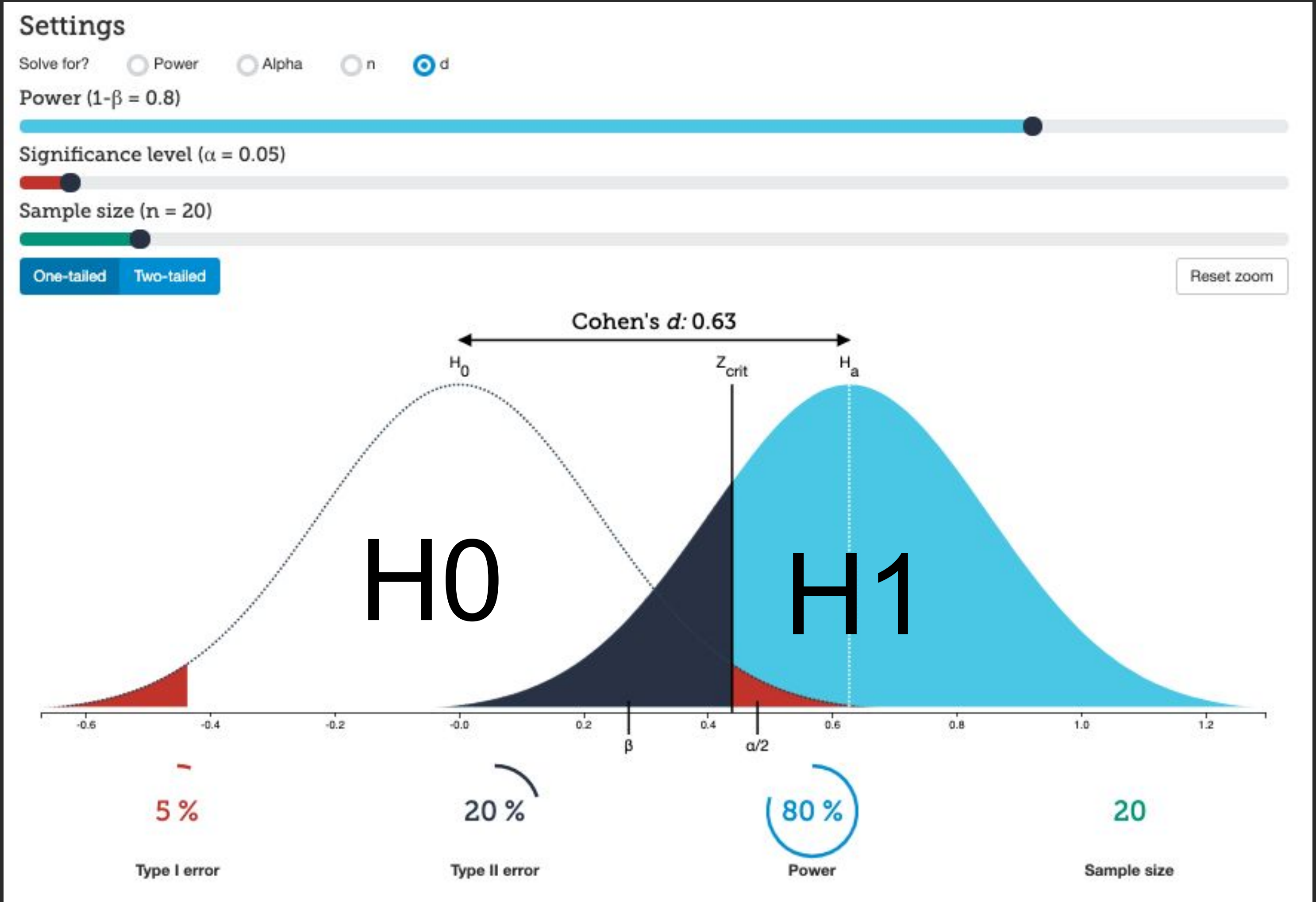
	H0 True (50%)	H1 True (50%)
Significant Finding α = 5%, 1-β=80%	False Positive 5%*50%=2.5%	True Positive 80%*50%=40%
Non-Significant Finding 1-α = 95%, β=20%	True Negative 95%*50%=47.5%	False Negative 20%*50%=10%

Rejection Region (One Tailed)



Rejection Region (Two Tailed)





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
Cohen's d: 0.63
H0
Zcrit
Ha
H0
H1
-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

Calculating a Test Statistic

- The general formula for a test statistic is …

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

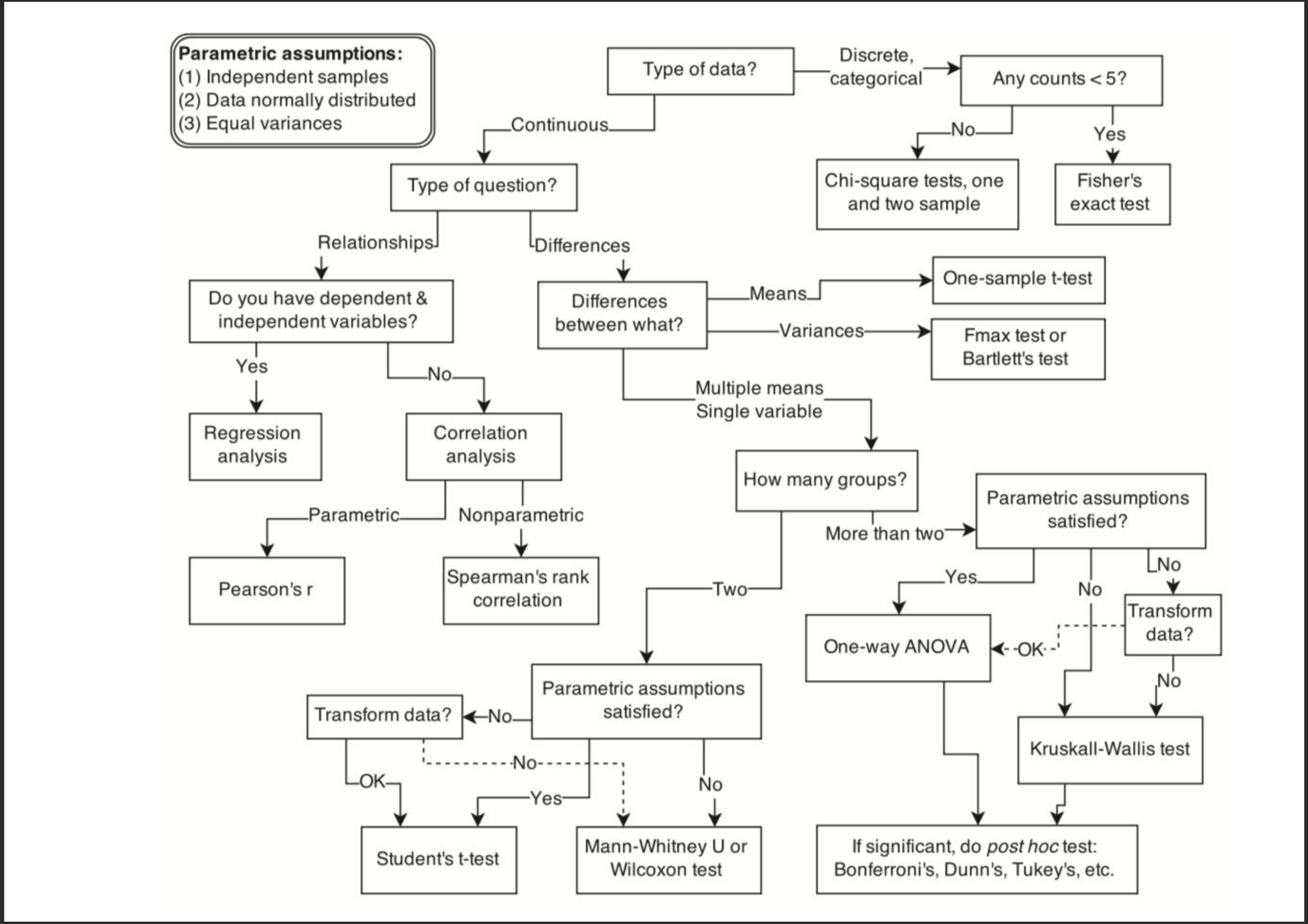
If we know standard deviation, we use z test

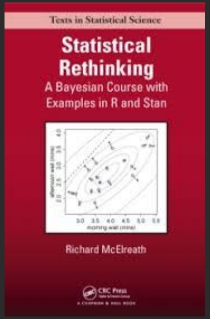
$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

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

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

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

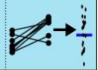
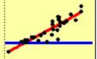
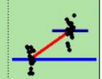
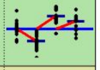
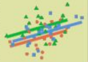

Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances
Type of data?
Discrete, categorical
Any counts < 5?
No
Yes
Chi-square tests, one and two sample
Fisher's exact test
Continuous
Type of question?
Relationships
Differences
Do you have dependent & independent variables?
Yes
No
Regression analysis
Correlation analysis
Parametric
Nonparametric
Pearson's r
Spearman's rank correlation
Differences between what?
Means
One-sample t-test
Variances
Fmax test or Bartlett's test
Multiple means Single variable
How many groups?
More than two
Parametric assumptions satisfied?
Yes
No
No
Transform data?
OK
No
One-way ANOVA
Kruskall-Wallis test
If significant, do post hoc test: Bonferroni's, Dunn's, Tukey's, etc.
Two
Parametric assumptions satisfied?
No
Transform data?
OK
No
Yes
No
Student's t-test
Mann-Whitney U or Wilcoxon test


Texts in Statistical Science
Statistical Rethinking
A Bayesian Course with Examples in R and Stan
Richard McElreath
CRC Press

Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise **y_2-y_1** differences. - (Same, but it predicts the *signed rank* of **y_2-y_1**.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=...[B])[A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x_1 + x_2 +...)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +...+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +...+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +...+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 + ... + G_N + S_2 + S_3+ ... + S_K + G_2*S_2+G_3*S_3+...+G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: **$G_{2\ to\ N}$** is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for **$S_{2\ to\ K}$** for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + ... + G_N + S_2 + S_3+ ... + S_K + G_2*S_2+G_3*S_3+...+G_N*S_K, family=...)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments: `glm(model, family=poisson())` As linear-model, the Chi-square test is $\log(y_i) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +...+ G_N, family=...)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for y = 1·b + a·x which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when Δx = 1 between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.

[B] Same model, but with one variance per group: `gls(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.

Jonas Kristoffer Lindeløv
https://lindeloev.net

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!

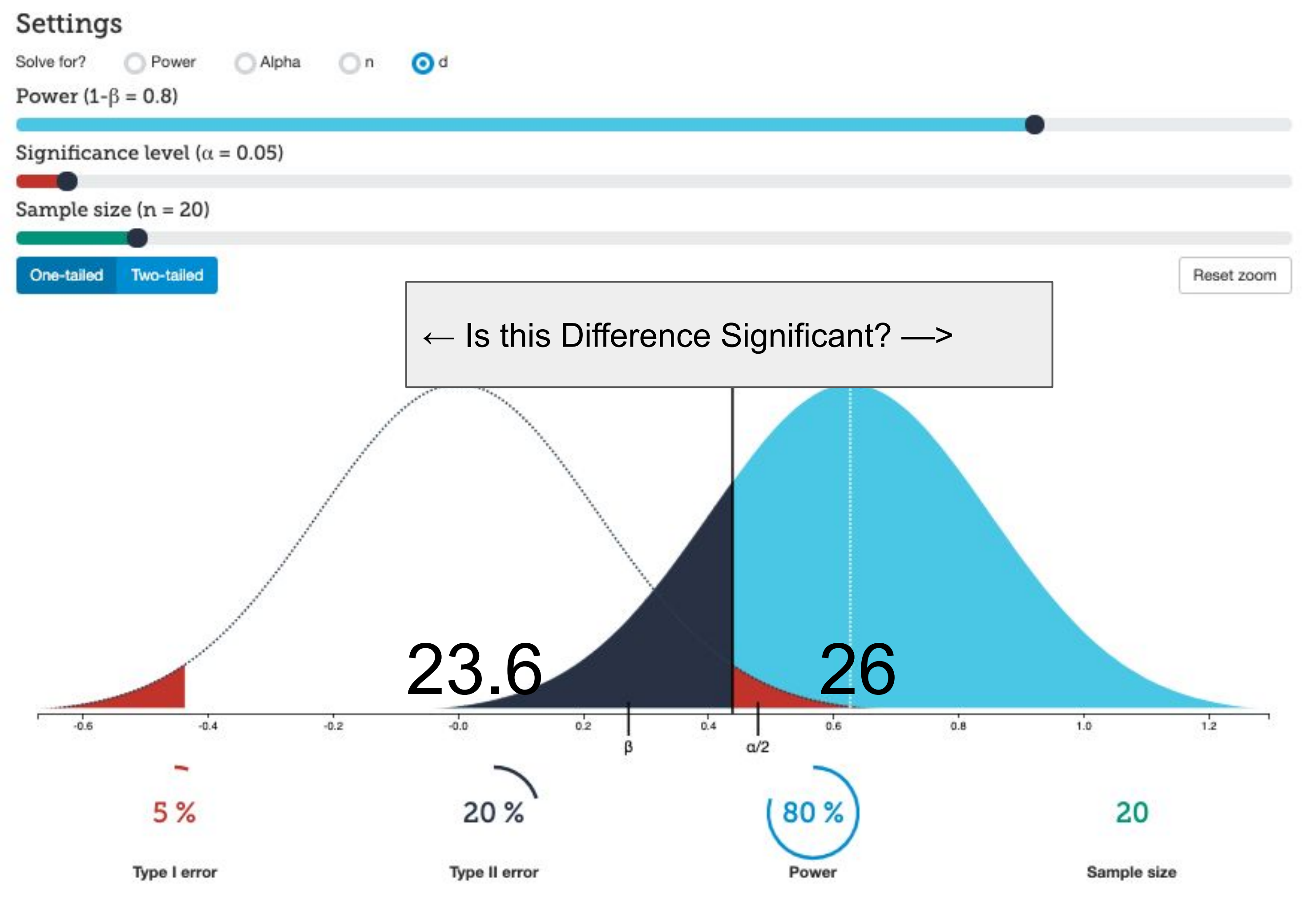
Step One

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? —>
23.6
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.

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

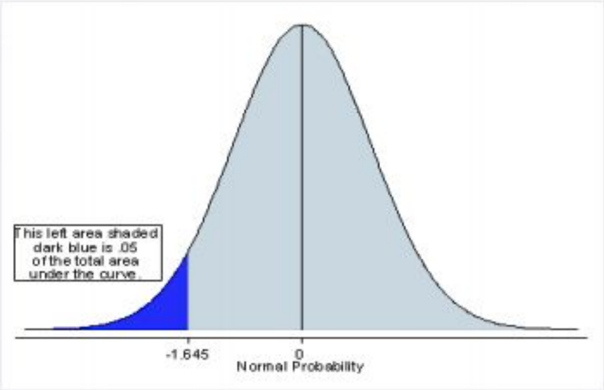
$$z = \frac{-2.64}{1.07} = -2.47$$

//

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?



Step Six (Only using positive critical test table…*)

Scenario: On a standardized anagram task, $\mu = 26$ anagrams solved with a $\sigma = 4$. A research… and will decrease… tested on the task…

f. **Step six**: Decid… than the critical…

See that our CRITICAL VALUE of -2.47 is GREATER THAN our ONE tailed test of (-)1.7771

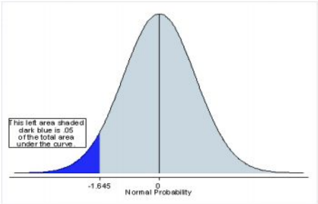
Significance level

Degrees of freedom						
Two-tailed test:	[illegible]	[illegible]	[illegible]	1%	0.2%	0.1%
One-tailed test:	5%	2.5%	1%	0.5%	0.1%	0.05%
1	[illegible]	[illegible]	[illegible]	[illegible]	318.309	636.619
2	2.920	4.303	6.965	9.925	22.327	31.599
3	2.353	3.182	4.541	5.841	10.215	12.924
4	2.132	2.776	3.747	4.604	7.173	8.610
5	2.015	2.571	3.365	4.032	5.893	6.869
6	1.943	2.447	3.143	3.707	5.208	5.959
7	1.894	2.365	2.998	3.499	4.785	5.408
8	1.860	2.306	2.896	3.355	4.501	5.041
9	1.833	2.262	2.821	3.250	4.297	4.781
10	1.812	2.228	2.764	3.169	4.144	4.587
11	1.796	2.201	2.718	3.106	4.025	4.437
12	1.782	2.179	2.681	3.055	3.930	4.318
13	1.771	2.160	2.650	3.012	3.852	4.221
14	1.761	2.145	2.624	2.977	3.787	4.140
15	[illegible]	2.131	2.602	2.947	3.733	4.073
16	1.746	2.120	2.583	2.921	3.686	4.015
17	1.740	2.110	2.567	2.898	3.646	3.965
[illegible]	1.734	2.101	2.552	2.878	3.610	3.922
19	1.729	2.093	2.539	2.861	3.579	3.883
20	1.725	2.086	2.528	2.845	3.552	3.850
21	1.721	2.080	2.518	2.831	3.527	3.819
22	1.717	2.074	2.508	2.819	3.505	3.792
23	1.714	2.069	2.500	2.807	3.485	3.768
24	1.711	2.064	2.492	2.797	3.467	3.745
25	1.708	2.060	2.485	2.787	3.450	3.725

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?



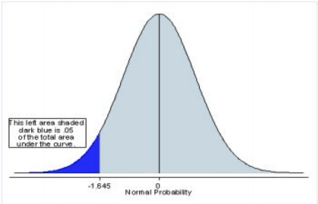
Since our TEST STATISTIC (2.47) is greater than our CRITICAL VALUE (1.77) we have reason to believe that the sample of data we gathered comes from a population so different from the original we can declare it to be significantly different (in the *statistical sense*).

//

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?



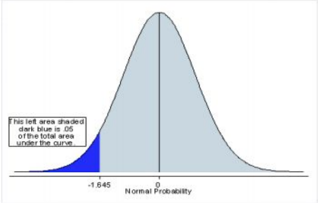
The blue region also corresponds to a very small p value since it's so far out in the tail. Here we did not compute the actual p value of the TEST STATISTIC, but know it's less than .05 (set earlier) since it is larger than our CRITICAL VALUE.

//

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?

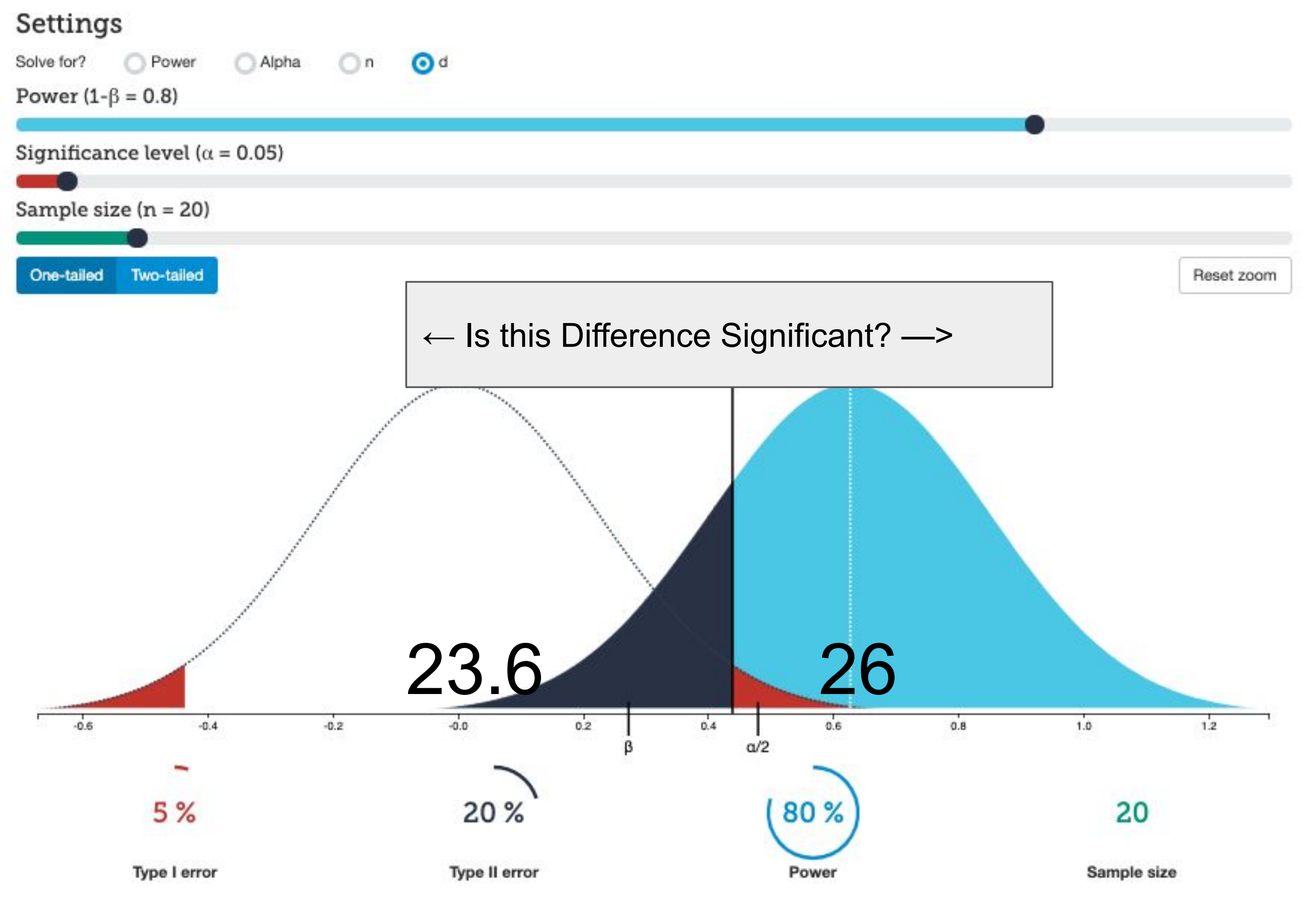


Computers will calculate your p value in the future!! All it tells you is "the probability of the observed, or more extreme, data, under the assumption that the null-hypothesis is true".

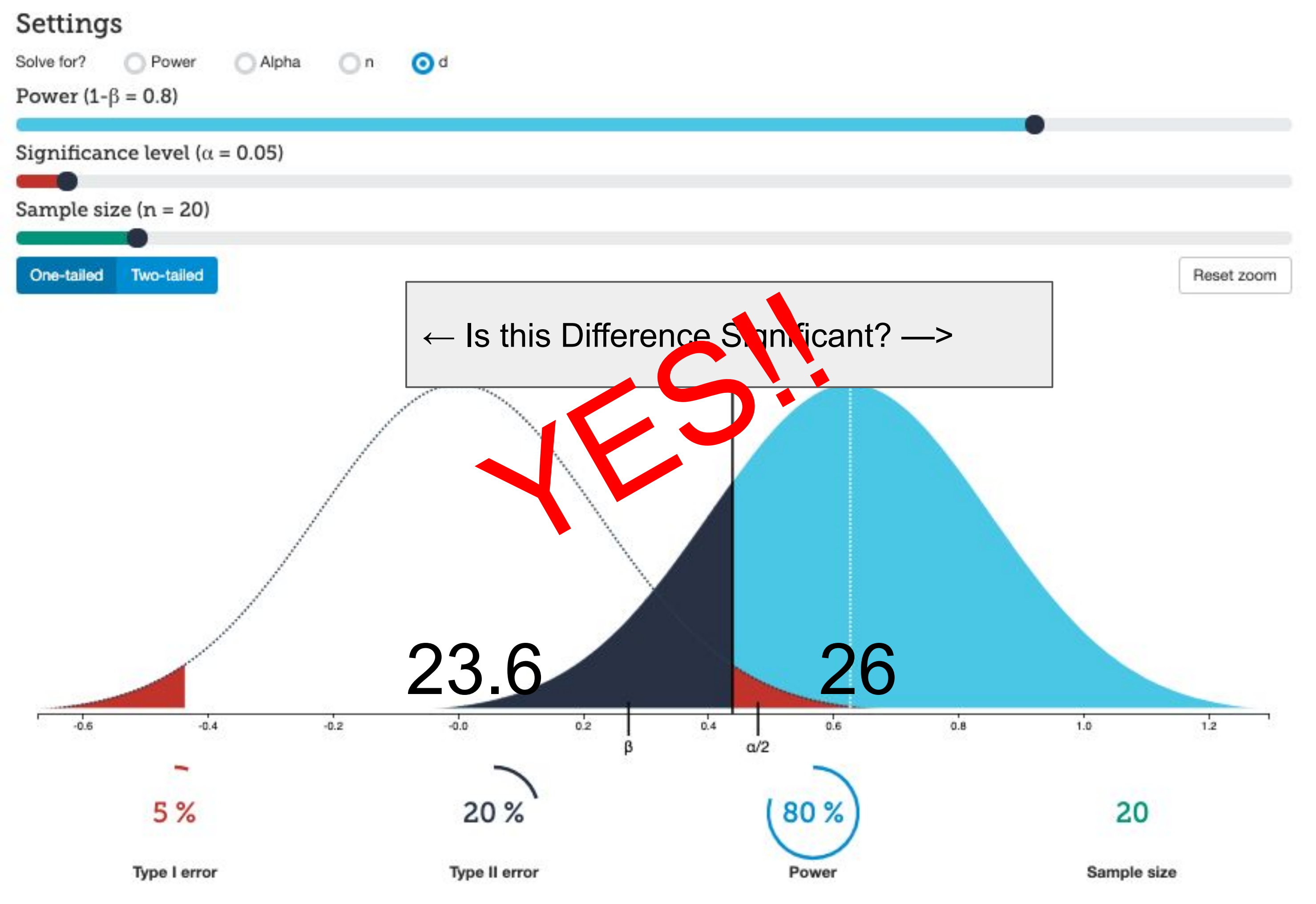
NO MORE. NO LESS. IT IS NOT THE PROBABILITY OF YOUR HYPOTHEIS BEING CORRECT! YOU NEED BAYESIAN STATS FOR THAT!!

http://daniellakens.blogspot.com/2017/12/understanding-common-misconceptions.html

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
← Is this Difference Significant? —>
23.6
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



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? —>
YES!!
23.6
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

Is it worth writing home about?

That depends on your theoretical framework, reliability of your tools, if the samples were representative of your larger population and didn't have any biases, repeats again under different conditions… and all that. Need to also consider your alpha, power, size of the effect,(how big the difference is in a standardized measure that incorporates errors) and how many people were in your sample!

Just because it's SIGNIFICANT does not mean it's MEANINGFUL.

Statistical significant JUST gives us a sanity check when we investigate this paradigm again and again and again to make sure we are on the right track!

Good statistics will never make up for bad theory!!

Settings

Solve for? ○ Power ○ Alpha ○ n ◉ d

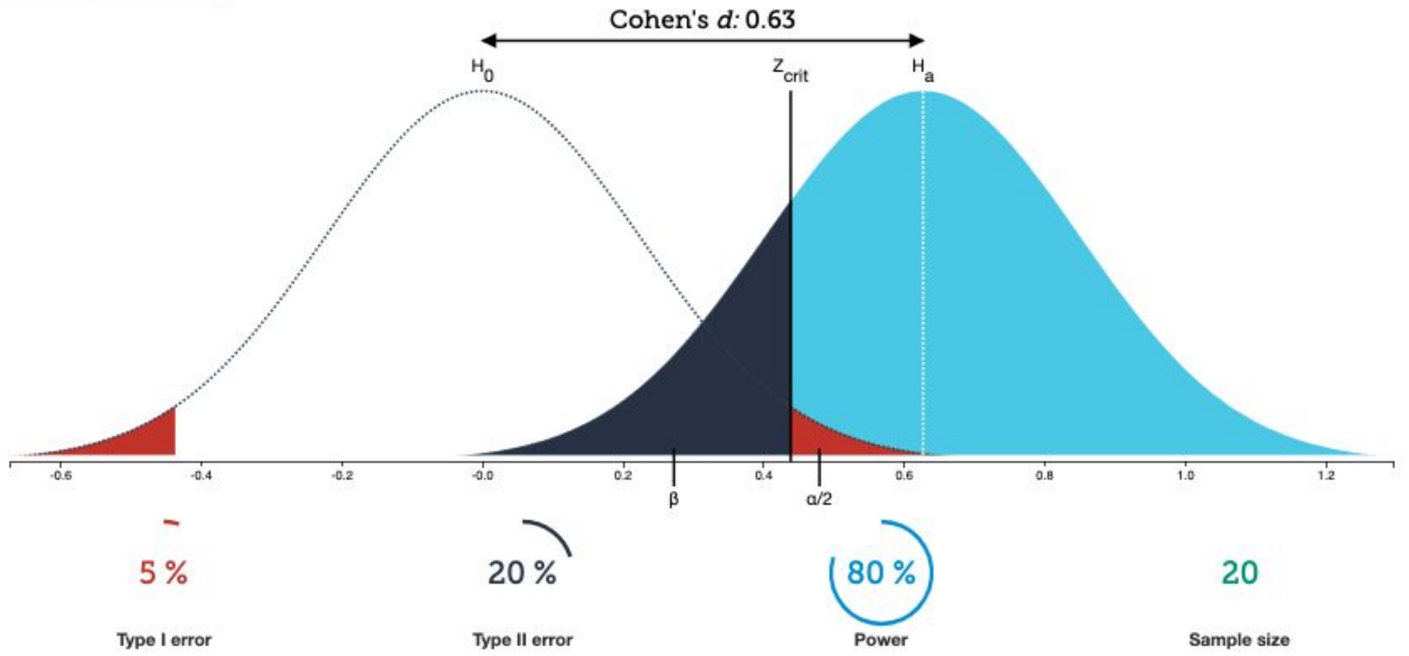
Power (1-β = 0.8)

Significance level (α = 0.05)

Sample size (n = 20)

One-tailed Two-tailed

Reset zoom



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

Turn to your partner, explain one of the following to them, write one question down about something you don't understand

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

//