0:00 / 2:30

Null Hypothesis Significance Testing (NHST)

Professor Andy Field

@profandyfield

www.youtube.com/user/ProfAndyField/

www.discoveringstatistics.com

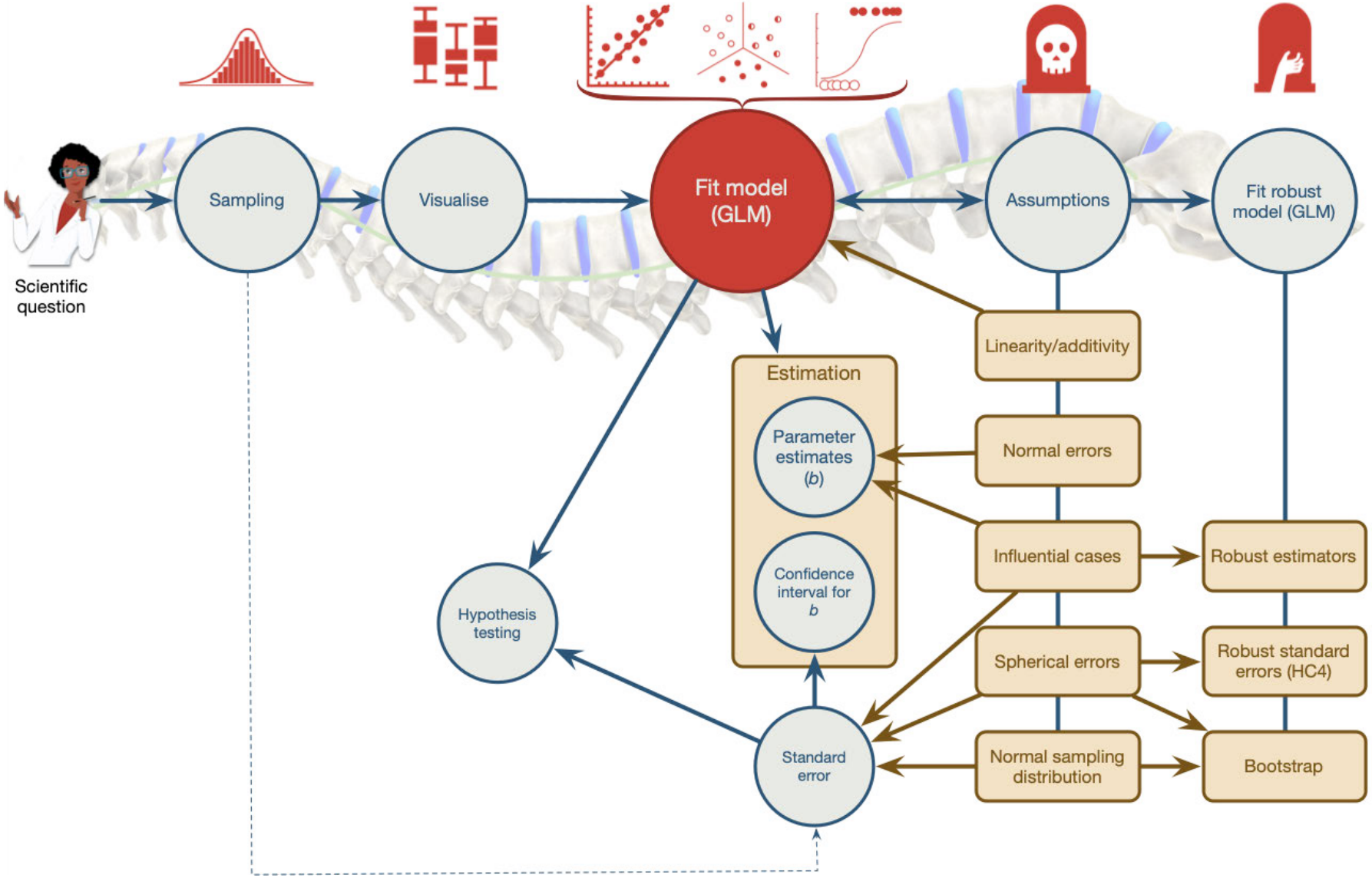
www.milton-the-cat.rocks

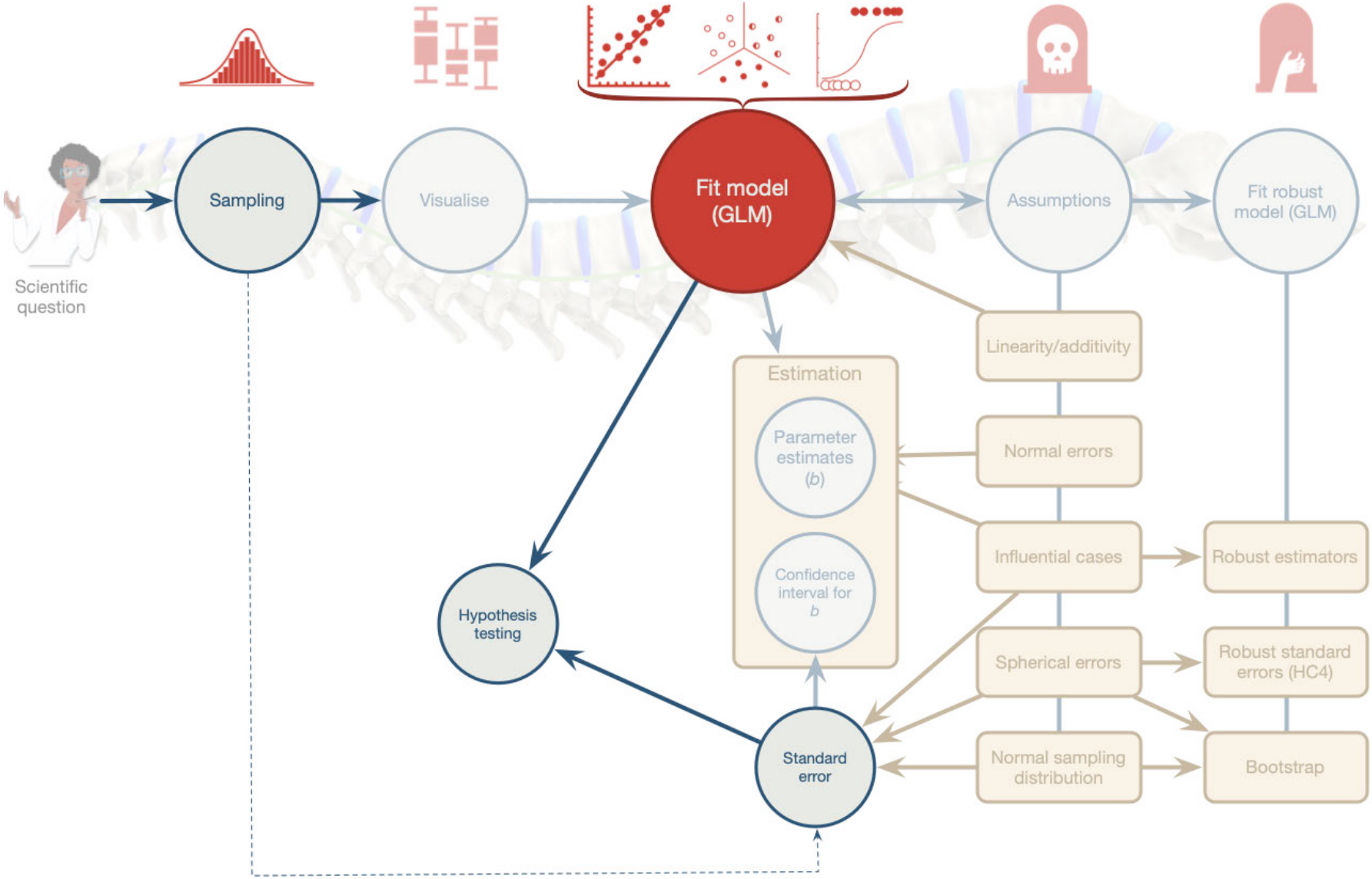
www.discovr.rocks

The SPINE of statistics

5 Key concepts

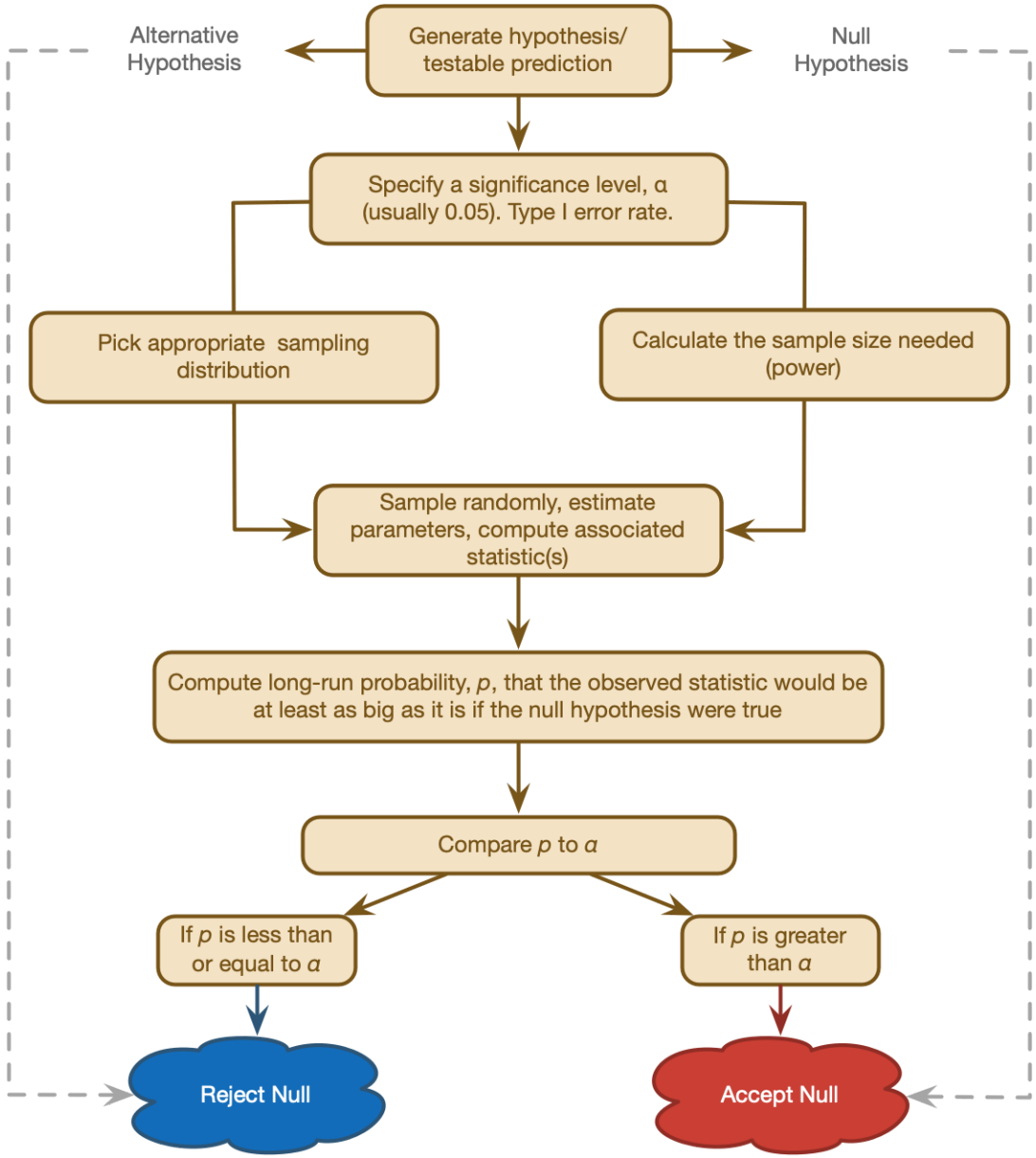
- Standard error
- Parameters
- Interval estimates
- Null hypothesis significance testing (NHST)
- Estimation


Scientific question
Sampling
Visualise
Fit model (GLM)
Assumptions
Fit robust model (GLM)
Linearity/additivity
Estimation
Parameter estimates (b)
Confidence interval for b
Normal errors
Influential cases
Robust estimators
Hypothesis testing
Spherical errors
Robust standard errors (HC4)
Standard error
Normal sampling distribution
Bootstrap

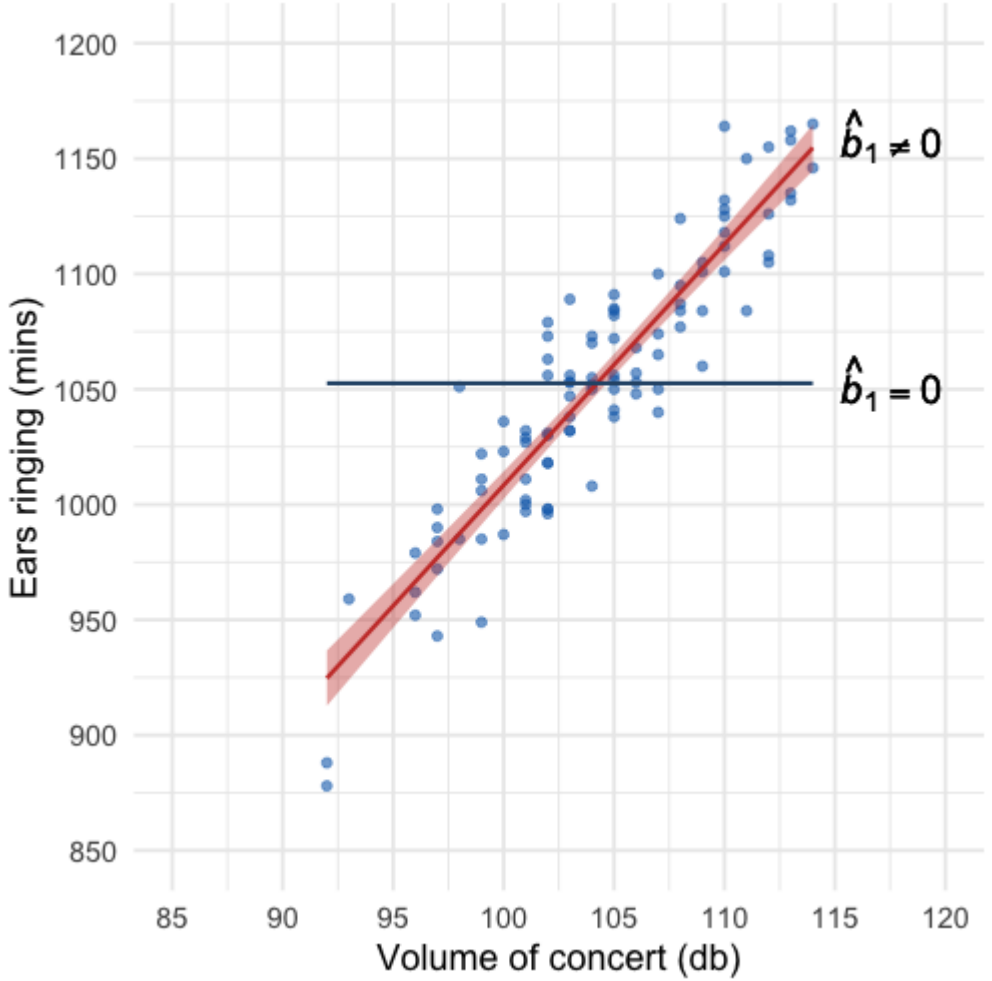

Scientific question
Sampling
Visualise
Fit model (GLM)
Assumptions
Fit robust model (GLM)
Linearity/additivity
Estimation
Parameter estimates (b)
Confidence interval for b
Normal errors
Influential cases
Robust estimators
Hypothesis testing
Spherical errors
Robust standard errors (HC4)
Standard error
Normal sampling distribution
Bootstrap

Learning outcomes

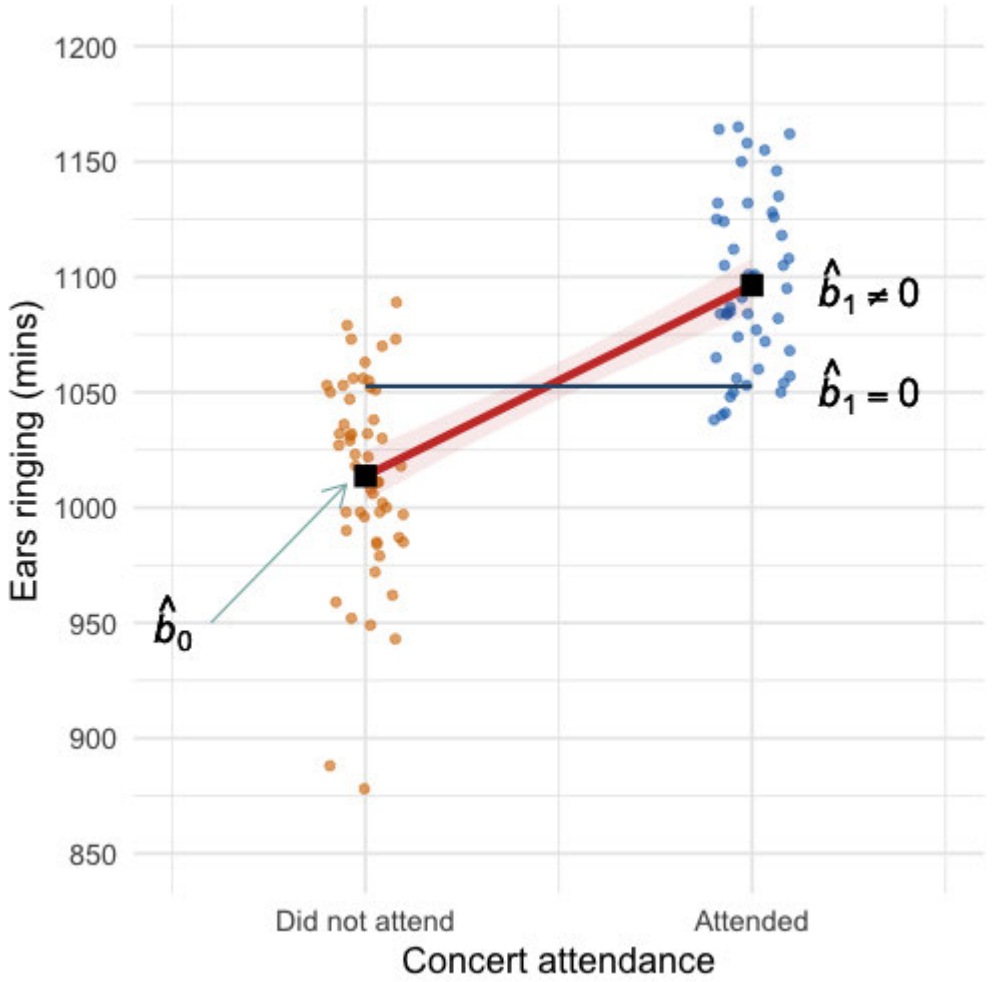
- Null hypothesis significance testing (NHST)
  - Understand the process of significance testing parameters
  - Understand what a *p*-value represents
  - Understand what a *p*-value does NOT represent
- Problems with NHST
  - Be able to articulate the limitations of NHST
- Understand what an effect size is and how it should be used to contextualise significance tests


Alternative Hypothesis
Generate hypothesis/ testable prediction
Null Hypothesis
Specify a significance level, α (usually 0.05). Type I error rate.
Pick appropriate sampling distribution
Calculate the sample size needed (power)
Sample randomly, estimate parameters, compute associated statistic(s)
Compute long-run probability, p, that the observed statistic would be at least as big as it is if the null hypothesis were true
Compare p to α
If p is less than or equal to α
If p is greater than α
Reject Null
Accept Null

$$\widehat{\text{ringing}}_i = \hat{b}_0 + \hat{b}_1 \text{volume}_i + e_i$$

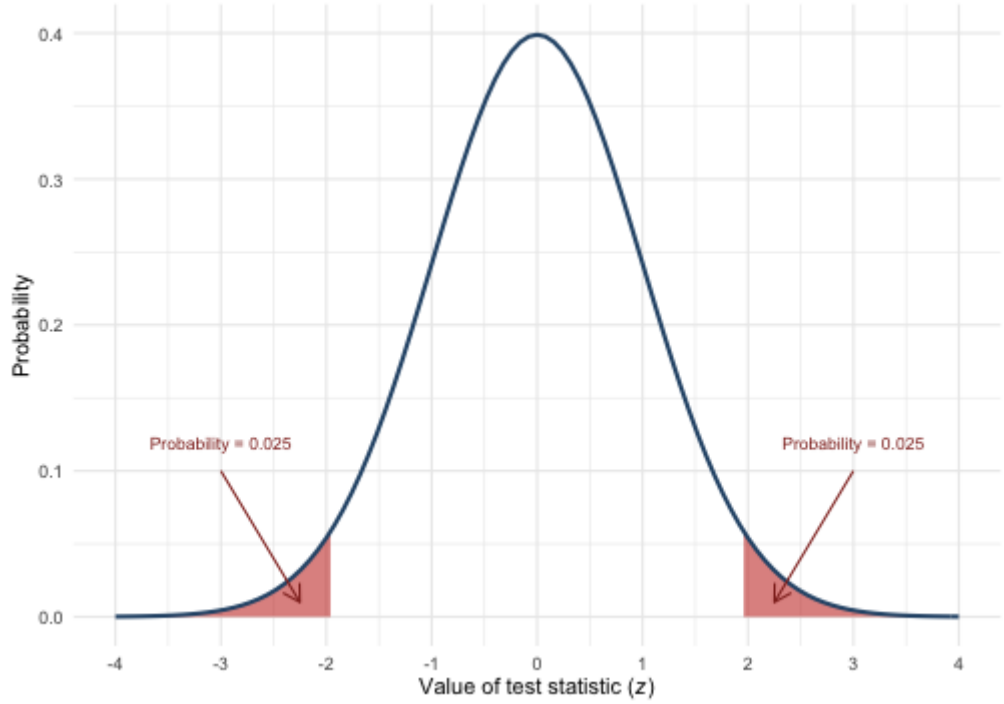


$$\hat{\text{ringing}}_i = \hat{b}_0 + \hat{b}_1 \text{attendance}_i + e_i$$



The long-run probability of the test statistic

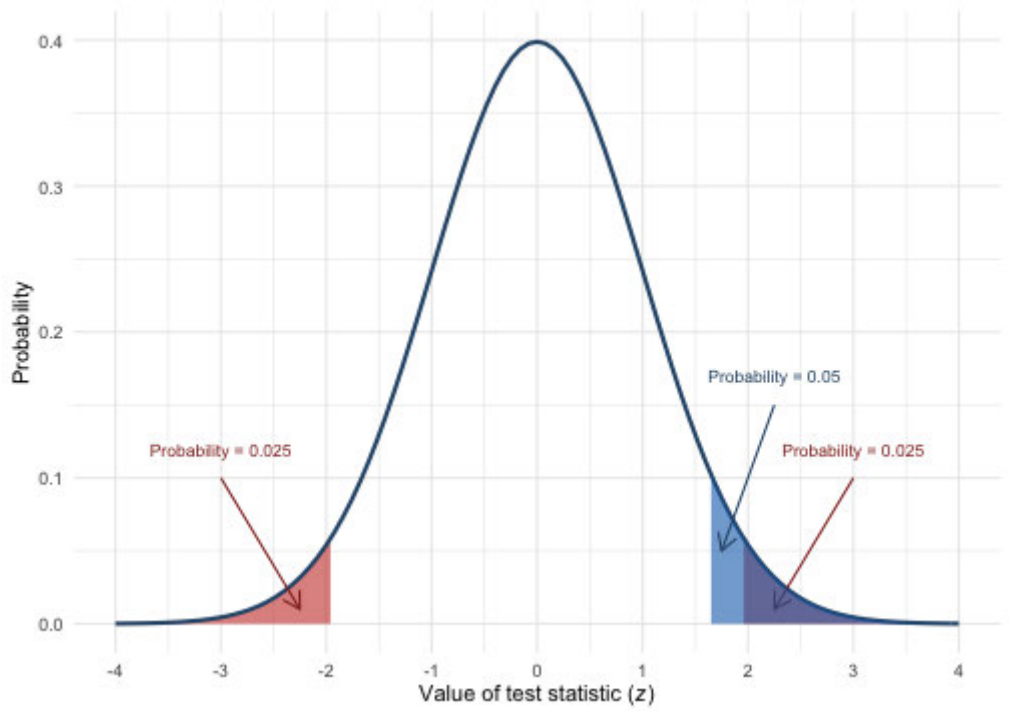
- Parameters represent effects:
  - Relationships between variables
  - Differences between means
- Parameters reflect hypotheses:
  - H_0: $b = 0$ or $b_1 = b_2$
  - H_1: $b \neq 0$ or $b_1 \neq b_2$
- All parameters have an associated sampling distribution
  - For any parameter, we can work out the probability of getting at least the value we have if the null hypothesis is true (e.g., if $b = 0$, or $b_1 \neq b_2$)
  - p < 0.05 is typically used as a threshold for 'significance'



$$t = \frac{b}{SE_b}$$

The long-run probability of the test statistic

- Parameters represent effects:
  - Relationships between variables
  - Differences between means
- Parameters reflect hypotheses:
  - H_0: $b = 0$ or $b_1 = b_2$
  - H_1: $b \neq 0$ or $b_1 \neq b_2$
- All parameters have an associated sampling distribution
  - For any parameter, we can work out the probability of getting at least the value we have if the null hypothesis is true (e.g., if $b = 0$, or $b_1 \neq b_2$)
  - p < 0.05 is typically used as a threshold for 'significance'



$$t = \frac{b}{SE_b}$$

What is a *p*-value?

Hypothesis

H_0: Alice does not want to date Zach

H_1: Alice wants to date Zach

Test statistic

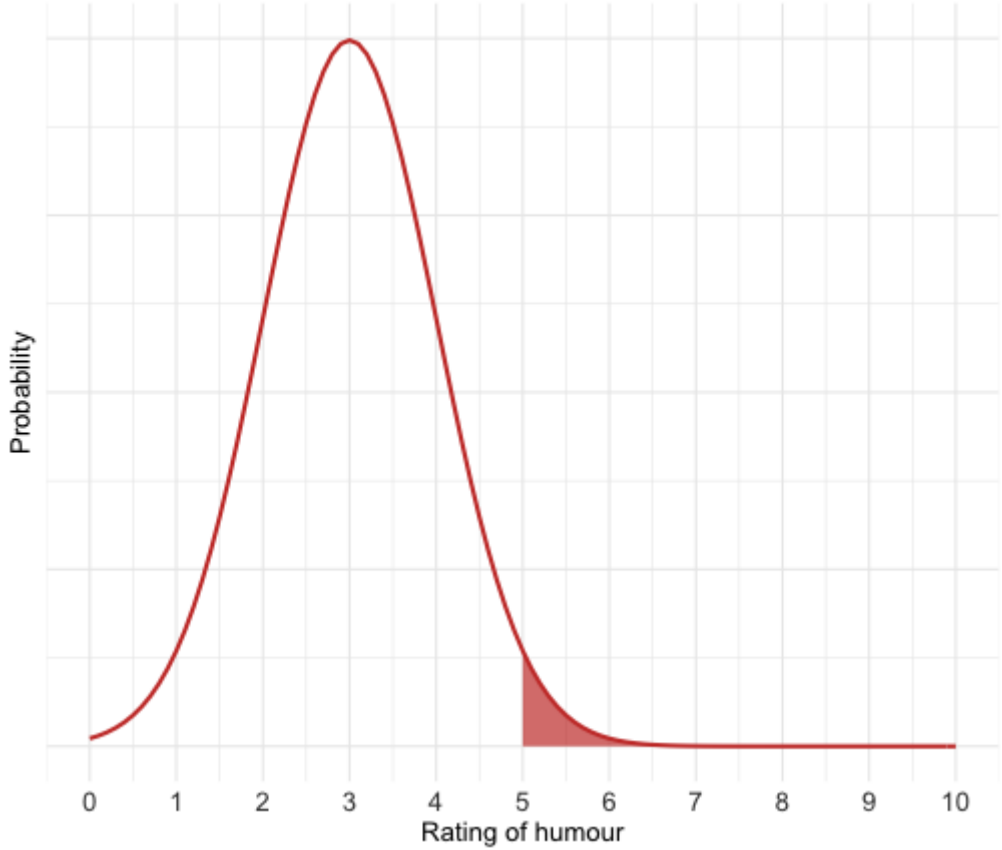
Humour rating = 5

What is a *p*-value?

Humour rating = 5

What is a *p*-value?

Humour rating = 5



The *p*-value

It is:

- The probability of getting a test statistic at least as big as the one you have observed given that the null hypothesis is true.

It is NOT:

- The probability of a chance result
- The probability that H_1 is true
- The probability that H_0 is true

Related constructs

Type I error

- Rejecting the null when it is true
- Believing in effects that don't exist
- Zach believing Alice wants to date him when in fact she doesn't (Awkward!)

Type II error

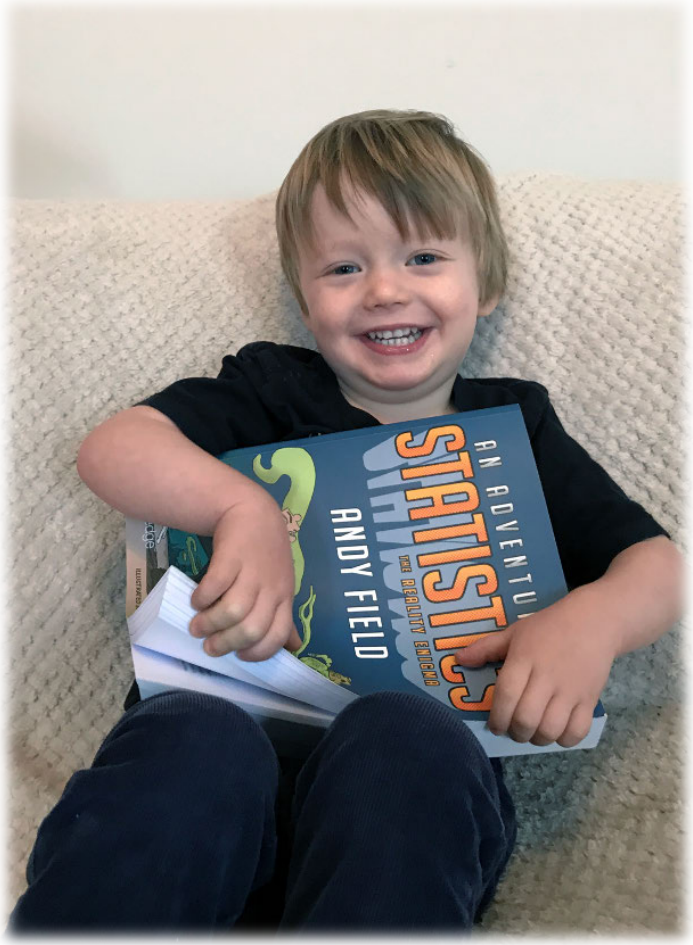
- Accepting the null when it's false
- Not believing in effects that do exist.
- Zach believing Alice doesn't want to date him when in fact she does. (Missed opportunity.)

Statistical power

- The probability of a test avoiding a Type II error
- The probability that a test detects an effect that is, in fact, true
- The probability of rejecting H_0 when H_1 is true

Teddy bear therapy

Control group



Problems with NHST

1. Tells us nothing about importance because *p* depends upon sample size.

Same effects, different *p*s

Study 1:

term	estimate	std.error	statistic	p.value
(Intercept)	12.89	0.525	24.565	0
groupBook	-5.00	0.742	-6.738	0

Study 2:

term	estimate	std.error	statistic	p.value
(Intercept)	12.8	2.054	6.233	0.000
groupBook	-5.0	2.904	-1.722	0.102

Zero effect (approx), significant *p*

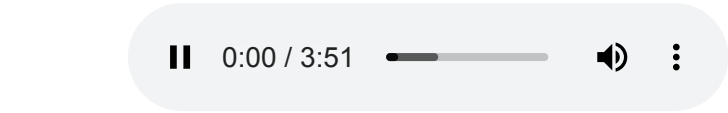
Study 3:

term	estimate	std.error	statistic	p.value
(Intercept)	12.113	0.018	660.082	0.000
groupBook	0.052	0.026	1.997	0.046

Problems with NHST

1. Tells us nothing about importance because *p* depends upon sample size
2. Provides little evidence about the null (or alternative) hypothesis
    - Assumes the null is true
    - $p > .05$ simply means the effect is not big enough to be found, not that it is 0
    - $p < .05$ means that the observed test statistic is unlikely given the null is true
    - Flawed logic

Logical flaw

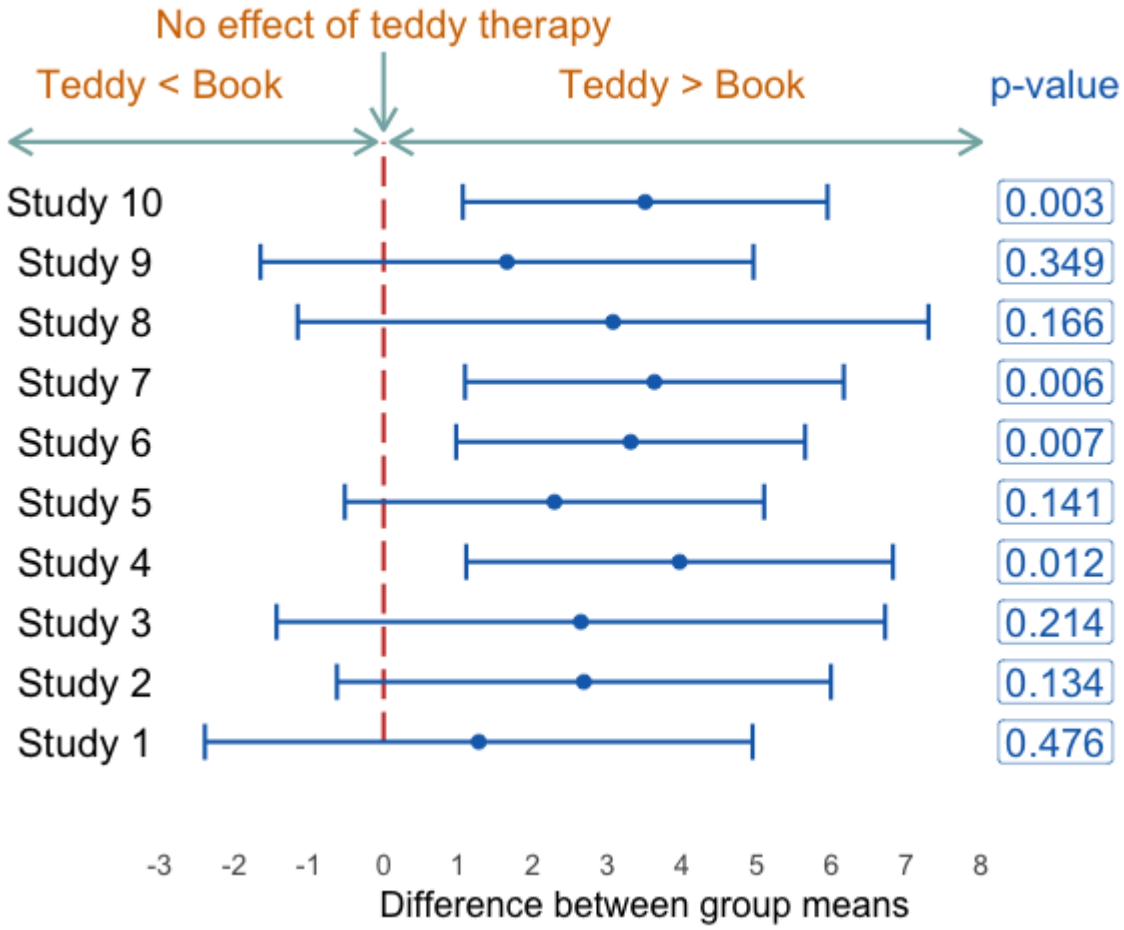


- If 'null hypothesis' is true, then it is highly unlikely to get this test statistic:
  - This test statistic has occurred.
  - Therefore, the null hypothesis is highly unlikely.'
- If 'person plays guitar' is true, then it is highly unlikely that he plays in Iron Maiden
  - This person plays in Iron Maiden
  - Therefore, 'person plays guitar' is highly unlikely.

Problems with NHST

1. Tells us nothing about importance because *p* depends upon sample size
2. Provides little evidence about the null (or alternative) hypothesis
3. Encourages all-or-nothing thinking

All or nothing thinking



Problems with NHST

1. Tells us nothing about importance because *p* depends upon sample size
2. Provides little evidence about the null (or alternative) hypothesis
3. Encourages all-or-nothing thinking
4. Based on long-run probabilities
    - *p* is the relative frequency of the observed test statistic relative to all test statistics from an infinite number of identical experiments with the exact same a priori sample size
    - The type I error rate is in a given study is either 0 or 1, but we don't know which

Estimation and effect sizes

Rather than obsessing over *p* we must also interpret the effects themselves Effect sizes

- Raw effect sizes (*b*)
- Standardized effect sizes
  - Cohen's *d*
  - Pearson's *r*
  - Standardized β

$$\hat{d} = \frac{\bar{X}_1 - \bar{X}_2}{s_p}$$

$$s_p = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}$$

Same effects, same effects!

Study 1: *n* = 200

term	estimate	std.error	statistic	p.value
(Intercept)	12.89	0.525	24.565	0
groupBook	-5.00	0.742	-6.738	0

Study 2: *n* = 20

term	estimate	std.error	statistic	p.value
(Intercept)	12.8	2.054	6.233	0.000
groupBook	-5.0	2.904	-1.722	0.102

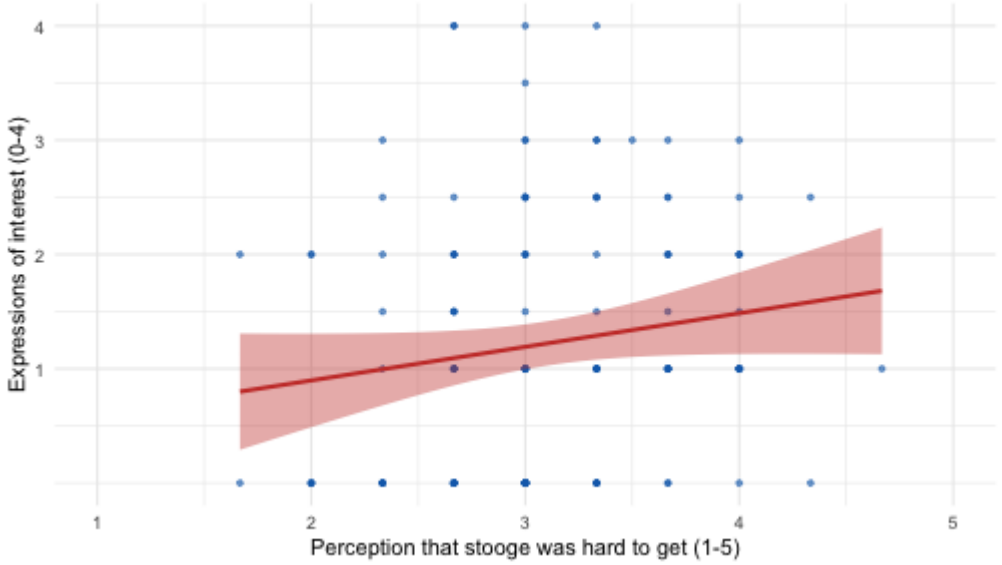
Zero effect (approx), zero effect

Study 3: *n* = 200,000

term	estimate	std.error	statistic	p.value
(Intercept)	12.113	0.018	660.082	0.000
groupBook	0.052	0.026	1.997	0.046

Raw effect size (*b*)

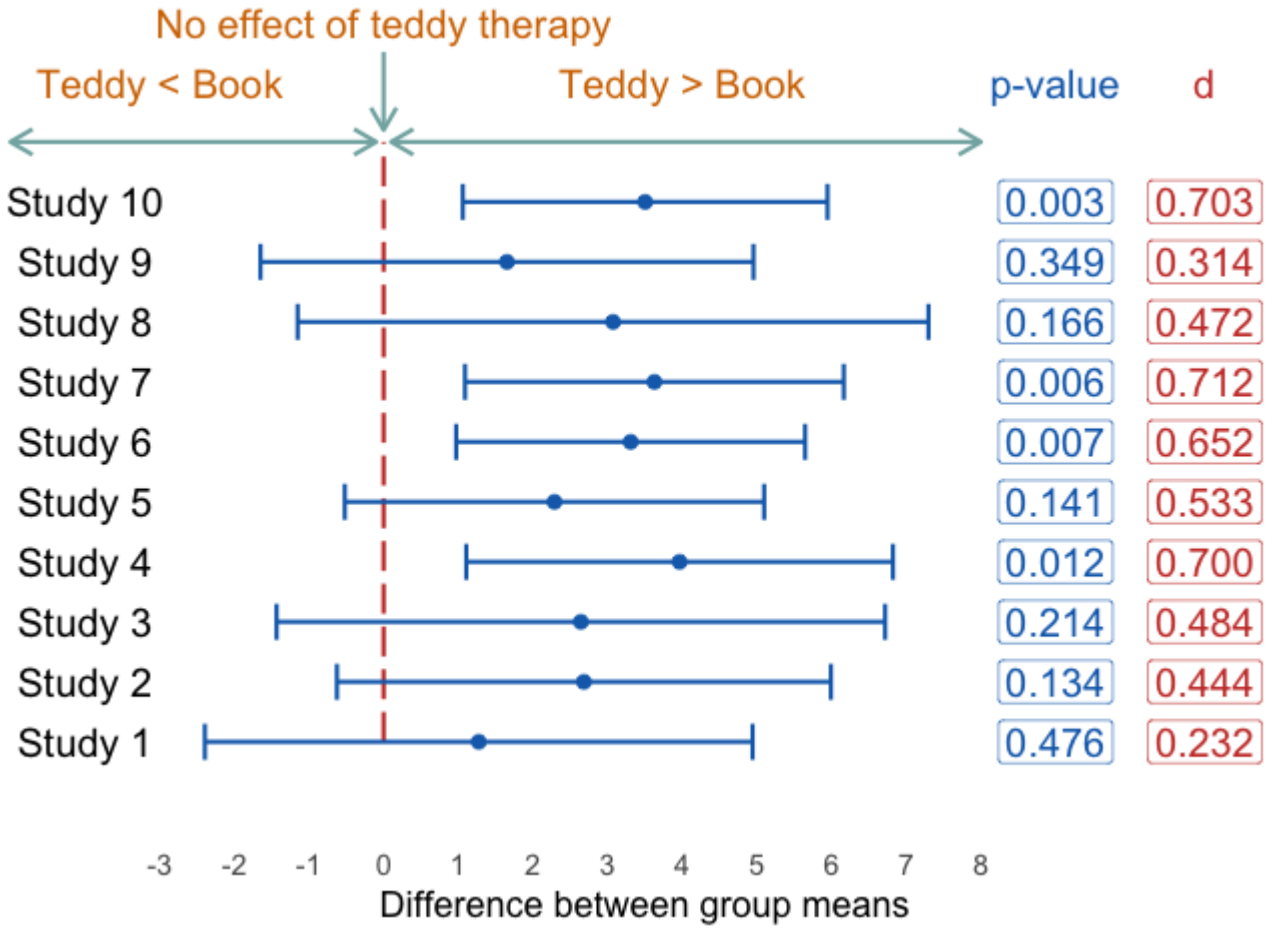
$$\hat{\text{interest}}_i = \hat{b}_0 + \hat{b}_1 \text{hard to get}_i + e_i$$



term	estimate	std.error	statistic	p.value
(Intercept)	0.310	0.524	0.591	0.556
hard_to_get	0.294	0.167	1.765	0.080

- As the perception that the other person was hard to get increased by 1 (on a scale from 1-5), 0.294 more expressions of interest were made.
- You'd need perceptions of 'hard to get' to increase by $\frac{1}{0.294} = 3.4$ on a 5-point scale to get 1 additional expression of interest.

All or nothing thinking



Summary

- Model parameters typically represent hypotheses
- We can 'test' these parameters/hypotheses by computing p
- The probability of observing a test statistic at least as large as the one you have given that the null hypothesis is true
- The process is problematic!
  - Address the wrong question
  - Depend on sample size
  - All-or-nothing thinking

Always interpret parameter estimates and effect sizes as well as *p*-values