Notes on Statistical Power

Ben Prytherch, Statistics Department, CSU

“The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.”

- John Tukey

Overview

Type I and Type II errors

Definition of power

What affects power?

How can I calculate power?

What are the consequences of low power?

Visualizing power

Deciding to reject or fail to reject H_0

- When conduction a statistical hypothesis test, we define a null hypothesis, e.g.

$$H_0: \mu_1 - \mu_2 = 0$$

- We then choose if we will “reject” or “fail to reject” this null

- There most popular methods for making this decision is to calculate a p-value and compare it to a “level of significance” (α), typically 0.05

$$If\ p < 0.05, reject\ H_0$$

$$If\ p > 0.05, fail\ to\ reject\ H_0$$

~~$If\ p = 0.05, use\ more\ precision\ until\ it\ doesn't$~~

Disclaimer

- I think hypothesis testing is overused, typically unnecessary, and causes more trouble that it’s worth.

- There is a way of reinterpreting “power” in terms of the width of confidence intervals, which I prefer.

- But, hypothesis testing is wildly popular and so it’s important to understand power.

Type I and Type II Errors

- If we reject H_0 when H_0 is true, we have committed a Type I error.

- If we fail to reject H_0 when H_0 is false, we have committed a Type II error.

“The truth”

“The statistical decision”

	$H_0\ is\ false$	$H_0\ is\ true$
$Reject\ H_0$	👍	Type I error
$FTR\ H_0$	Type II error	👍

Type I and Type II error trade off

- We can reduce Type I errors by lowering α, but doing this would increase Type II errors.

- This is because lowering α makes it harder to reject H_0, which is good when H_0 is true but bad when H_0 is false.

- Similarly, we can reduce Type II errors by increasing α, but doing this would increase Type I errors.

Power

- Power is the probability of **not** committing a Type II error.

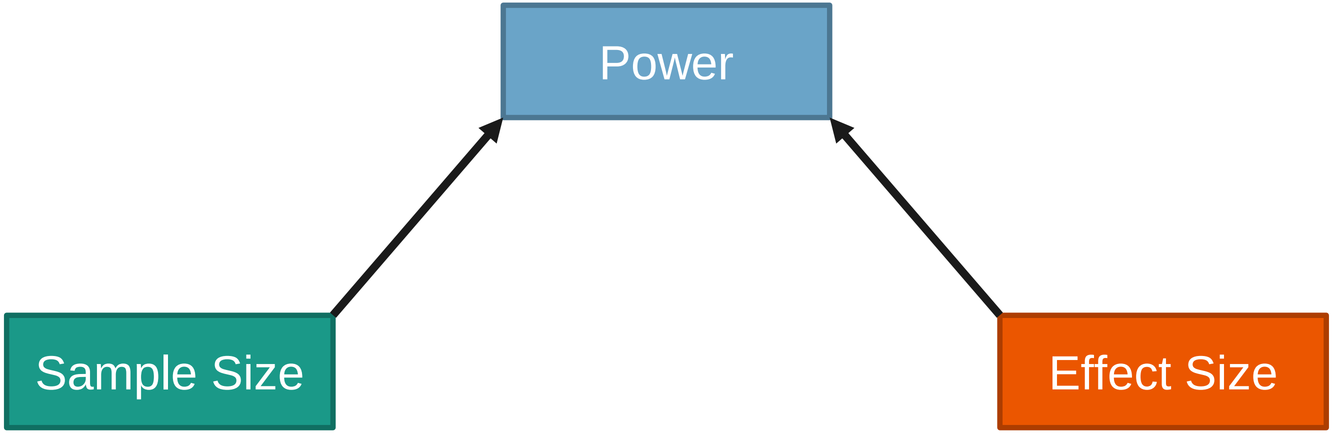
$$\boldsymbol{Power} = P(Reject\ H_0|H_0\ is\ false)$$

or

$$Power = 1 - P(Type\ II\ Error)$$

What determines power?

Power is determined by sample size and effect size



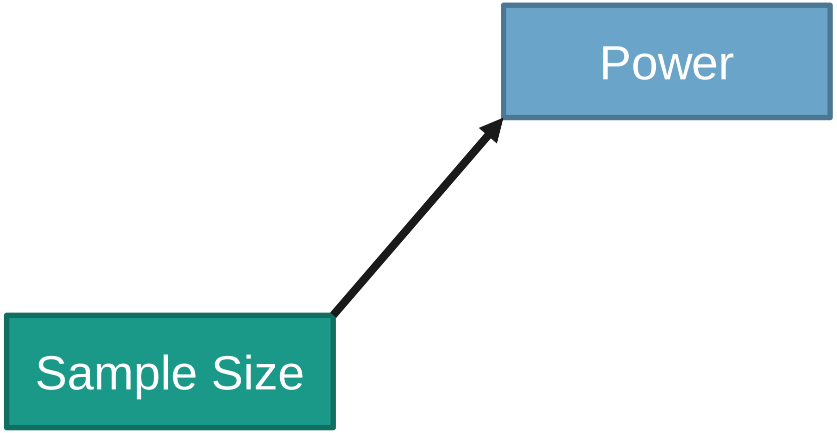
Power

Sample Size

Effect Size

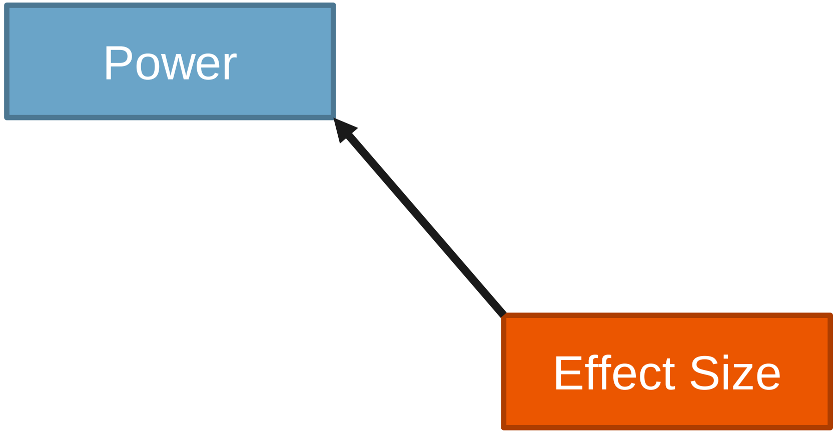
Sample Size and Power

As sample size increases, power increases.



Effect Size and Power

As effect size increases, power increases



“Effect size”

- Suppose we compute a difference in means between a treatment group and a control group. A large difference in means could be interpreted as a large “effect” for the treatment.

- However, a seemingly large difference in means might not seem large if the data are highly dispersed.

- Example: if we are comparing one mile running times among elite athletes, a difference of 10 seconds may be considered large. But among casual joggers, it may be considered small.

Quantifying effect size: Cohen's d

- For a two-sample test, a popular measure of effect size is Cohen's d:

$$d = \frac{|\mu_1 - \mu_2|}{\sigma_p}$$

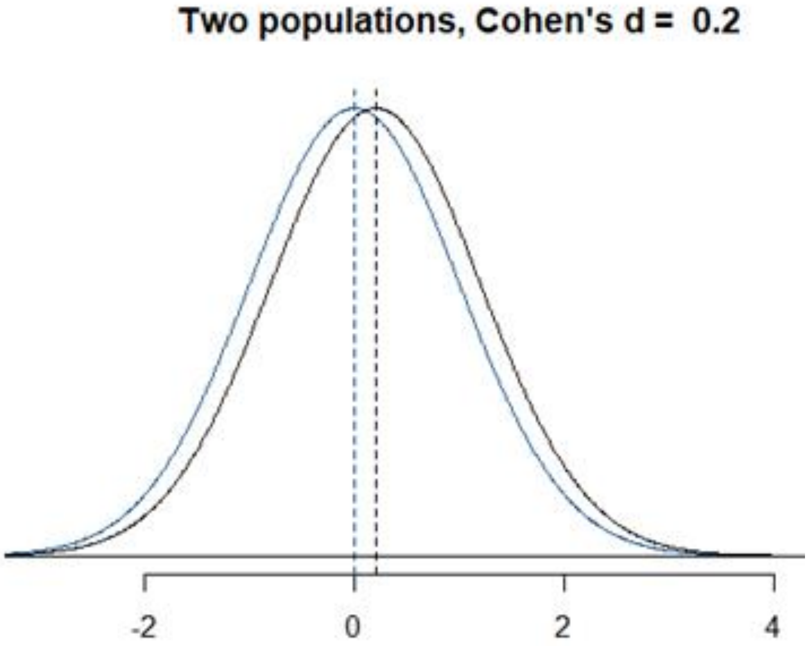
- Where μ_1 and μ_2 are the population means for groups one and two, and σ_p is the combined (or "pooled") standard deviation for both groups. Don't worry how σ_p is computed.

- So, Cohen's d tells us how far apart two means are from one another, in terms of the number of standard deviations.

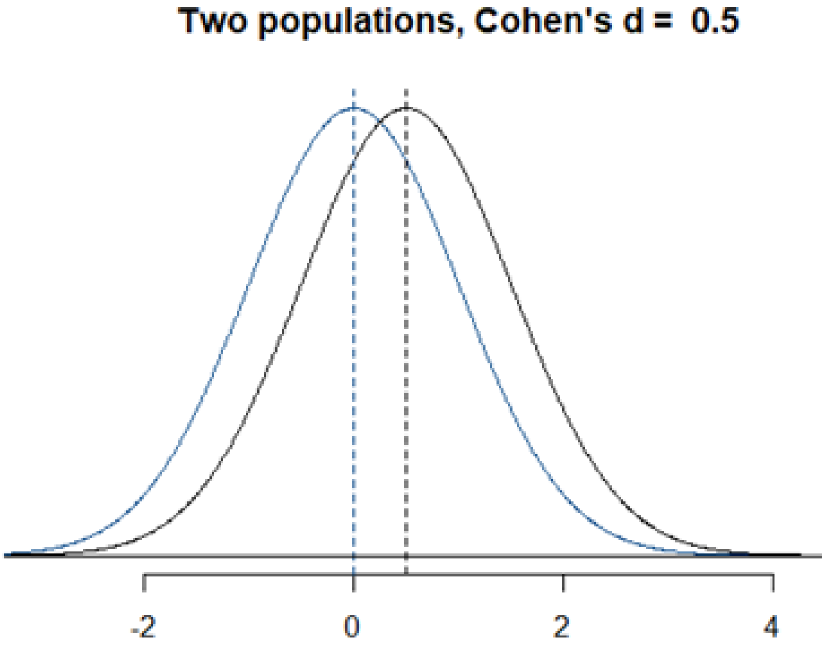
“Large” vs. “medium” vs. “small” effect sizes

- What counts as a “large” or “small” effect size is subjective, but some common guidelines are:
  - ➢$d < 0.2$ is “small”
  - ➢$0.2 < d < 0.8$ is “medium”
  - ➢$d > 0.8$ is “large”
- As with all arbitrary cutoffs, these should not be taken too seriously. Nothing magical happens when d changes from 0.79 to 0.81.
- What counts as “large” or “small” will also depend on the problem at hand. Maybe I shouldn’t have included this slide…

Visualizing Cohen’s d

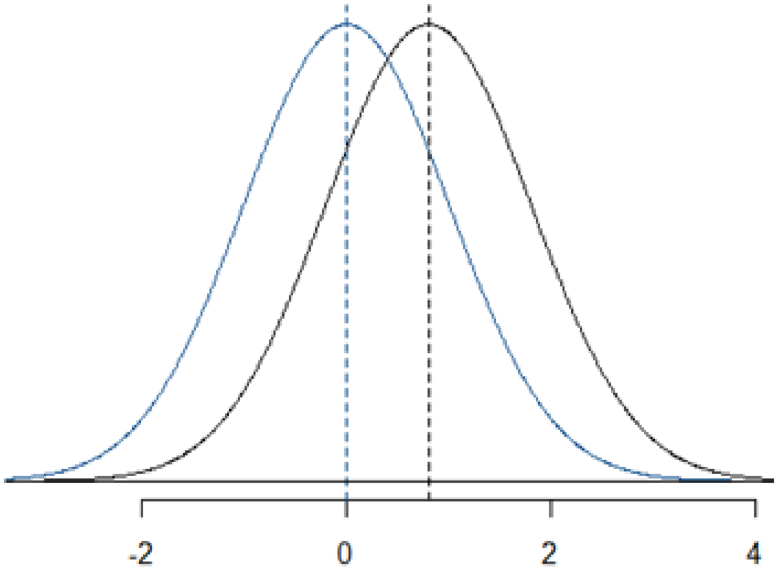


Two populations, Cohen's d = 0.2

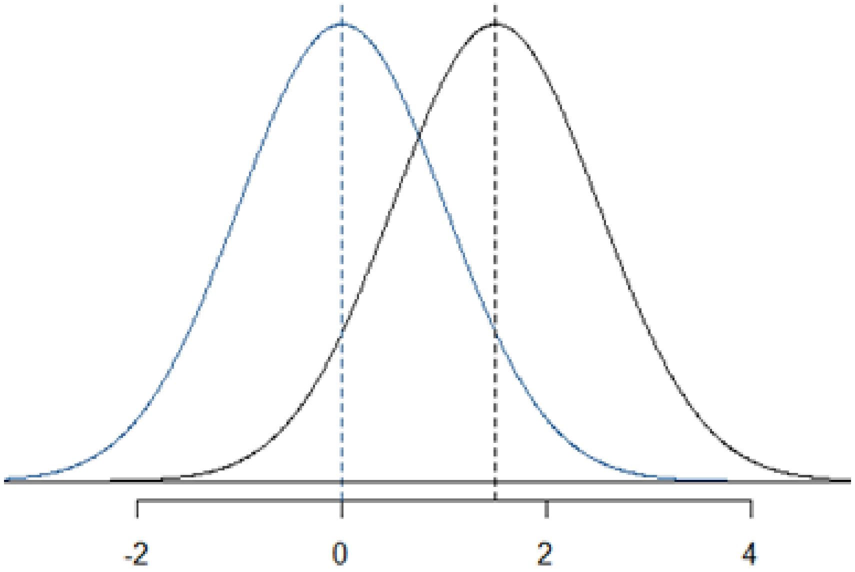


Two populations, Cohen's d = 0.5

Visualizing Cohen’s d



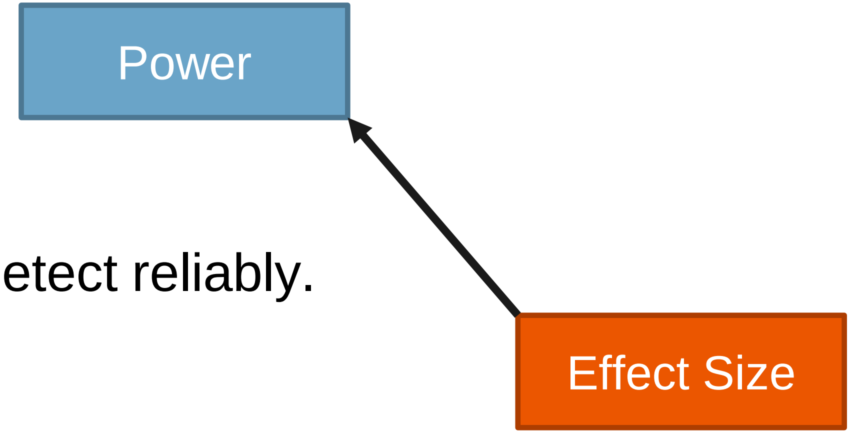
Two populations, Cohen's d = 0.8



Two populations, Cohen's d = 1.5

Effect Size and Power

As effect size increases, power increases.

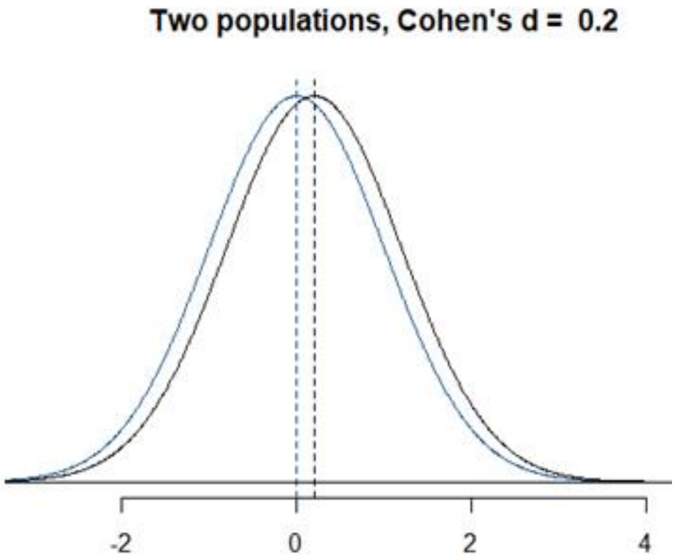


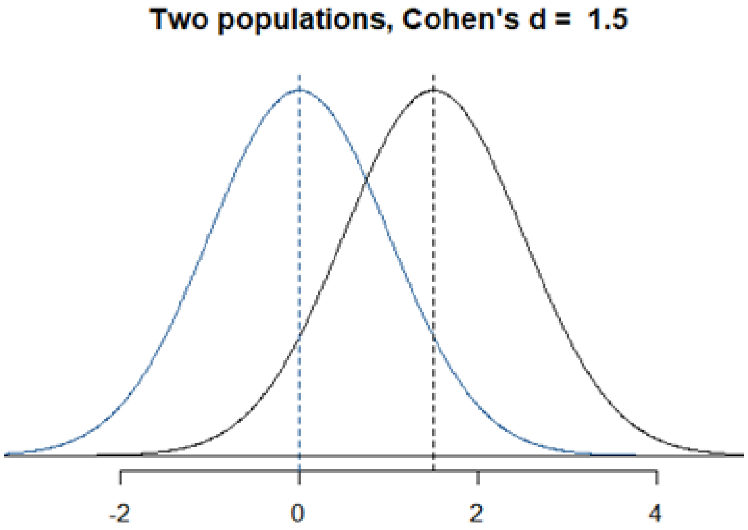
Larger effects are easier to detect reliably.

Effect Size and Power

We might miss this:

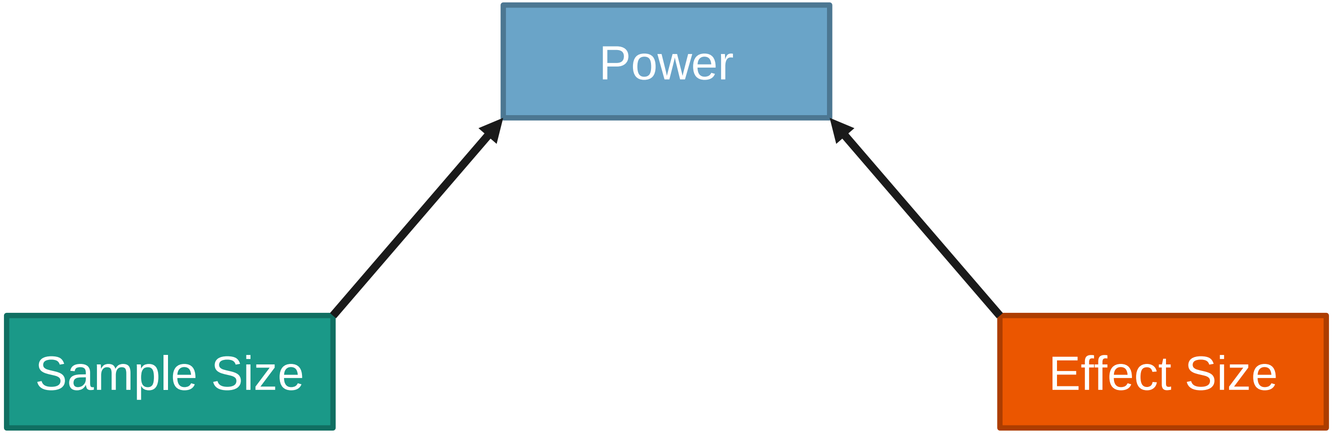
But it would be hard to miss this:





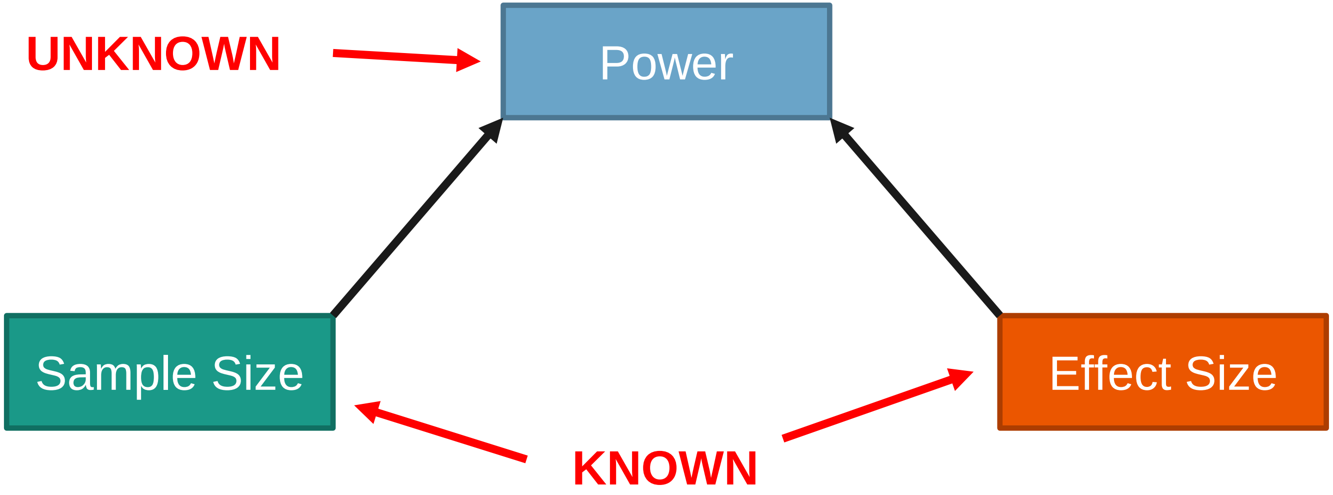
Power Analysis Questions

We can use any two of these pieces to find the third. This leads to three types of power analysis questions we can ask. We ask these questions *before collecting any data*.



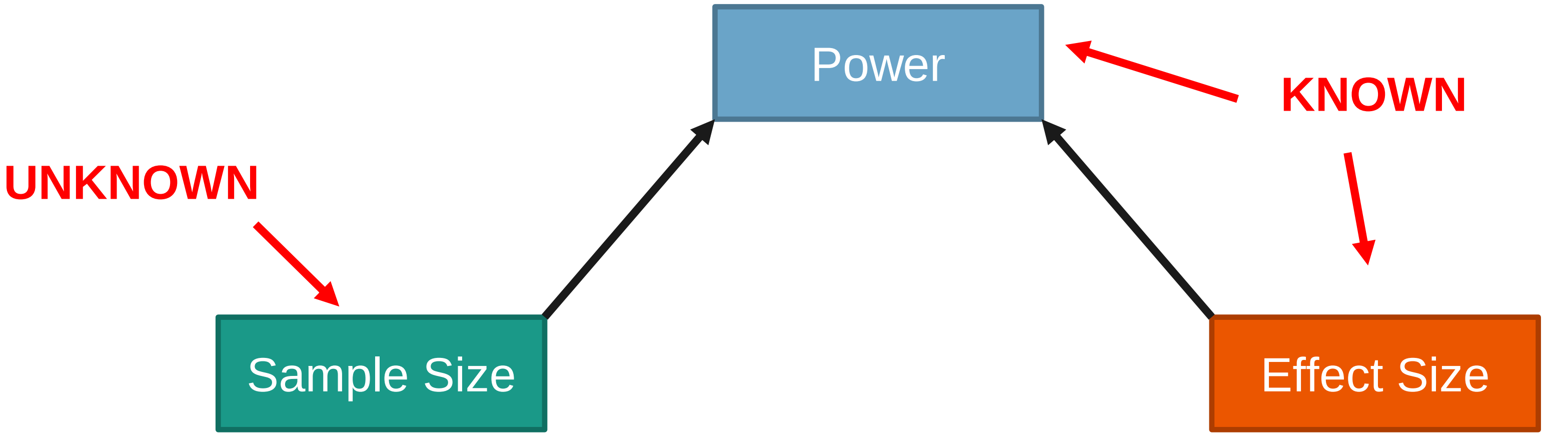
Power Analysis #1: computing power

"If we have n data points, and we would want to detect a difference, effect size = d, what is the power of the test?"



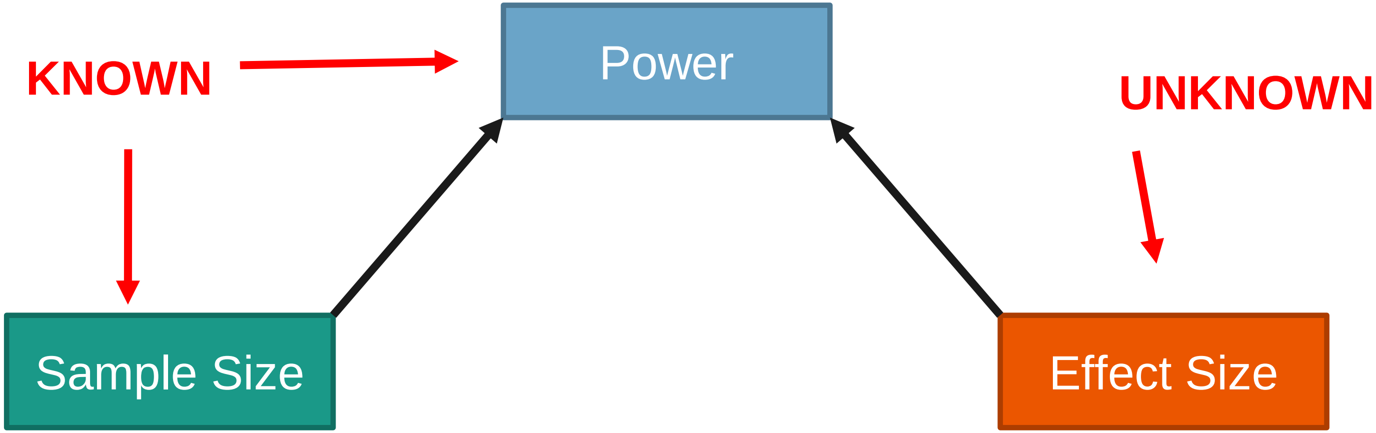
Power Analysis #2: computing sample size

“If we want to detect effects of size = d with power = p, how many samples do we need to collect?”



Power Analysis #3: computing sample size

“If we have n observations, what size effect could we detect with power = p?”



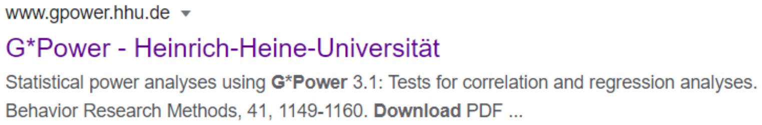
KNOWN

UNKNOWN

Power in G*Power

- G*Power is free software for power analysis.

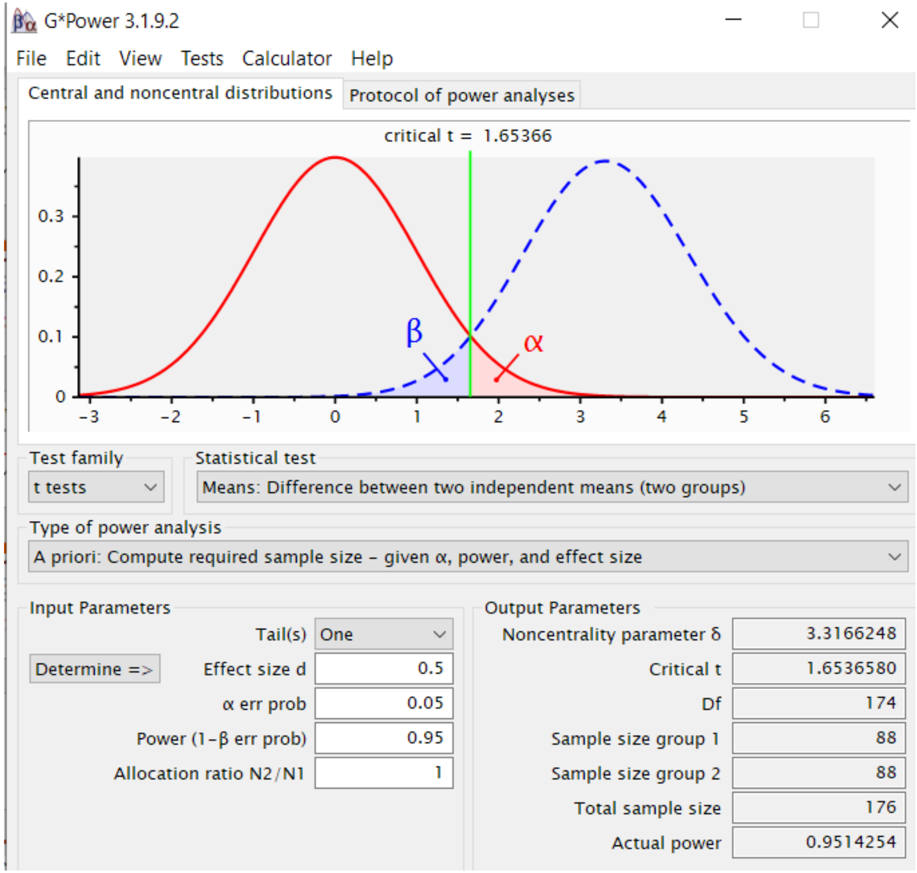
- If you google "G*Power download", it's this link:



www.gpower.hhu.de ▾

G*Power - Heinrich-Heine-Universität

Statistical power analyses using **G*Power** 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41, 1149-1160. **Download** PDF ...



Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s): One

Effect size d: 0.5

α err prob: 0.05

Power (1−β err prob): 0.95

Allocation ratio N2/N1: 1

Output Parameters

Noncentrality parameter δ: 3.3166248

Critical t: 1.6536580

Df: 174

Sample size group 1: 88

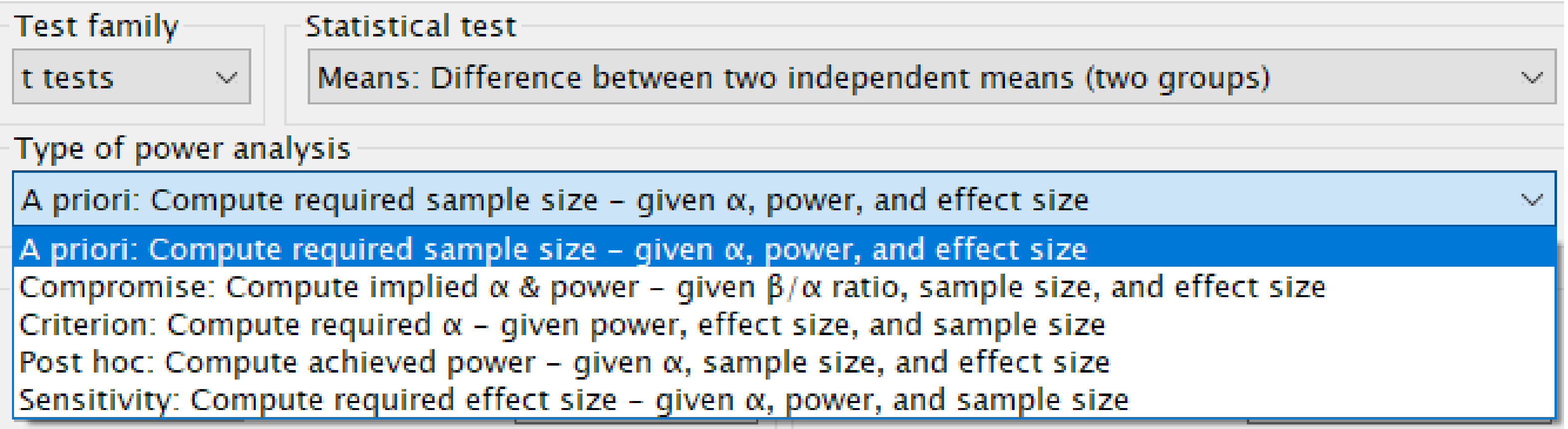
Sample size group 2: 88

Total sample size: 176

Actual power: 0.9514254

Options in G*Power

- Choose the type of test statistic (e.g. t, F, chi-square)
- Choose the type of test (e.g. correlation, difference in means…)
- Choose the type of analysis



Test family: t tests

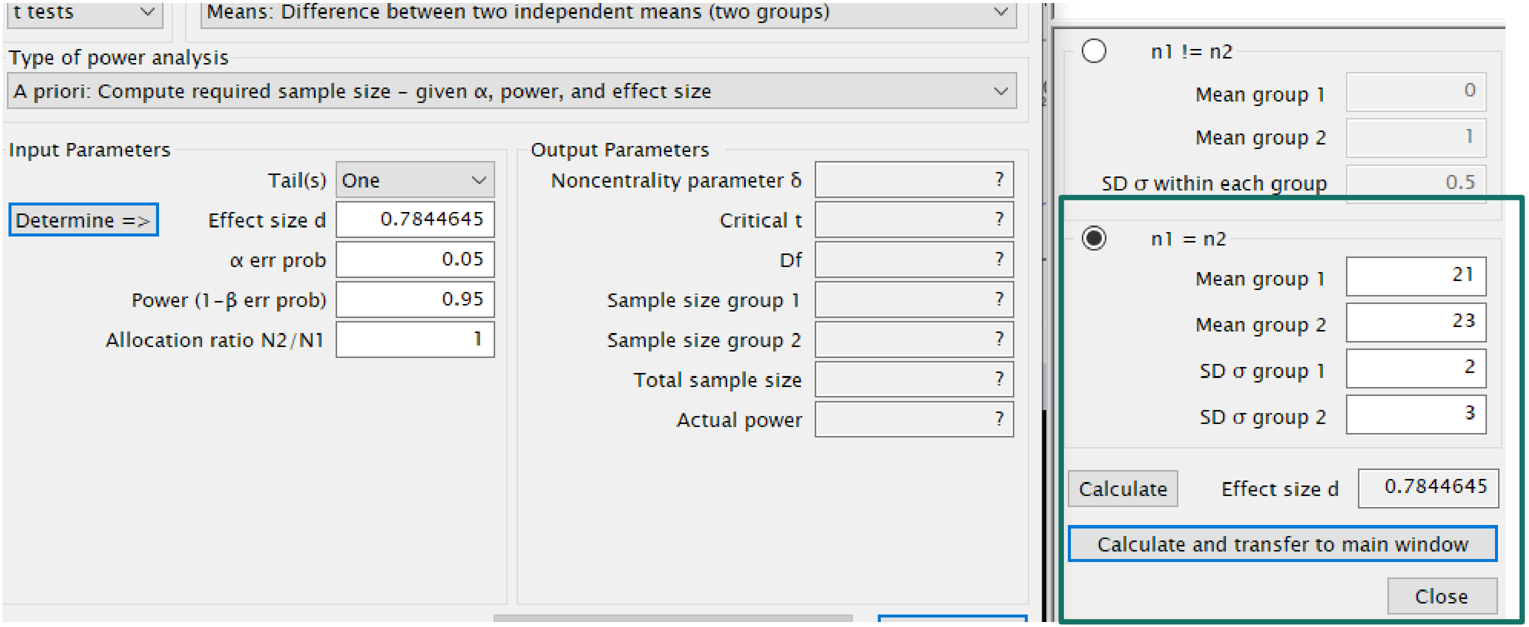
Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

A priori: Compute required sample size – given α, power, and effect size
Compromise: Compute implied α & power – given β/α ratio, sample size, and effect size
Criterion: Compute required α – given power, effect size, and sample size
Post hoc: Compute achieved power – given α, sample size, and effect size
Sensitivity: Compute required effect size – given α, power, and sample size

Effect size in G*Power

- You can give G*Power a standardized effect size (Cohen's d, R^2, partial eta-squared…), or you can give G*Power the components of the effect size (means, standard deviations, correlations…) and G*Power will determine the effect size:



t tests | Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s): One

Determine =>

Effect size d: 0.7844645

α err prob: 0.05

Power (1−β err prob): 0.95

Allocation ratio N2/N1: 1

Output Parameters

Noncentrality parameter δ: ?

Critical t: ?

Df: ?

Sample size group 1: ?

Sample size group 2: ?

Total sample size: ?

Actual power: ?

○ n1 != n2

Mean group 1: 0

Mean group 2: 1

SD σ within each group: 0.5

● n1 = n2

Mean group 1: 21

Mean group 2: 23

SD σ group 1: 2

SD σ group 2: 3

Calculate

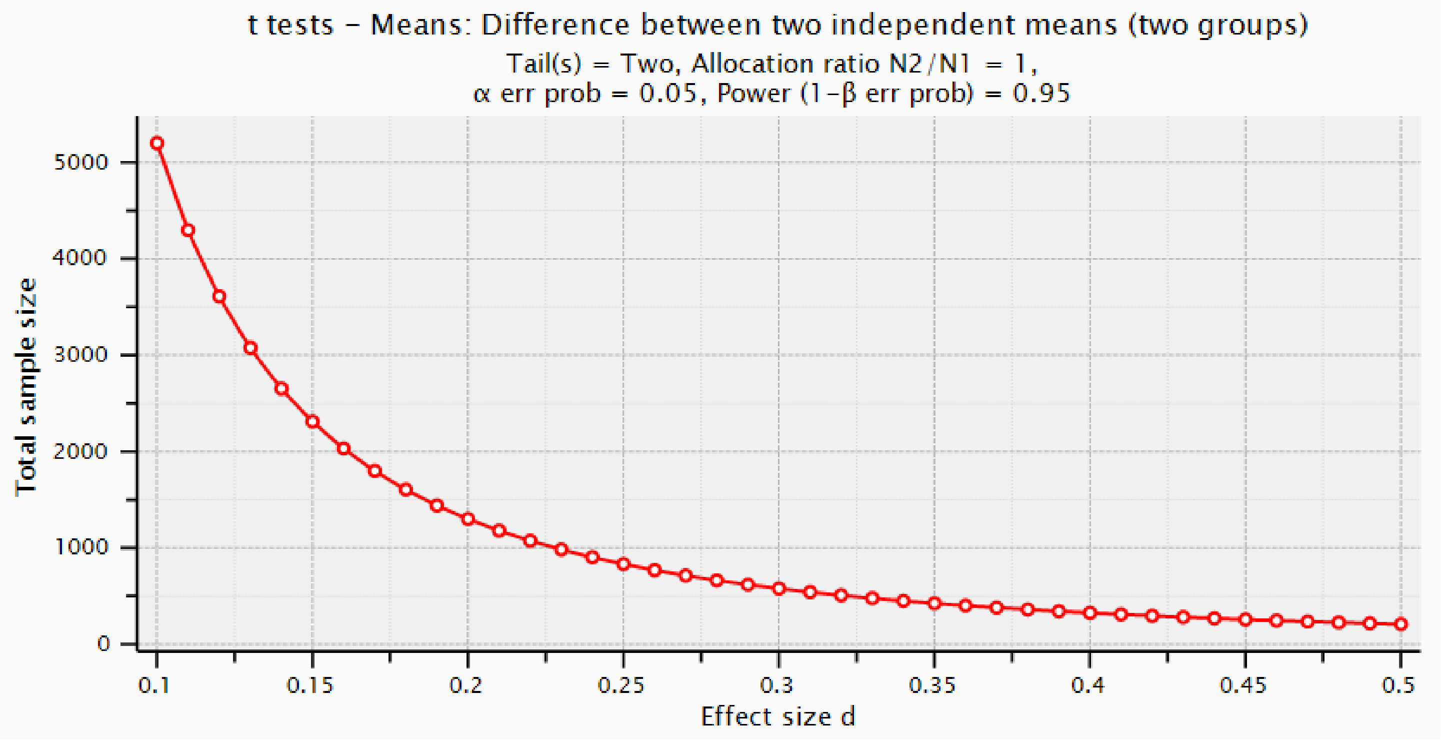
Effect size d: 0.7844645

Calculate and transfer to main window

Close

Even better: make a plot

- G*Power will also plot a graph of two elements of power, given the value of the 3rd:



Power analysis without using G*Power

- G*Power will do power analysis for “basic” analyses.
- If you want to do power analysis for something more complicated, I recommend using simulation.
- Basic idea: write code that generates fake data using your model and some assumed parameter values (e.g. means, standard deviations, slopes, correlations)
- Simulate a large number (e.g. 100,000) of fake data sets. For each data set, perform the hypothesis test. Estimated power is the proportion of these simulated data sets in which the null hypothesis is rejected.

Why We Care About Power: Efficiency

Determining sample size is often a battle between two forces:

Budget

Each additional research subject costs money.

publicdomainpictures.net

Results

Each additional research subject increases power and lowers the probability of Type II errors.

maxpixel.net - Creative commons

Efficiency and power

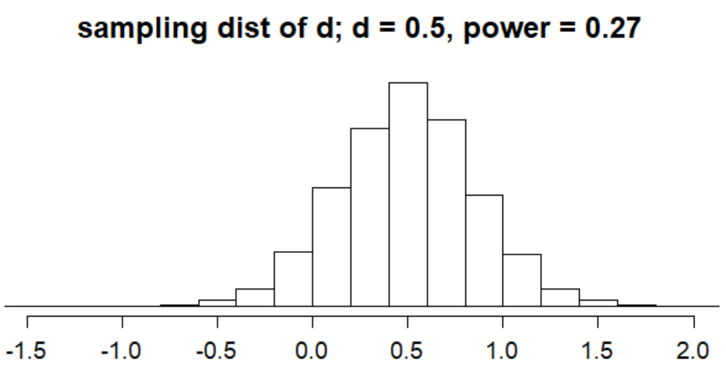
- If power is too low, we will be very likely to commit a Type II error.
- What’s the point of conducting a hypothesis test if we would have no chance of detecting a result, even if one existed?
- With low power, it is tough to know what to make of a “FTR H_0” result. Was it because the null hypothesis is really true, or because power is low?
- Conducting a test with low power is a waste of time and money, which is generally considered a **bad thing to do**.

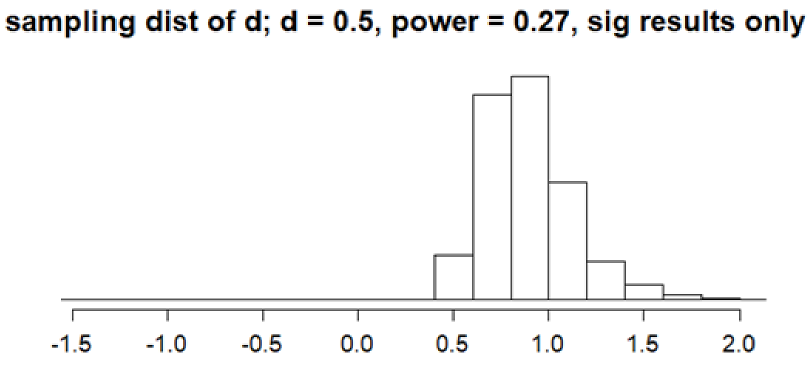
Why We Care About Power: Publication Bias

- Some journals will only publish a paper with “statistically significant” results. This is called **publication bias**.

- Publication bias, when combined with low power studies, can result in overestimates of effect sizes in published papers.

Same idea, using sampling distributions





- Left histogram: sampling distribution of Cohen’s d, when population d = 0.5 and power is low (0.27)
- Right histogram: same thing, but after removing all d statistics that do not achieve statistical significance.
- The mean of the distribution on the right is larger than the mean of the distribution on the left. So “significant” results are biased upward.

Why We Care About Power: Ethics

“The ethical statistician strives to **avoid** the use of **excessive or inadequate** numbers of research subjects by making informed recommendations for **study size**.”

Ethical Guidelines for Statistical Practice (2018),
American Statistical Association

(emphasis added)

Power and ethics

- If sample size is too small, **power is low**. Our research subjects will have gone through our study for nothing, since we are unlikely to get a good estimate of what we want to measure. Is this ethical if the study involves risk to the subjects? (Like a drug trial?)

- If sample size is too large, **power is very high**. We made a whole lot of research subjects participate when a smaller number would have been enough. Is this ethical if the study involves risk to our subjects?

Visualizing power

We can visualize power by looking at the sampling distribution of the p-value.

Recall that a “sampling distribution” is the distribution of values that a statistic takes on, under repeated sampling.

The p-value is a statistic, and so we can think about its sampling distribution.

Visualizing power (sampling distribution of p-value)

We reject the null hypothesis when $p < 0.05$.

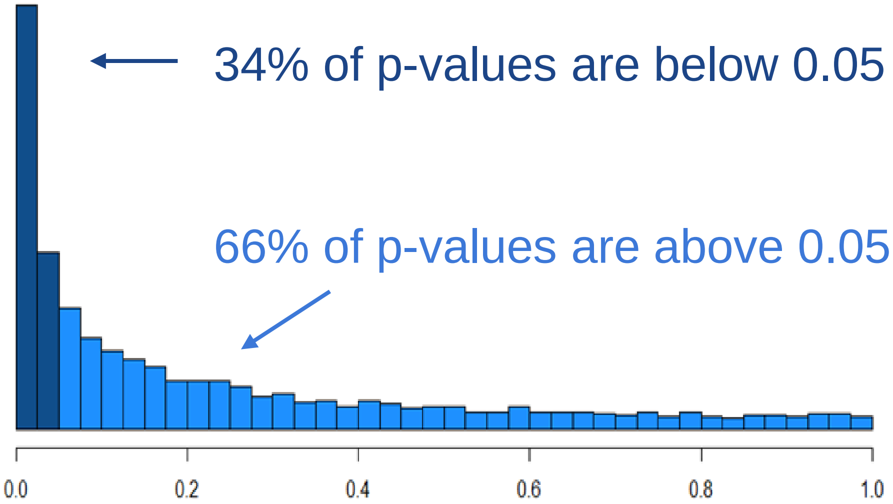
So the sampling distribution of the p-value will have more values below 0.05 when power is high, and more values above 0.05 when power is low.

These pictures are made using this app:

https://csu-statistics.shinyapps.io/visualize_power/

Visualizing power (sampling distribution of p-value)

Here is the sampling distribution of the p-value when power = 0.34



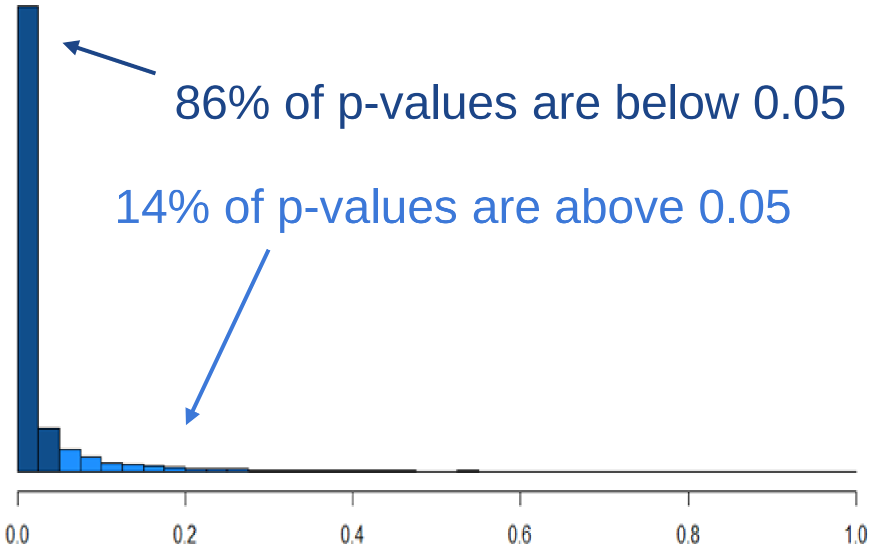
34% of p-values are below 0.05

66% of p-values are above 0.05

- The dark shaded bars are for $p < 0.05$ (reject H_0)
- The light shaded bars are for $p > 0.05$ (FTR H_0)
- Though the smallest p-values occur more frequently than the largest ones, p-values < 0.05 occur only 34% of the time. So this would be considered “low power”.

Visualizing power (sampling distribution of p-value)

Here is the sampling distribution of the p-value when power = 0.86



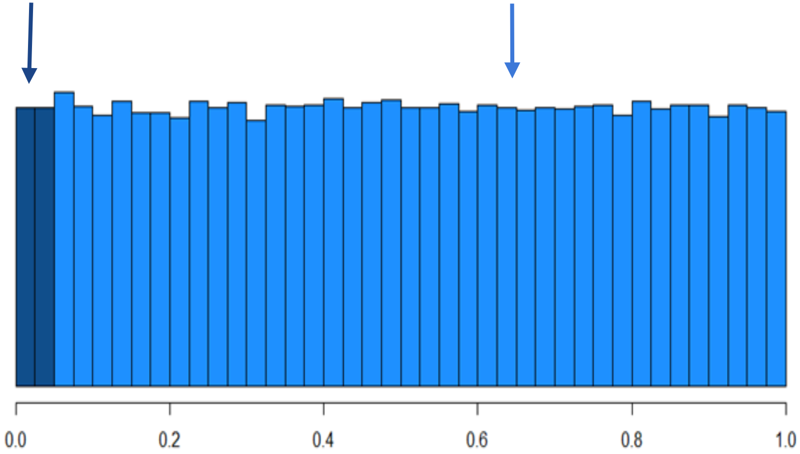
- Now 86% of p-values are below 0.05.
- Note that the distribution of the p-value is right skewed whether power is high or power is low.
- If H_0 is false, the p-value will always be right skewed, with small values occurring more frequently than large values. Higher power produces higher frequencies of small p-values.

Visualizing power (sampling distribution of p-value)

Here is the sampling distribution of the p-value when H_0 is true:

5% of p-values are below 0.05

95% of p-values are above 0.05



- In this distribution, 5% of p-values are below 0.05. This corresponds to a Type I error rate of $\alpha = 0.05$.

- Note that all p-values are equally likely when H_0 is true.

- P(p-value < 0.05) = P(p-value > 0.95) = P(0.05 < p-value < 0.10) = P(0.60 < p-value < 0.65), etc.