

Power and effect sizes

2023/03/07

Replicability and reproducibility in psychology


Scientists Replicated 100 Psychology Studies, and Fewer Than Half Got the Same Results

The massive project shows that reproducibility problem plagues top scientific journals

A question of trust: fixing the replication crisis

The crisis of non-replications in experimental social psychology is a crisis of trust. What's the solution?

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson & Marcus R. Munafò 

Nature Reviews Neuroscience **14**, 365–376 (2013) | [Download Citation](#) 

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • <https://doi.org/10.1371/journal.pmed.0020124>

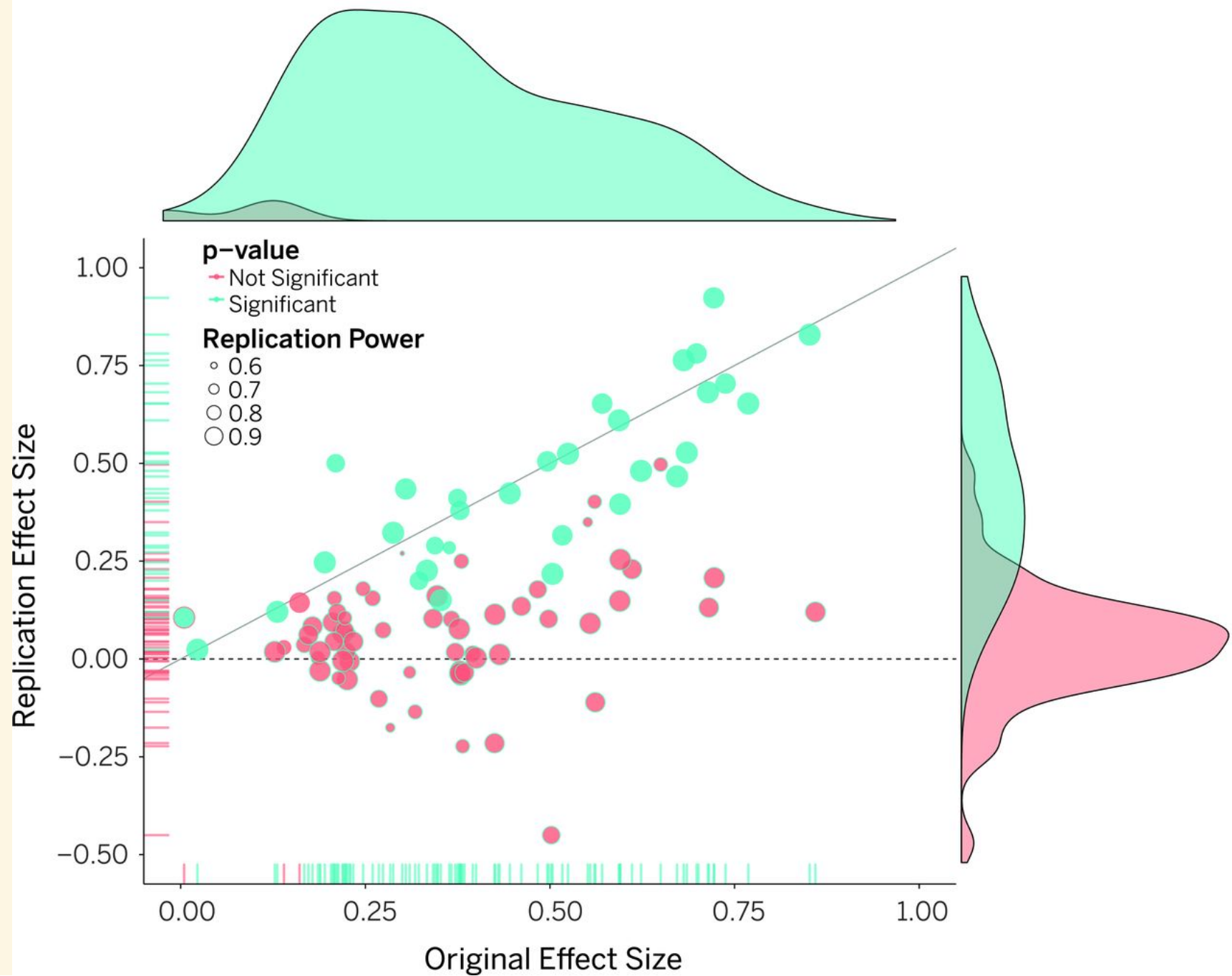
RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration^{*,†}

⁺ See all authors and affiliations

Science 28 Aug 2015;
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716



Null Hypothesis Significance Testing (again)

Setting up our studies

When we embark on a programme on research, we begin by identifying a **research question**.

- Do people who have been a victim of crime express higher fear of crime?
- Are people faster at saying the names of colours when the names are written in the same colour?

Null Hypothesis Significance Testing (NHST) (again)

Our typical way of answering questions such as these is to set up a Null Hypothesis as an alternative.

Typically, this hypothesis is that of **zero** effect.

We then pose the question:

If there were no difference or relationship between these two variables in the population, how likely is it that we would observe this data in our sample?

The process of NHST

```
a <- rnorm(50)
b <- rnorm(50, mean = 1)
t.test(a, b)
```

```
##
##      Welch Two Sample t-test
##
## data:  a and b
## t = -4.531, df = 98, p-value = 0.00001658
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.2478334 -0.4877117
## sample estimates:
##  mean of x  mean of y
## 0.04359069 0.91136323
```


Interpreting p-values

Significance versus non-significance

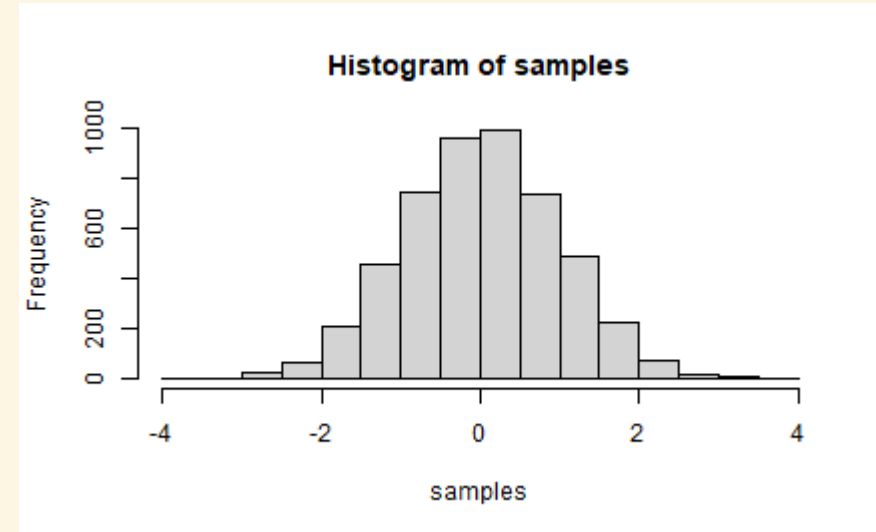
In our field, we typically set our significance criterion - *alpha*, or α - at .05.

If the p-value of our test falls **below** this threshold, we say we have a *significant* result, and we get all excited and break out the bubbly. 🍾, 🍾, 🍾

If the p-value falls **above** this threshold, we say we have a *non-significant* result, and we get extremely upset. 😭 😞

(both these reactions are a little bit over-the-top)

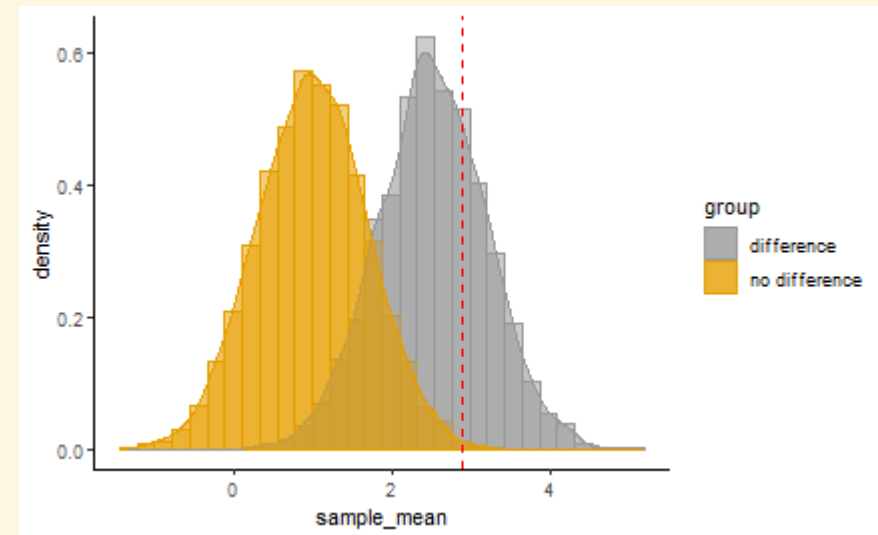
In other words, assuming the null hypothesis is true (no difference): out of 1000s of samples, each with a distribution, mean and statistic, our sample is only one. If it is a sample with a mean and statistic in the 5% of the tails of the distribution of all samples, then our result is considered significant.



Another way to look at it

Your sample may be extreme *assuming the null hypothesis is true*, but it might be perfectly average *assuming the null hypothesis is false*.

```
## `stat_bin()` using `bins = 30`. Pick better value
```



Type I and Type II errors

Under NHST, we are trying to decide whether our statistical results match reality. There are **two** basic types of error we can make.

	Null hypothesis is false	Null hypothesis is true
$p \leq .05$	True positive	False positive
$p > .05$	False negative	True negative

False positives - a significant result when there is *no* real effect - are called **Type I errors**.

False negatives - a non-significant result when there *is* a real effect - are called **Type II errors**.

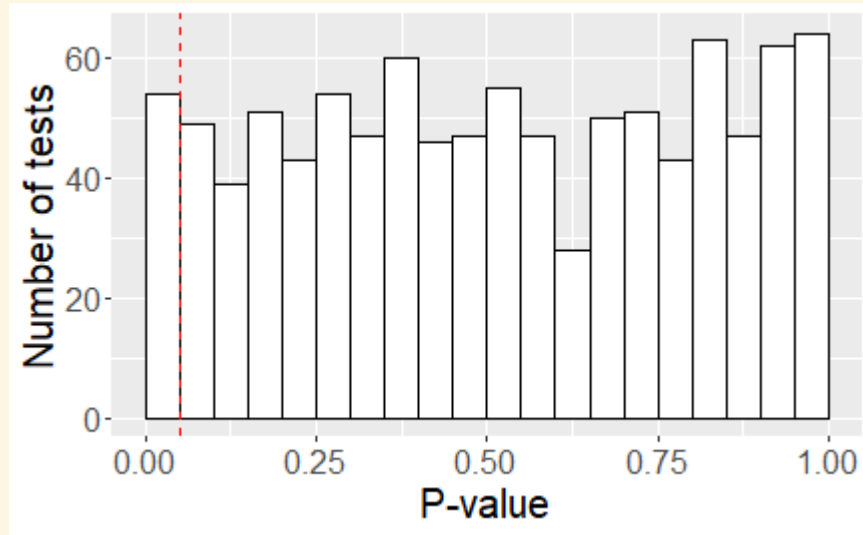
The false-positive rate

When the null hypothesis is **true**, *any specific p-value is as likely as any other*.

So if the null hypothesis is *true*, and there is *no* real effect, we will get a **significant result** 5% of the time - **1 in every 20** repeats.

In other words, setting α at .05 means we accept a **false positive rate** of 5%.

The false-positive rate



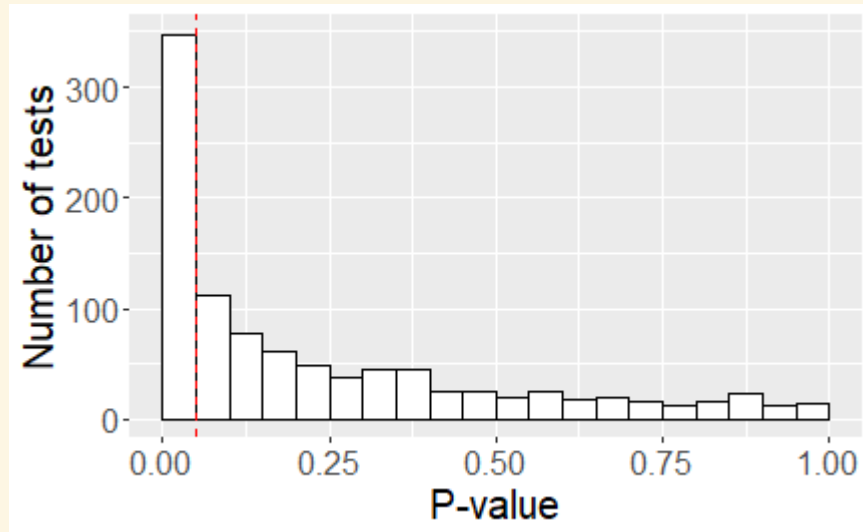
Here, I simulated *normally distributed* data from 100 participants with a mean of zero, one thousand times.

Each time, I tested whether that data was significantly different from zero.

Approx 5% of the p-values of these 1000 tests were $< .05$.

The false-negative rate

When the null hypothesis is false, p-values lower than our threshold become *more likely*. But it's still not certain we'll get a *significant effect*.

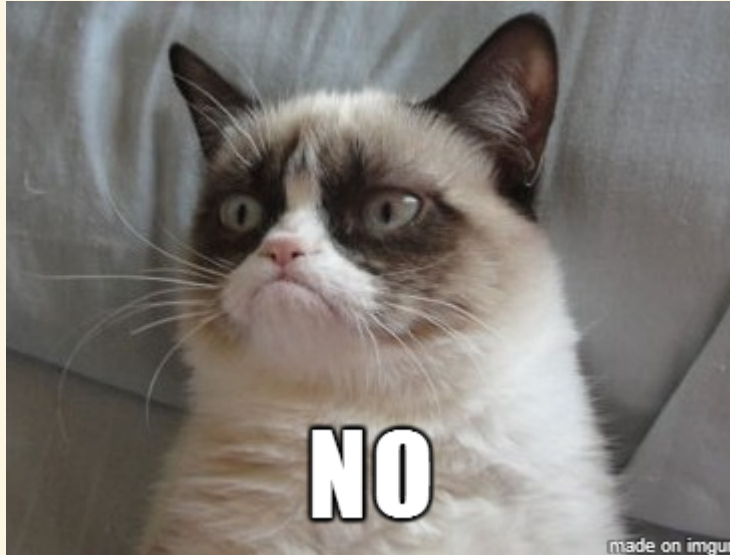


Here, I simulated data from 100 participants with a mean of 0.15, 1000 times, each time testing whether the data differs from zero.

Approx 320 tests out of 1000 were significant - a true-positive rate of .32, and a false-negative rate of .68 - 68%.

What does the p-value tell us?

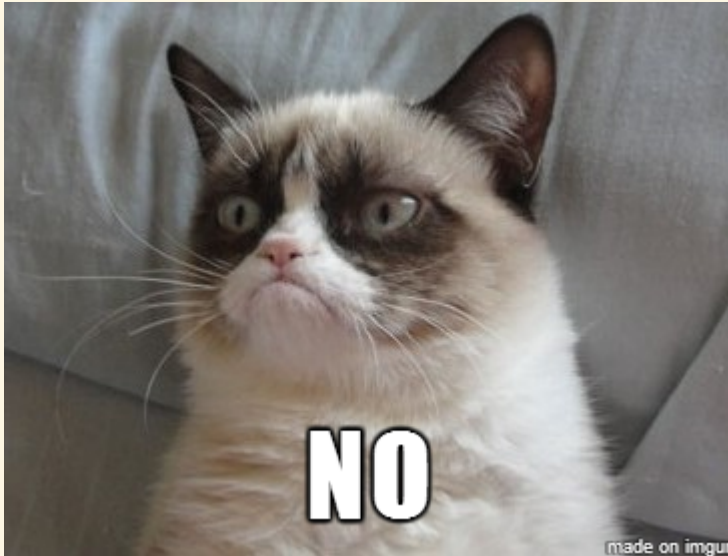
Does a significant p-value tell us how likely it is our experimental hypothesis - that there is really an effect - is *true*?



- 1) Even if there was an effect, it may not be for the reason we think.
- 2) A significant finding may also be a **false positive**.

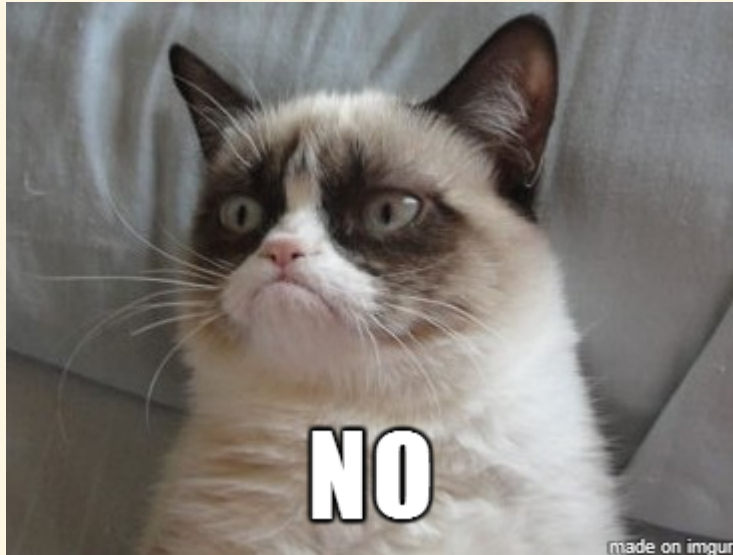
What does the p-value tell us?

Does a non-significant p-value tell us that there is no effect, or that our hypothesis was false?



- 1) A non-significant effect only tells us that we **failed to reject** the null hypothesis.
- 2) A non-significant effect can be a **false negative**.

Ok, does the p-value tell us how big the effect is?



P-values tell us *absolutely nothing* about the size of the effect.

All they tell us that the data is *unlikely* if the null hypothesis is true, not whether the effect is large or small. Tiny effects can have p-values just as tiny as large effects can.

Right. So does it tell us how important the effect is?



A tiny p-value does not mean the effect is any way *important*.

Essentially meaningless effects can have very small p-values.

So what *does* the p-value tell us?



**The probability that
the alternative
hypothesis is true**



**The probability that
you would have
observed this data if
the null hypothesis
is true**

Effect sizes

Effect sizes

p-values tell us how likely it was that the data we observed would happen if the null hypothesis were true. But to understand what our tests are really telling us, we need to look at *effect sizes*.

Effect sizes:

- 1) Communicate the practical significance of a result.
- 2) Enable comparison across different studies and different scales.
- 3) Allow you to perform *power analysis*.

Unstandardized effect sizes

Unstandardized effect sizes are effects on the *measurement scale*.

Easy to understand and interpret, so they're particularly helpful for understanding the real-world relevance of statistical effects.

e.g. Usain Bolt ran .12 seconds faster than Yohan Blake.



A screenshot of a sports broadcast showing the final results of the Men's 100m race. The table lists the top 8 finishers with their rank, country, name, and time. Usain Bolt is the winner with a time of 9.63 seconds, marked as a World Record (WR). The wind speed is noted as +1.5m/s.

MEN'S 100M		WIND +1.5M/S
RESULT - FINAL		
1	JAM 	USAIN BOLT WR 9.63
2	JAM 	YOHAN BLAKE 9.75
3	USA 	JUSTIN GATLIN 9.79
4	USA 	TYSON GAY 9.80
5	USA 	RYAN BAILEY 9.88
6	NED 	CHURANDY MARTINA 9.94
7	TRI 	RICHARD THOMPSON 9.98
8	JAM 	ASAFA POWELL 11.99

Standardized effect sizes

Standardized effect sizes place effects on a common scale - they're helpful when the dependent measure is measured in different units across different studies.

1	2	3	4	5
Definitely agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Definitely disagree

1	2	3	4	5	6	7
Definitely agree	Somewhat agree	Slightly agree	Neither agree nor disagree	Slightly disagree	Somewhat disagree	Definitely disagree

Standardized effect sizes

There are two major families of standardized effect size:

Measure	<i>Cohen's d</i>	<i>r</i>
Definition	Size of mean differences	Strength of association
Statistical tests	t-tests	correlation, ANOVA, regression
Variations	d , d_z , Hedge's g	r , r^2 , η^2 , ω^2

Standardized mean differences

Cohen's d

Cohen's d ranges from 0 to ∞ (infinity!)

The basic calculation is pretty simple - it's the *mean difference* divided by the *standard deviation pooled across conditions*.

$$\frac{\mu_1 - \mu_2}{SD_{pooled}}$$

All variations of Cohen's d for different types of design (e.g. d_z for within-subjects designs) are variants of this formula.

Interpreting Cohen's d

The website linked here provides a great interactive tool to visualize what Cohen's d is [RPsychologist Cohen's d](#)

Pooled standard deviation

The pooled standard deviation is calculated using this formula:

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}}$$

```
sqrt((sd(a)^2 + sd(b)^2) / 2)
```

```
## [1] 0.9575889
```

Quick example

```
t.test(a, b)
```

```
##  
##      Welch Two Sample t-test  
##  
## data:  a and b  
## t = -4.531, df = 98, p-value = 0.00001658  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
##  -1.2478334 -0.4877117  
## sample estimates:  
##  mean of x  mean of y  
## 0.04359069 0.91136323
```

```
(mean(a) - mean(b)) / sqrt((sd(a)^2 + sd(b)^2) / 2)
```

```
## [1] -0.9062057
```


The `effectsize` package

A simpler way to calculate Cohen's d is to use the `cohens_d` function from the `effectsize` package.

```
library(effectsize)
cohens_d(a, b)
```

```
## Cohen's d |          95% CI
## -----
## -0.91      | [-1.32, -0.49]
##
## - Estimated using pooled SD.
```

This also gives us *confidence intervals* around the effect size - a helpful reminder that these are *estimates*.

The Facebook study

Experimental evidence of massive-scale emotional contagion through social networks



Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

PNAS June 17, 2014 111 (24) 8788-8790; published ahead of print June 2, 2014

<https://doi.org/10.1073/pnas.1320040111>

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

The Facebook study

When positive posts were reduced in the News Feed, the percentage of positive words in people's status updates decreased by $B = -0.1\%$ compared with control [$t(310,044) = -5.63$, $P < 0.001$, Cohen's $d = 0.02$], whereas the percentage of words that were negative increased by $B = 0.04\%$ ($t = 2.71$, $P = 0.007$, $d = 0.001$).

The Facebook study

When positive posts were reduced in the News Feed, the percentage of positive words in people's status updates decreased by $B = -0.1\%$ compared with control [$t(310,044) = -5.63$, $P < 0.001$, Cohen's $d = 0.02$], whereas the percentage of words that were negative increased by $B = 0.04\%$ ($t = 2.71$, $P = 0.007$, $d = 0.001$).

$P = .007$, $d = .001$. This is an absolutely **tiny** effect size.

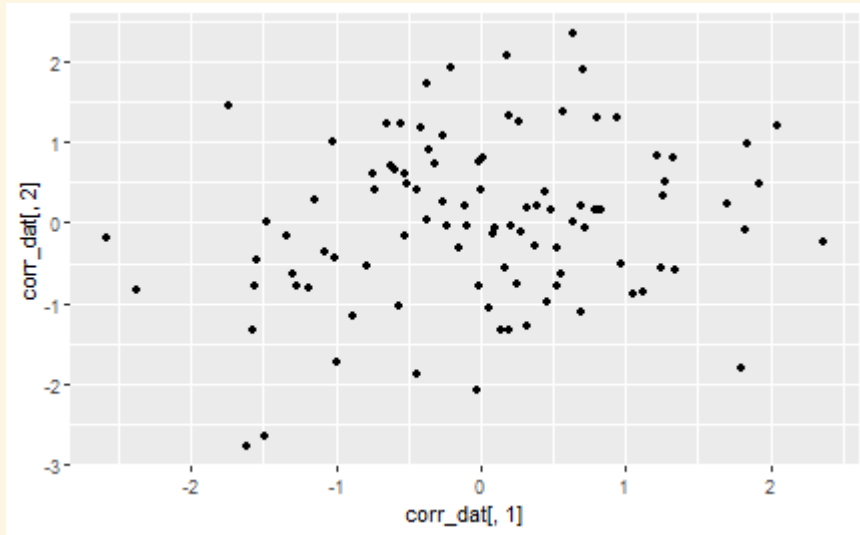
Approx 1 extra negative word for every 3570 words typed.

```
effectsize::interpret_cohens_d(.001)
```

```
## [1] "very small"  
## (Rules: cohen1988)
```

Strength of associations

Guess the correlation

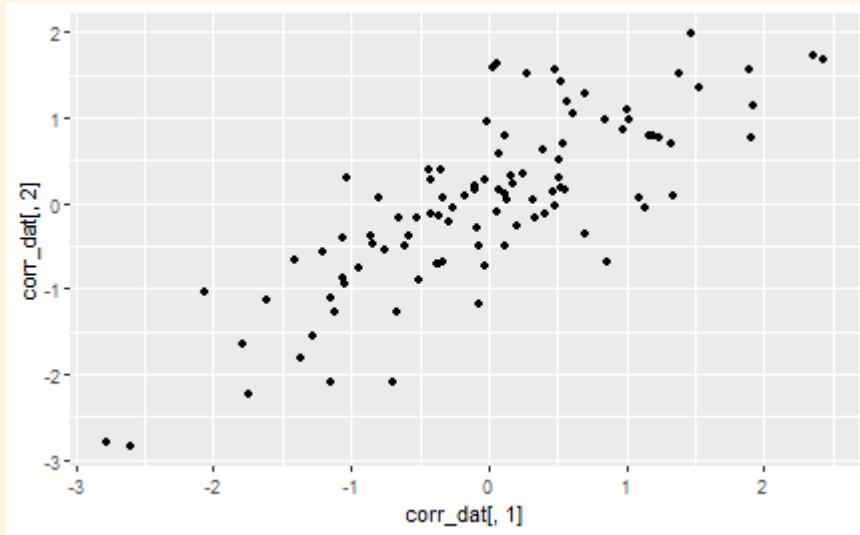


$r = .2$

```
interpret_r(.2)
```

```
## [1] "medium"  
## (Rules: funder2019)
```

Guess the correlation



$r = .8$

```
interpret_r(.8)
```

```
## [1] "very large"  
## (Rules: funder2019)
```

Guess the correlation

RPsychologist correlation visualizations <https://rpsychologist.com/d3/correlation/guessthecorrelation.com>

Converting from r to d

Although the scale is different, r and d are closely related.

The formula below can be used to convert between them.

$$r = \frac{d}{\sqrt{d_s^2 + \frac{N^2 - 2N}{n_1 n_2}}}$$

```
r_to_d(.8) # from the effectsize package again!
```

```
## [1] 2.666667
```

```
interpret_cohens_d(r_to_d(.8))
```

```
## [1] "large"
```

```
## (Rules: cohen1988)
```

Proportion of variance explained

For regressions, and ANOVAs, we don't use the correlation coefficient on its own. Rather, we use one of the various *proportion of variance explained* effect sizes.

Symbol	name
r^2	r-squared
η^2	eta-squared
η_p^2	partial eta-squared
η_g^2	generalized eta-squared

Proportion of variance explained

Every one of these measures is a variation on the same thing: how much does the relationship between our variables reduce the error of our model.

Remember the formula for R-squared (r^2)?

$$r^2 = \frac{SS_m}{SS_t}$$

It's the ratio of the variance explained by the model to the total variance in the model. Thus, it's the *percentage of variance explained by the model*.

Reporting effect sizes in your results

Reporting effect sizes in your results

When reporting your statistical results, it's best practice (though not always followed...) to report both standardized **and** unstandardized effect sizes.

- 1) Reporting **unstandardized** effect sizes helps understand how big the effect is in *real terms*.
- 2) Reporting **standardized** effect sizes helps compare the effect to effects in different studies and on different scales.
- 3) Always interpret the effect sizes. Take care of the difference between *statistical* and *practical* significance.
- 4) Many of the standardized effect sizes are somewhat *interchangeable*, but always try to report the right one for your test (e.g. Cohen's d for t-tests, r^2 / *adjusted* $- r^2$ for regression)

Which standardized effect should you report?

Statistical test	Standardized effect size
t-test	d (between), d_z (within), d_s
ANOVA	η^2 (one-way), η_p^2 (factorial), η_g^2 (Any)
Correlation	r
Linear (simple or multiple) regression	r^2 , <i>adjusted</i> $- r^2$

Note - not every possible test and every possible effect size can fit! We'll cover some more later in the course...

Rules of thumb for interpreting standardized effect sizes

Effect size	small	medium	large
d	0.2	0.5	0.8
r	.1	.3	.5
r^2	.1	.19	.25
η^2	.01	.06	.14
η_p^2	.01	.09	.25
η_g^2	.02	.13	.26

These are *guidelines*, not *rules*. (you can also try the `interpret` functions from `effectsize`)

Designing for statistical power

Statistical power

Statistical power is the inverse of the **false-negative rate**.

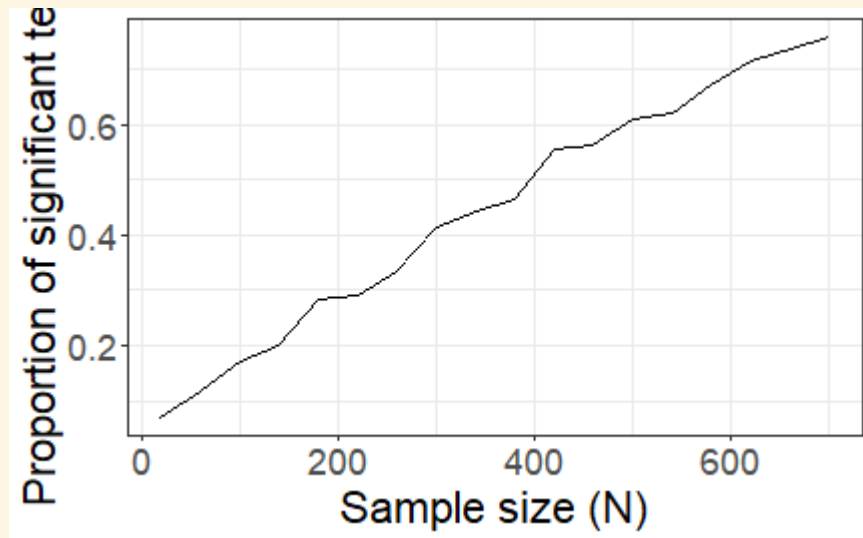
Also termed *beta*, or β , power is the probability of getting a significant result with a given *sample size*, *statistical test*, and *effect size*.

By convention, psychological studies aim for **80% power** - we accept a 20% false-negative rate!

In this treatment, the only specification for power is .80 (so $\beta = .20$), a convention proposed for general use. (SPABS provides for 11 levels of power in most of its *N* tables.) A materially smaller value than .80 would incur too great a risk of a Type II error. A materially larger value would result in a demand for *N* that is likely to exceed the investigator's resources. Taken with the conventional $\alpha = .05$, power of .80 results in a $\beta:\alpha$ ratio of 4:1 (.20 to .05) of the two kinds of risks. (See SPABS, pp. 53--56.)

Statistical power and sample size

Sample size is an important factor determining statistical power:



Here I simulate the effects of **increasing sample size** on statistical power.

The **effect size** stays constant - there's a 0.1 difference between the means of each group.

Estimating sample size

We can estimate the *required sample size* of a specific statistical test if we know the desired *power* and the *expected effect size*. The hardest part of this is typically **knowing what effect size you expect**.

```
library(pwr)
pwr.t.test(power = .8, # this is a proportion
           d = .3, # Cohen's d
           type = "one.sample") # we are doing a one-sample t.test

##
##      One-sample t test power calculation
##
##              n = 89.14938
##              d = 0.3
##      sig.level = 0.05
##              power = 0.8
##      alternative = two.sided
```

Why don't we aim for 100% power?

```
pwr.t.test(power = 1, # this is a proportion
           d = .3, # Cohen's d
           type = "one.sample") # we are doing a one-sample t.test
```

```
##
##      One-sample t test power calculation
##
##              n = 10000000000
##              d = 0.3
##      sig.level = 0.05
##              power = 1
##      alternative = two.sided
```

100% power is often *just a little* impractical.

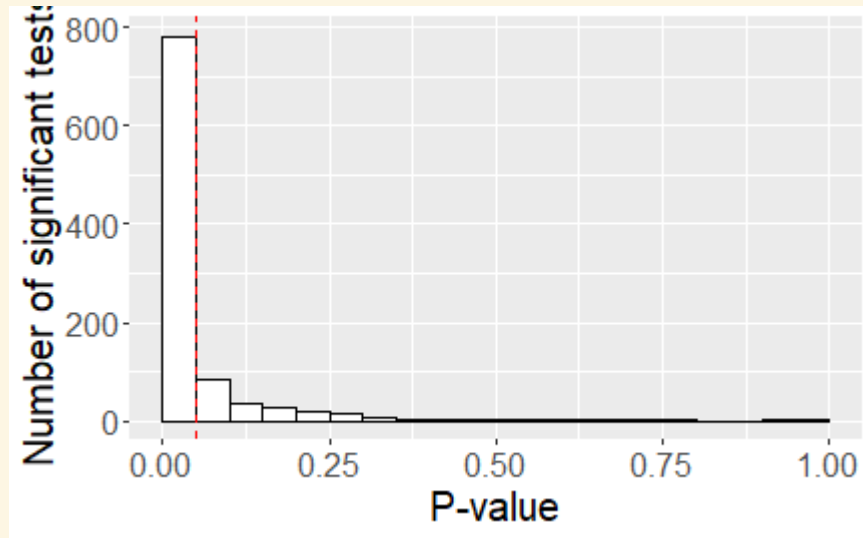
Estimating effect size

We can estimate the *effect size we'd have power to detect* if we know the *power* and the *sample size*. Suppose we *know* we'd have 100 participants - we can't get more, and we won't get **fewer**.

```
pwr.t.test(power = .8, # this is a proportion
           n = 100, # Cohen's d
           type = "one.sample") # we are doing a one-sample t.test
```

```
##
##      One-sample t test power calculation
##
##              n = 100
##              d = 0.2829005
##      sig.level = 0.05
##              power = 0.8
##      alternative = two.sided
```

Statistical power and effect size

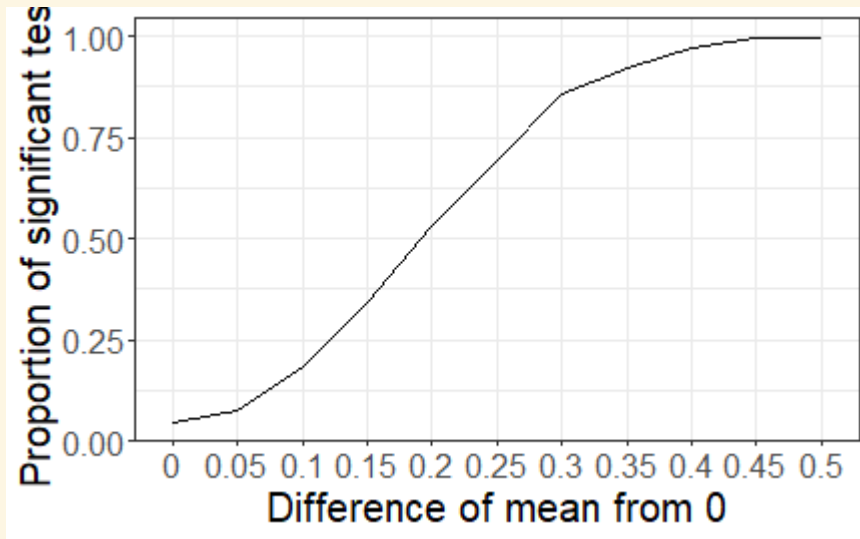


Here, I simulate data with a mean of 0.283, 1000 times, and test whether it differs from zero.

The *sample size* remains constant at 100 participants.

Approx 800 tests are significant - a true-positive rate of .80, and thus a statistical power - β - of 80%.

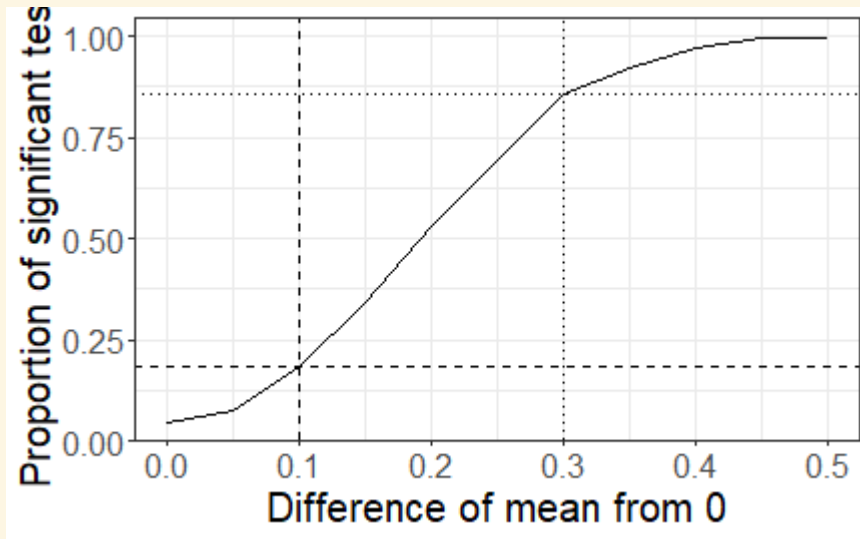
Statistical power and effect size



In this simulation, the general design of the study stays the same - there are 100 participants, we test against zero with a t-test.

As the effect size increases, the **power** of the study increases.

Statistical power and effect sizes



The study has approximately 18.3% power to detect a 0.1 difference in means.

But the study also has approximately 85.5% power to detect a 0.3 difference in means.

Studies have a **power curve**, not a single *power*.

Estimating power

We can estimate the *power* of a specific statistical test if we know the *sample size* and the *effect size*.

```
pwr.t.test(n = 100, # this is a proportion
           d = .3, # Cohen's d
           type = "one.sample") # we are doing a one-sample t.test
```

```
##
##      One-sample t test power calculation
##
##              n = 100
##              d = 0.3
##      sig.level = 0.05
##              power = 0.8439471
##      alternative = two.sided
```

... but this is not generally what you want to do. After you've done the study, it's too late. Before you run the study, you want to estimate sample size to know how many people you need.

Critiquing the statistical power of a study

A common critique of studies is that their sample size is **too low**, and thus that they lack statistical power.

But any given study always has 80% power to detect **something**: power is a **curve**.

A better critique is that a study has insufficient sample size to reliably detect a **meaningful, important effect**.

Further (suggested, not required) reading

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00863/full>

Perugini, M., Gallucci, M., & Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, 31(1), 20. DOI: <http://doi.org/10.5334/irsp.181>

Next session

Next week, the topic is **Multiple regression** and looking at **logistic regression and Generalized linear models**.

Chapter 8 of Field et al., Discovering Statistics Using R.