

Power and Sample Size Calculation

Presented by

Jim Matthews

Sydney Informatics Hub

Core Research Facilities

The University of Sydney



During the workshop



Ask **short questions** or clarifications during the workshop (either by Zoom chat or verbally). There will be breaks during the workshop for longer questions.



Slides with this **blackboard icon** are mainly for your reference, and the material will not be discussed during the workshop.

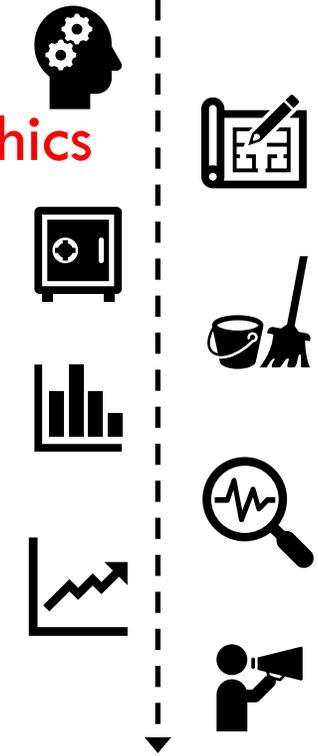


Challenge questions will be encountered throughout the workshop.



General Research Workflow

1. Hypothesis Generation (Research/Desktop Review)
2. **Experimental and Analytical Design** (sampling, power, ethics approval)
3. Collect/Store Data
4. Data cleaning
5. Exploratory Data Analysis (EDA)
6. Data Analysis aka inferential analysis
7. Predictive modelling
8. Publication



Outline

- Statistical power and sample size calculation – concepts
- Statistical Workflow
- Software tools and workflows – Statulator, Power and Sample Size (PS), others
- Four worked examples including difference between means and difference between proportions
- Power calculation for other designs
- References

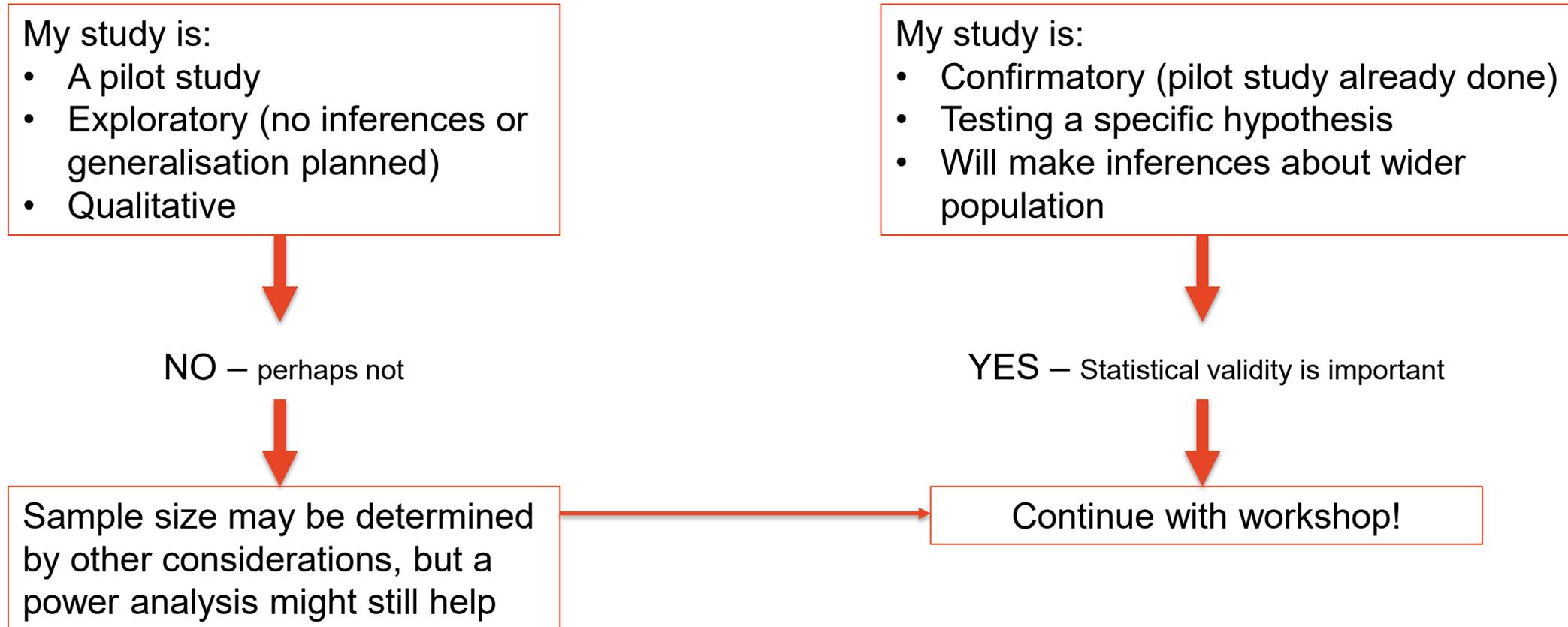
Why do we need to calculate power and sample size?

Why do we want to estimate the power of a study?

- To know that it is worth doing the study
- ...and make sure we are not wasting our time
- To plan the time and resources necessary
- To get a grant application approved
- To make sure the study design is ethically acceptable (Animal Ethics, Human Ethics)

But do I really need to calculate power?

What type of study are you planning?



What is the power of a study design?



[This Photo](#) by [Un](#) [ND](#)



Statistical power: The power to know...

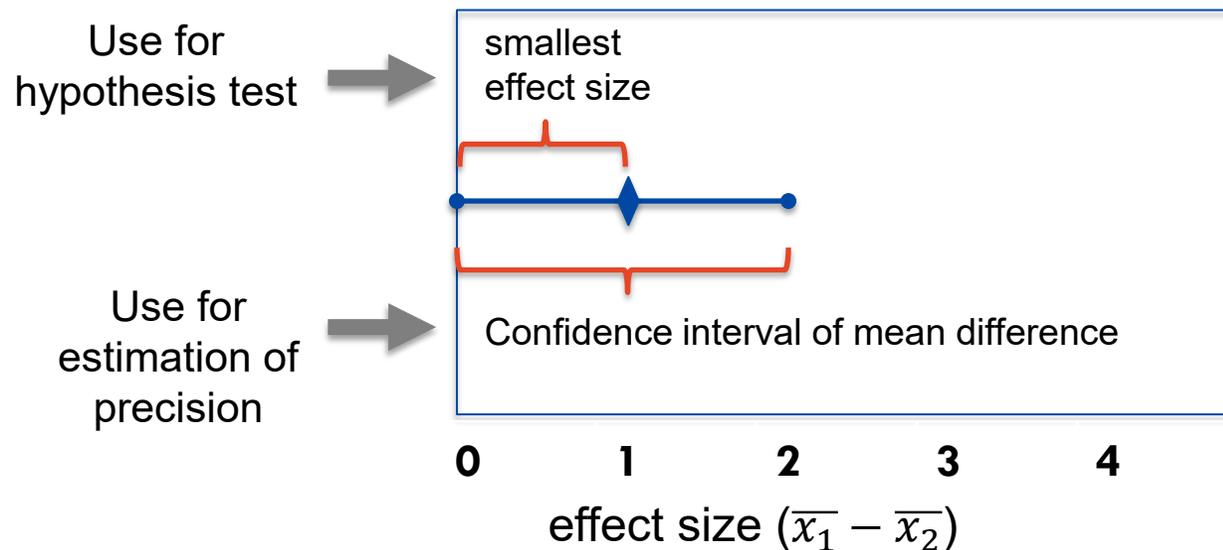


What do we want to know?

- Statistics involves using a sample to estimate (or make inference about) a population parameter (a mean, a mean difference, an odds ratio, or some other property).
- Usually we want to test a hypothesis: e.g. mean effect of drug X is different to the mean effect of drug Y
- Sometimes, we only want to estimate the population property *with some level of precision* and not perform a hypothesis test (e.g. the prevalence of disease A in the population is 10.5% [9.5%-11.5%])

What do we want to know?

- Sample sizes can be calculated for either research goal
- In this workshop, examples A-C are based on hypothesis tests of two groups, while example D is based on estimation of a proportion to a specified level of precision



Further reading on the use of CI for sample size calc: see chapter 3 of **“Determining Sample Size Balancing Power, Precision, and Practicality”** by **Dattalo**

We propose an alternative hypothesis but test a null hypothesis

Start with the hypothesis that you have generated, for example:

“Novel drug X lowers blood pressure more than standard-of-care drug Y”

In statistics, this is referred to as the alternative hypothesis (H_1). This is the hypothesis we are interested in. Classically, we test the veracity of the null (H_0) hypothesis:

“Drug X and drug Y lower blood pressure by the same amount”

A statistical test of the null hypothesis is always subject to **uncertainty**.

Errors can't be avoided, but can be controlled

Despite this **uncertainty**, when we perform a null hypothesis test, we are setting up a binary choice that can result in making a correct decision, or an error

We can't completely avoid errors, but we can choose the rate of error that is acceptable to us given **uncertainty**



[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

Types of Statistical Error

Table of error types		Reality	
		Null hypothesis (H_0) is	
		True	False
<u>Decision</u> about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)



Statistical power!



Alpha and beta are our chosen error rates

Type I error “False positive”

- Rate of false positives designated by the significance level, α
- We want the rate of false positives to be as low as possible
- The *convention* is to set the significance level to $\alpha = 0.05$, at this rate, we accept that even when the null hypothesis is true, it will be rejected one in every twenty runs of a study

Type II error “False negative”

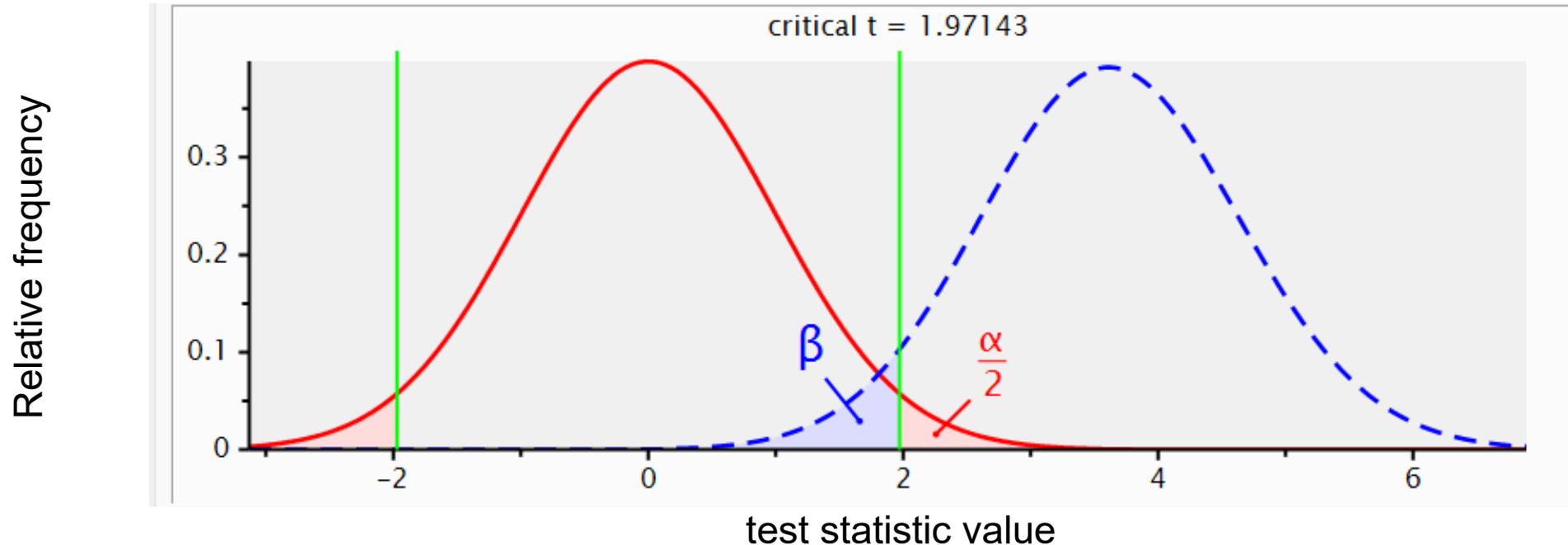
- Rate of false negatives designated by β
- Statistical power is the complement of β , denoted by $1 - \beta$
- We want power to be as high as possible
- The *convention* is $1 - \beta \geq 0.8$, at 0.8, even when the null hypothesis is false, it will not be rejected one in every five runs of an experiment



P-values allow us to make decisions and quantify evidence

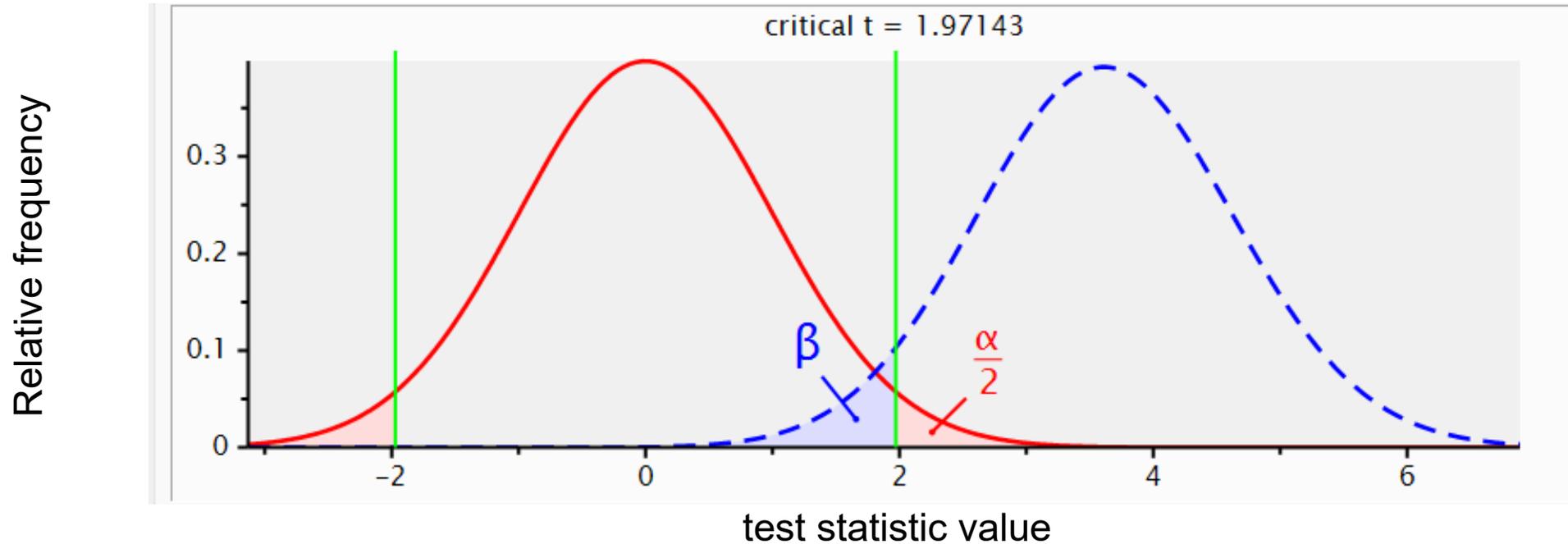
- Maligned, misunderstood and occasionally spurned, p-values have a simple purpose: *“The probability of observing a test statistic at least as extreme as the one observed, assuming the null hypothesis is true”*
- Remember that p-values are produced from a “null-is-true” perspective, so we always assume that the null hypothesis is true when calculating them
- If a p-value falls below the significance threshold α (and especially if it falls far below a given threshold), we may suspect that we are in fact in the “null-is-false universe” and reject the null hypothesis
- Counterintuitive: We don’t test the alternative hypothesis directly, so the smaller the p-value, the more plausible the alternate hypothesis becomes
- A discussion of how p-value thresholds relate to α and β is given on the following slides for your reference

Alpha is also the p-value threshold



- Here is a plot from G*Power software
- The red curve shows how a given test statistic (a t-statistic in this case) is distributed in the “null-is-true universe”, where the null hypothesis is always true
- The blue and dotted curve shows how this test statistic is distributed in a “null-is-false universe” for a particular effect size – more on effect sizes later
- The chosen α determines the critical value of the test-statistic (green vertical line)

Alpha is also the p-value threshold

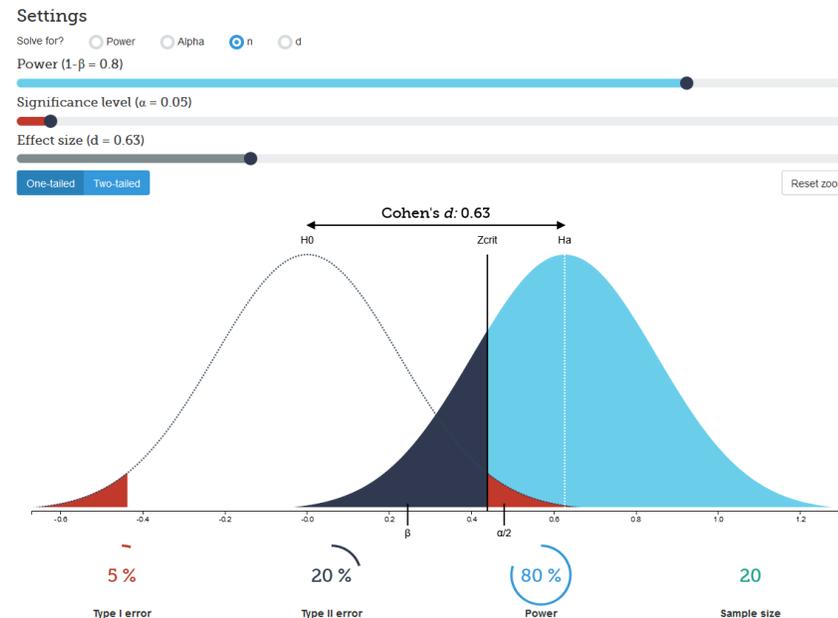


- For the usual two-tailed hypothesis, the α area is distributed across the two tails of the red solid line distribution (called the null distribution). The total area is equal to α
- A p-value is calculated based on the area of the null distribution that covers test statistic values more extreme than that observed (on both sides for a two-sided test)
- There are no overlaps between the α and β areas, reflecting the binary choice made of whether to reject the null hypothesis or not
- $1 - \beta$ is the area under the blue dotted distribution that is not shaded

Building your intuition of a deeply unintuitive procedure



- For a more interactive view of these distributions to build your intuition check out: <https://rpsychologist.com/d3/nhst/>
- Note that the author is “deeply skeptical about the current use of significance tests”, but null-hypothesis statistical testing (NHST) is a mainstay of modern research. So, we must know how to use it even if we don’t like it.



Not every hypothesis is a superiority test

The most recognised and widely-used hypothesis test is whether two measures are *exactly* equal or different by any amount (a superiority test).

H_0 : Difference = 0

H_1 : Difference \neq 0

Other study objectives will lead to other types of hypothesis test.

The types below are frequently found in clinical trials (e.g. a novel drug performs *no worse* than the existing drug: Non-inferiority test)

Similar tests:

- Equivalence test
- Minimum Effect test

See reference for further details: Julious, Steven A. Sample Sizes for Clinical Trials . Boca Raton: CRC Press/Taylor & Francis, 2010. Print.

We need inputs to calculate statistical power

- It will depend on:
 - Sample size
 - Chosen significance level (α)
 - Minimum effect size to detect
 - Variance within groups
 - Experimental design and type of statistical hypothesis test

- Usually, we want to calculate a sample size given a required minimum power.



Decisions regarding the study design can be critically important in determining statistical power. This is covered in the “**Experimental Design**” workshop.

Power calculation can become sample size calculation

- It will depend on:
 - ~~Sample size~~
 - Chosen significance level (α)
 - Minimum effect size to detect
 - Variance within groups
 - Experimental design and type of statistical hypothesis test
 - Statistical Power ($1-\beta$)
- Usually, we want to calculate a sample size given a required minimum power.



Decisions regarding the study design can be critically important in determining statistical power. This is covered in the “**Experimental Design**” workshop.

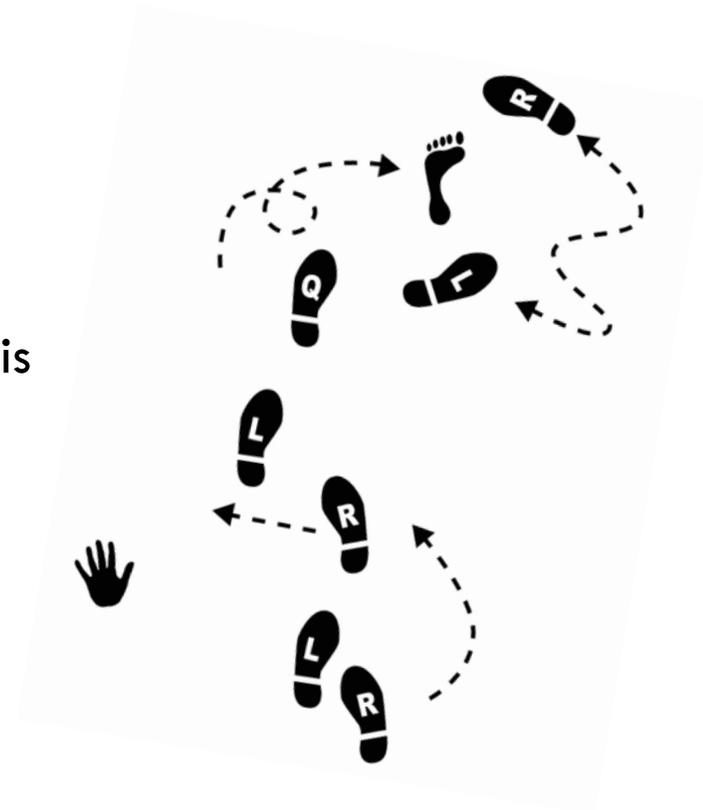
What Have We Learned So Far?

- Why power matters: Avoid wasted effort, ensure ethical design, secure funding
- Key concepts: Power = $1 - \beta$; Type I error (α), Type II error (β)
- Inputs for power/sample size: α , desired power, SESOI, variance, design
- Effect size & precision: SESOI aligns statistical and scientific significance
- Confidence intervals matter for estimation goals



What is a workflow?

- Every statistical analysis is different, but all follow similar paths. It can be useful to know what these paths are
- We have developed practical, step-by-step instructions that we call ‘workflows’, that you can follow and apply to your research
- We have a general research workflow that you can follow from hypothesis generation to publication
- And statistical workflows that focus on each major step along the way including **study design and power**



Statistical Workflows

- Our **statistical workflow** will be outlined within this workshop
- **Statistical workflows** are software agnostic, in that they can be applied using any statistical software
- There are accompanying **software workflows** that show you how to perform the sample size statistical workflow using software packages:
 - **Online calculators: Statulator and PS**
 - **G*Power**
 - **SPSS**
- Download these from the [Statistical Consulting website](#)



Sample size calculation statistical workflow

Sample size calculation workflow steps

1. Determine experiment type and statistical test
2. Set α and $1 - \beta$
3. Set the smallest effect size of interest
4. Estimate the variance
5. Calculate the minimum sample size
6. Explore scenarios

1. Determine experiment type and statistical test

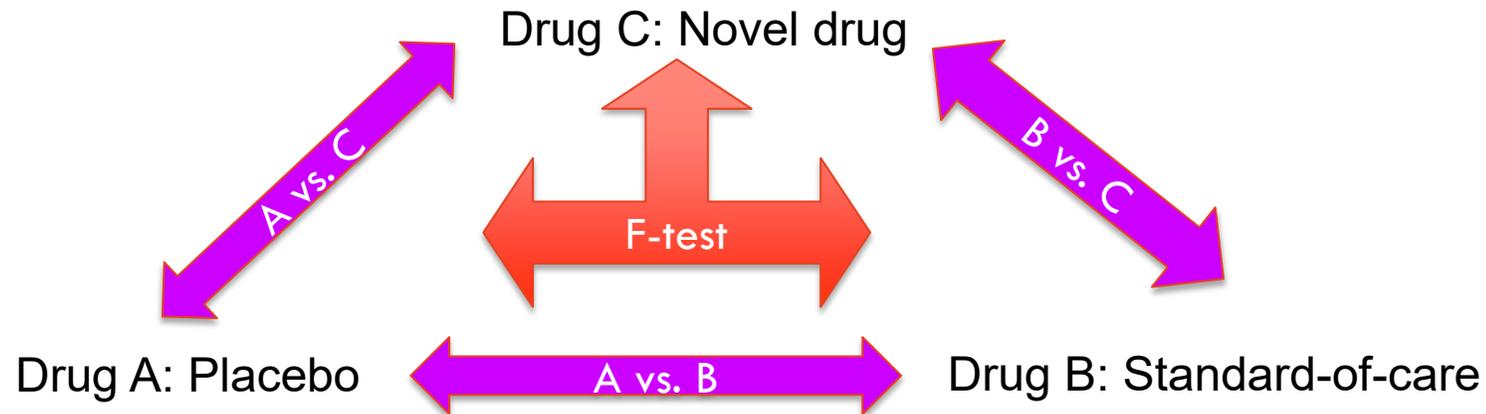
For example:

Experimental design	Main assumptions	Proposed statistical test
Comparison of 2 means	independent groups, normally distributed outcome	t-test
Comparison of 2 means	independent groups, no assumption of normality	Mann-Whitney U test
Comparison of 2 proportions	independent groups	z-test/Fishers exact
Comparison of means, more than 2 groups	independent groups normally distributed outcome	ANOVA, F-test

1. Determine experiment type and statistical test

- Tip from the consulting room:

- Your study design may lead to a series of hypothesis tests of interest
- You need to choose which of these hypothesis tests should be used to “power your experiment”
- A typical example is an ANOVA design experiment
- With three groups this would result in four hypothesis tests: an **F-test**, and three **post-hoc t-tests**



Which of these tests should be used to power the experiment (i.e. ensure sufficient sample size)

2. Set α and $1 - \beta$

Setting values of parameters

- Typically choose $\alpha = 0.05$ (or lower)
- Typically choose power $(1 - \beta) = 0.8$ (or higher)

You should have a justification for choosing a particular α and $1 - \beta$. There is no reason why the conventional values must be used... consider the two examples on the next slide.

2. Set α and $1 - \beta$



Let's consider a couple of scenarios:

- You are investigating the rate of detection of a pathogen in hospitals that can be fatal in immunocompromised patients. It is thought that the pathogen is very rare, but you suspect it is more common. Your study is very expensive to set up, but it is relatively inexpensive to collect more samples. If there is a negative finding, it is unlikely the study will be repeated. Should you use the conventional power ($1 - \beta$) of 80%?
 - *Consider increasing $1 - \beta$ (decreasing β) as 80% means a one in five chance of a false negative even if the alternative hypothesis is true*
- You are investigating the use of novel disinfectants on hospital pathogens by screening a panel of potential disinfectants. If a disinfectant is found to be effective, the disinfectant will be tested again in a new and larger sample. Should you use the conventional α of 0.05?
 - *Consider increasing α . A false positive is less of a problem, because the finding will be replicated*



2. Set α and $1 - \beta$

A good example of how a personal preference of an influential figure can become an iron-clad convention.

*“It is proposed here as a convention that, when the investigator has no other basis for setting the desired power value, the value .80 be used. This means that β is set at .20. This arbitrary but reasonable value is offered for several reasons (Cohen, 1965, pp. 98-99). **The chief among them takes into consideration the implicit convention for α of .05.** The β of .20 is chosen with the idea that the general relative seriousness of these two kinds of errors is of the order of .20/.05, i.e., that Type I errors are of the order of four times as serious as Type II errors. **This .80 desired power convention is offered with the hope that it will be ignored whenever an investigator can find a basis in his substantive concerns in his specific research investigation to choose a value ad hoc.**”*

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.



Multiple testing

- In high-throughput biology, which involves many performing many statistical tests for each outcome being measured, type I error is often controlled using a False Discovery Rate (FDR) rather than the usual Family-Wise Error Rate (FWER). This is effectively setting the α to a much higher level. Any differences detected on the high throughput platform will be verified using a lower throughput technology.
- A common and familiar form of multiple testing is adjusting for multiple post-hoc tests following an ANOVA F [omnibus] test
- Bonferroni is the most stringent multiple testing correction, but it easy to apply
- Other forms of multiple testing correction can be applied in other contexts. Come to a consult for further guidance.

3. Set the smallest effect size of interest

Effect size examples: difference in means, difference in proportions, odds ratio...

What is the smallest effect size of interest?

- Decide on a smallest effect size of interest (SESOI) as a target to detect in your hypothesis test. This should be based on the smallest effect size that is of *scientific interest/significance*
- You will need to use your domain expertise to decide this, e.g.:
 - What constitutes a *clinically* meaningful difference between patient groups in a score measuring depression?... What is a *biologically* meaningful difference in gene expression in two different cell types?... What difference in number of seconds to complete a task would you consider important in a mouse *behaviour* model?
- If it is unknown, you may need to consider a range of plausible effect sizes and evaluate the required sample size as each of these (as described in the examples)

3. Set the smallest effect size of interest

too small

Effect size chosen is smaller than necessary



- The sample size is larger than necessary
- Possible waste of resources
- Can achieve statistical significance with an effect that is too small to be interesting or useful

just right

Effect size chosen is based on sesoi



- The sample size is just right
- If statistical significance is achieved, then it will align with scientific significance
- Most efficient use of resources

too large

Effect size chosen is larger than necessary

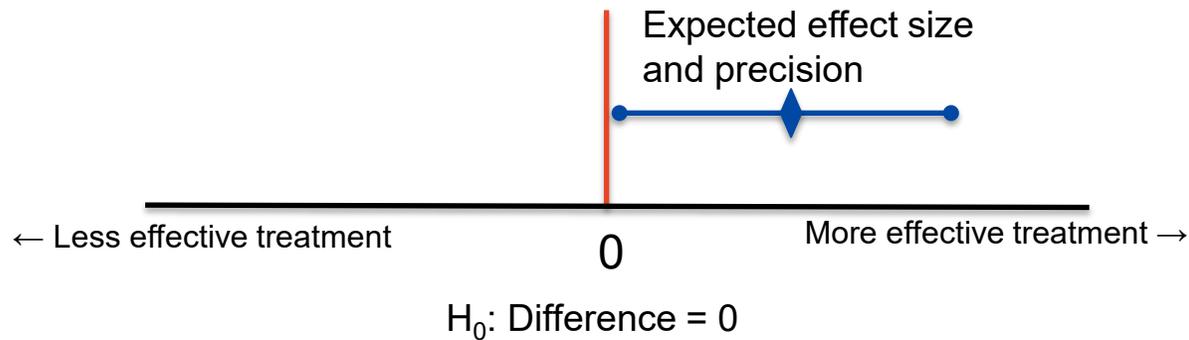


- The sample size is too small
- Not able to achieve statistical significance for small effect sizes of interest
- May detect large effects only
- Possible waste of resources

goldilocks

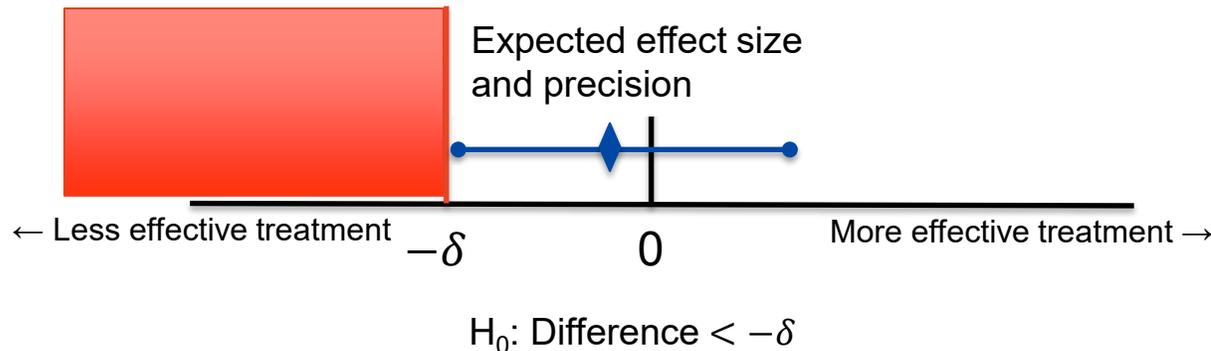


3. The SESOI may not be the only effect size you need



Effect size – note that your SESOI may not be the same as your expected effect size. Depending on the hypothesis test you may need to think of these as different quantities:

For the typical test (superiority), you use only your SESOI to calculate the sample size (statistical significance and scientific significance align)



For a non-inferiority test, your SESOI determines delta (δ): the boundary of the range of effect sizes that have no clinical/practical/scientific significance.

You also need an expected effect size to calculate the sample size.

https://lakens.github.io/statistical_inferences/09-equivalencetest.html

4. Estimate the variance

Within study variance is often the big unknown in a sample size calculation

How to estimate it?

- Estimate standard deviation (or proportions) from previous experiments?
- Consider theoretical bounds (eg proportions, for 5pt scales)
- Seek expert knowledge?
- If no idea, may be best to do pilot study

4. Estimate the variance **Standardised Effect Size**

Alternative: Use the Standardised Effect Size

Many effect sizes can be “standardised” by considering the ratio of the effect size to a within group standard deviation.

For example: Cohen’s d is the ratio of the difference in means to the pooled standard deviation

$$d = \frac{\overline{x_1} - \overline{x_2}}{s}$$

Cohen’s d is therefore analogous to the number of standard deviations difference, or the z-score difference. Also called the standardised mean difference (SMD).

4. Estimate the variance **Standardised Effect Size**

Alternative: Use the Standardised Effect Size

Instead of deciding on effect size and an estimate of SD, we can choose a value of Cohen's d based on accepted interpretations of relative size.

<i>Effect size</i>	<i>d</i>	Reference
Very small	0.01	Sawilowsky, 2009
Small	0.20	Cohen, 1988
Medium	0.50	Cohen, 1988
Large	0.80	Cohen, 1988
Very large	1.20	Sawilowsky, 2009
Huge	2.0	Sawilowsky, 2009

Other guidelines are published for other standardised effect sizes.

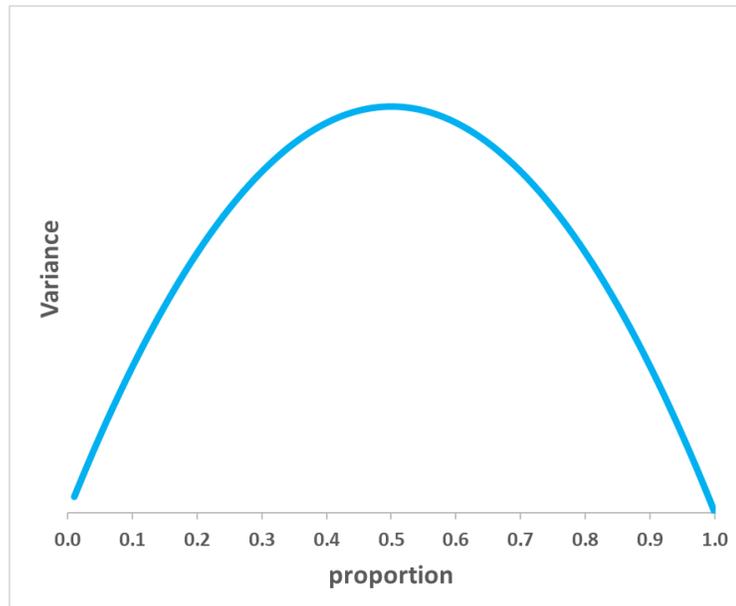
Note however that interpretation can vary across different fields of study.

4. Estimate the variance **Theoretical upper bound**

For proportions the maximum variance occurs when $p=50\%$ and is at a minimum when $p=0\%$ and 100% .

So we can use $p=50\%$ to find a theoretical upper bound.

$$\text{Variance}(p) = p(1 - p)$$



Why is this the case?

It might be strange to think about the variance of a proportion. You can think about it as a feature of a binomial (0 or 1) outcome.

What proportion of subjects are 0 and how many are 1 for some outcome? (e.g. 0 = non-smoker and 1 = smoker)

Proportions must be between 0 and 1. When our estimate is close to these limits, there is less variability in our estimates for the same number of subjects surveyed.



4. Estimate the variance **Theoretical upper bound**

For ordinal responses such as 5pt scales a similar limit applies:

Possible responses are: 1, 2, 3, 4 or 5

Mean=3 Min=1 Max=5

$$\text{Variance}(5\text{pt scale}) = (\max - \text{mean})(\text{mean} - \min)$$

$$\text{Max Variance}(5\text{pt scale}) = (5 - 3)(3 - 1) = 4$$

In practice the actual variance will be smaller than the max. A rule of thumb is explained on StackExchange

<https://stats.stackexchange.com/questions/23519/how-do-i-evaluate-standard-deviation>



5. Calculate the minimum sample size

- This is typically done using a software package
- Formulae for the calculation vary with the type of experimental design and the statistical test. We won't look at these too closely, but let's note some common features to reinforce the theory we have learned.
- The formula below is for a difference in means of two groups, where the standard deviation is known (good for illustration, though not often used in practice).

$$n = \frac{2 (Z_{\alpha/2} + Z_{1-\beta})^2}{\left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2}$$

The Z are critical values based on the chosen α and β . The smaller α or β are, the larger their critical values.

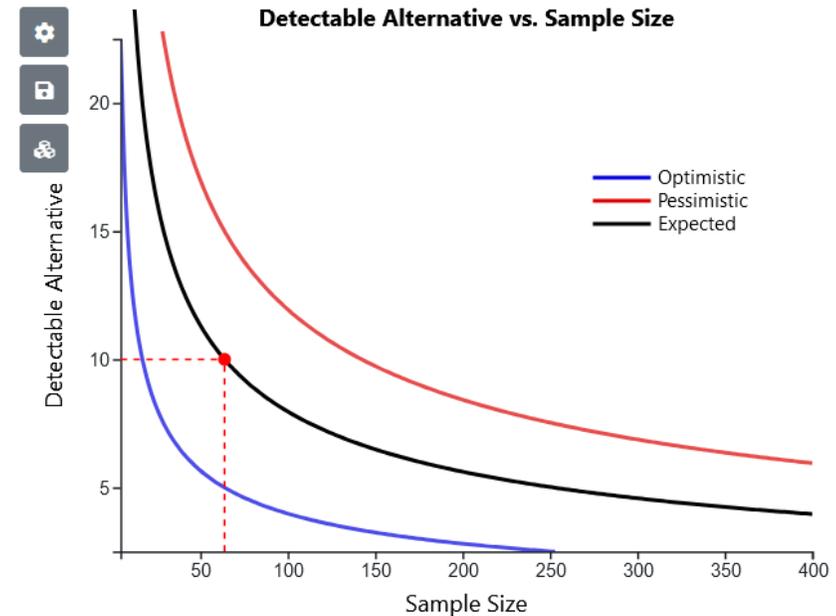
Note that for a fixed sample size if we increase alpha (accept higher FP) rate, we would increase power $1 - \beta$

Note that the expected difference in means $\mu_1 - \mu_2$ is divided by the expected SD σ . The denominator resembles Cohen's D, the standardised effect size. We use the greek letters here instead to indicate these are 'known' values. The whole expression is squared. The smaller the expected difference, the larger the sample size required to detect it.

6. Explore scenarios

Tip from the consulting room:

- Don't just calculate a single sample size n for a **power calculation**
- Use the software to calculate n for a range of scenarios to explore uncertainty in the input values used in the calculation. You may have a large uncertainty in the variance, and/or the expected effect size.
- A **Power Analysis** incorporates informative plots to design and plan your study



Useful for experiment planning: *Consider also the shape of the cost curve for sample data collection*

In the above example increasing sample size up to ~ 100 yields big effect size detection benefit but increasing sample size beyond ~ 100 yields diminishing returns. You can use this information in combination with feasibility considerations to choose a target sample size.

Examples using Statulator and PS

This section shows four simple worked examples:

A: Difference between two means (continuous response)

B: Difference between two means (survey response)

C: Difference between two proportions

D: Estimation of a single proportion

We will step through example A and leave B, C and D for you to do later.

Power calculation software used in this workshop

Statulator

- Free on-line statistical calculator
- Developed by epidemiologists and biostatisticians at Sydney University
- Easy to use
- Live interpretation provided for each calculation
- Visualisations to help you explore scenarios
- Incorporates other types of hypothesis test beyond the usual superiority type

Power and Sample Size (PS)

- Free on-line power and sample size calculator
- Useful for power analysis plots
- Live interpretation provided for each calculation

The screenshot shows the Statulator website interface. At the top, there is a navigation bar with the Statulator logo and menu items: Sample Size, Descriptive Analysis, Statistical Tests, About, Blog, and Contact us. The main heading is "Sample Size Calculator for Comparing Two Independent Means". Below this, there are three bullet points: "Provides live interpretations.", "Evaluates the influence of changing input values.", and "Adjusts sample sizes for continuity and clustering." There are four tabs for hypothesis testing: Equality (selected), Non-inferiority, Superiority, and Equivalence. Below the tabs are three buttons: Calculate, Visualise, and Tabulate. The "Input Values" section asks the user to select one of two options: "Expected Means" (selected) or "Expected Difference between Means". A note at the bottom explains the Options button. At the very bottom, there are four buttons: Calculate, Options, Adjust, and Reset.

PS Power and Sample Size



Statulator



A. Difference between two means

Example: Chicken Welfare – Bone density

The bone density of chickens is an important indication of their welfare. We want to test to see if (mineral) bone density can be improved from 120 to at least 130 mg/cm³



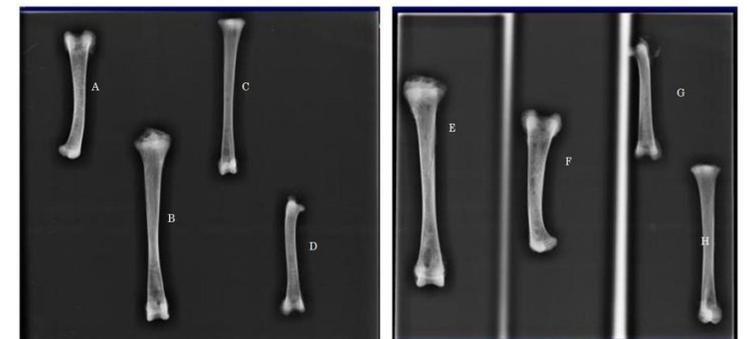
Control Group (1) = normal diet

Treatment Group (2) = high mineral diet

Response variable: Measure the tibia bone density after 6 weeks growth.

How many chickens do I need to detect a difference in bone density of 10 mg/cm³?

What type of statistical test will we perform?



TY - JOUR AU - Mabelebele, Monnye AU - Norris, Dannah AU - Siwendu, Ndyebo AU - Ng'ambi, Jones AU - John, Alabi AU - Mbajjorgu, C.A. PY - 2017/01/01 SP - 1387 EP - 1398 T1 - Bone morphometric parameters of the tibia and femur of indigenous and broiler chickens reared intensively VL - 15 DO - 10.15666/aeer/1504_13871398 JO - Applied Ecology and Environmental Research ER -

A. Difference between two means

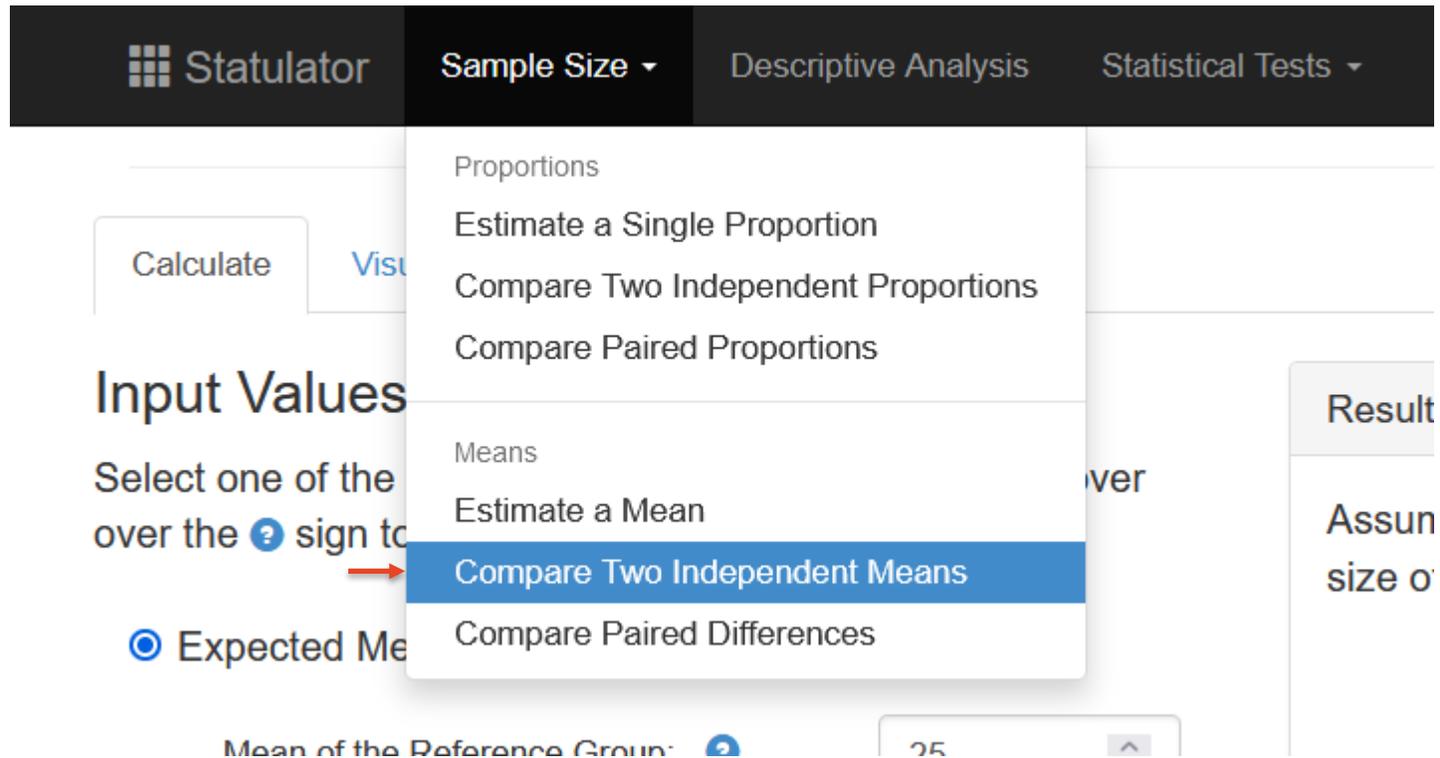
Example: Chicken Welfare – Bone density

- Step 1: We will use a t-test (assume normality)
- Step 2: $\alpha=0.05$ and $1 - \beta=0.8$ (leave values at their conventional levels)
- Step 3: Smallest Effect Size of interest is 10 mg/cm^3
- Step 4: Estimate the variance
 - We know from previous studies what the typical variation in bone density is for the control diet. We don't know about the treatment diet. We will use an estimate from the control diet of $SD=20 \text{ mg/cm}^3$
- Assume we will have equal size groups, $n_1 = n_2$

A. Difference between two means

Step 5: Statulator

Select Sample Size -> Compare Two Independent Means



A. Difference between two means

Step 5: Calculate the minimum sample size

- Put all the information into Statulator
 - Choose expected difference between the means
 - Difference between Two Means = 10 mg/cm^3
 - Expected Standard Deviation = 20 mg/cm^3
 - Leave all other Options at their defaults (80% power, 5% alpha)
 - Click “Calculate”

Statulator

Calculate

Visualise

Tabulate

Input Values

Select one of the two options to specify input values. Hover over the ? sign to obtain help.

Expected Means ?

Expected Difference between Means ?

Difference between Two Means: ?

10

Expected Standard Deviation: ?

20

Click the Options button to change the default options for Power, Significance, Alternate Hypothesis and Group Sizes. Use the Adjust button to adjust sample sizes for t-distribution (option applied by default), and clustering.

▶ Calculate

Options

Adjust

↻ Reset

A. Difference between two means

Step 5: Calculate the minimum sample size

- Group sample sizes are $N_1=63$, $N_2=63$
- Statulator provides a plain-English explanation of the calculation. Relevant paragraphs can be included in your grant or ethics application
- Also a good opportunity to check that Statulator has done the calculation you think it has (in this case a two-sided test, using a t-distribution)

Statulator

Results and Live Interpretation Download

Assuming a pooled standard deviation of 20 units, the study would require a sample size of:

63

for each group (i.e. a total sample size of 126, assuming equal group sizes), to achieve a power of 80% and a level of significance of 5% (two sided), for detecting a true difference in means between the test and the reference group of 10 units.

In other words, if you select a random sample of 63 from each population, and determine that the difference in the two means is 10 units, and the pooled standard deviation is 20 units, you would have 80% power to declare that the two groups have significantly different means, i.e. a two sided p-value of less than 0.05.

Reference: Dhand, N. K., & Khatkar, M. S. (2014). Statulator: An online statistical calculator. Sample Size Calculator for Comparing Two Independent Means. Accessed 13 January 2025 at <http://statulator.com/SampleSize/ss2M.html>

Note: Statulator used the input values of a power of 80%, a two sided level of significance of 5% and equal group sizes for sample size calculation and adjusted the sample size for t-distribution. You may change the options by clicking [here](#) or the 'Options' button and the adjustments by clicking [here](#) or the 'Adjust' button.

A. Difference between two means

Example: Chicken Welfare – Bone density

Step 6: Explore scenarios

Power Analysis

- It is advisable to explore some different scenarios to incorporate uncertainty in our inputs
- Consider how much your within study standard deviation could vary from your point estimate
 - Our estimate is $SD = 20 \text{ mg/cm}^3$ (expected)
 - Possible min value = 15 mg/cm^3 (optimistic)
 - Possible max value = 30 mg/cm^3 (pessimistic, conservative)
- Consider how much the difference between the groups (effect size) could vary from our point estimate
 - Our estimate is 10 mg/cm^3
 - Possible min value = 5 mg/cm^3 (pessimistic, conservative)
 - Possible max value = 15 mg/cm^3 (optimistic)

A. Difference between two means

- Click the visualise tab
- Enter the ranges into Statulator and click 'Customize'

Statulator

The screenshot shows the 'Statulator' interface with the 'Visualise' tab selected. A red arrow labeled '1' points to the 'Visualise' tab. Below it, the 'Customize Visualisation' section is shown with a red arrow labeled '2' pointing to the 'Expected Pooled Standard Deviation (x-axis)' section. This section contains three input fields: 'From Min' (15), 'To Max' (30), and 'By' (1). A second red arrow labeled '3' points to the 'Difference of Means Between Groups' section, which contains three input fields: '1st Series' (5), '2nd Series' (10), and '3rd Series' (15). A third red arrow labeled '4' points to the 'Customize' button at the bottom of the form. The form also includes an 'Options' button, an 'Adjust' button, and a 'Reset Form' button.

Calculate Visualise Tabulate

Customize Visualisation
Customize the plot by changing input values from here.

Expected Pooled Standard Deviation (x-axis): ?

From Min To Max By

15 30 1

Difference of Means Between Groups: ?

1st Series 2nd Series 3rd Series

5 10 15

Note: You may change the default options for Power, Significance, Alternate Hypothesis and Group Sizes by clicking the 'Options' button. Click the 'Adjust' button below to adjust sample sizes for the t-distribution (option applied by default), and clustering.

Customize Options Adjust Reset Form

A. Difference between two means

- We get a plot showing how the specified scenarios affect the required sample size.
- Although useful, for some this is not a complete power analysis and you may need additional plots
- For plots of sample size vs. power and sample size vs. detectable effect we can use [Power and Sample Size \(PS\)](#)

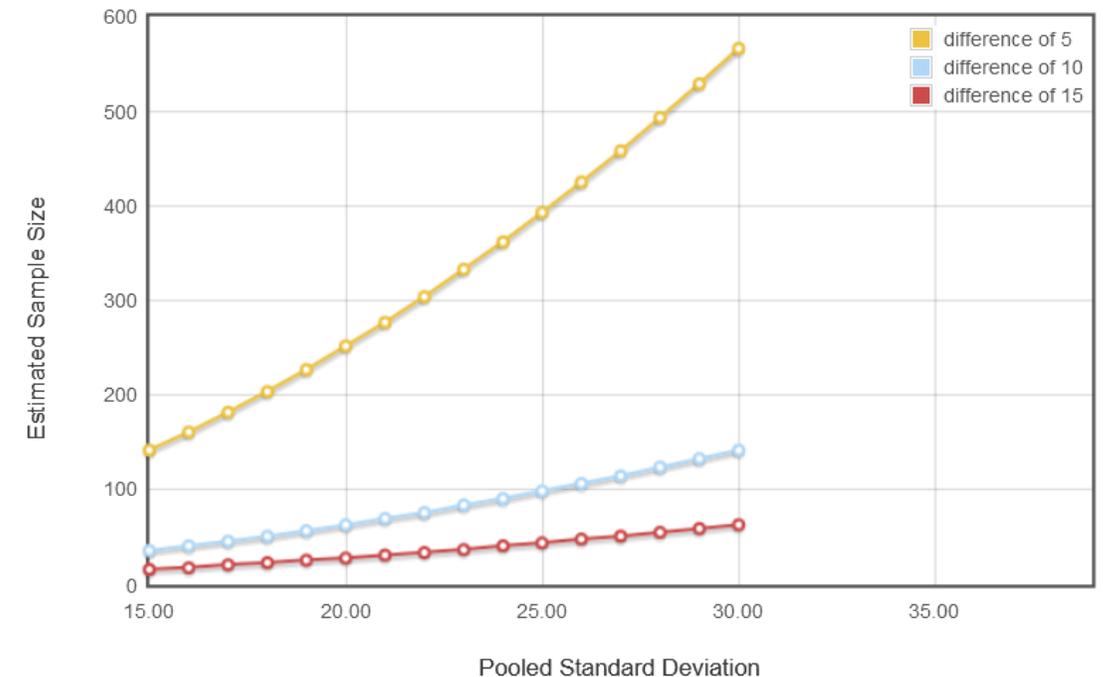
Statulator

Visualisation

This is a plot of sample sizes for a range of pooled Standard Deviations and for three values of Difference of means between groups. Customize the plot by changing input values from the 'Customize Visualisation' panel.

Note that the sample sizes are displayed for only one of the two groups.

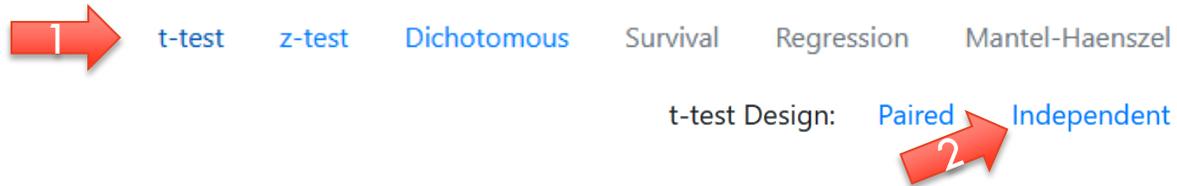
[Download Figure](#)



Note: Statulator used the input values of a power of 80%, a two sided level of significance of 5% and equal group sizes for sample size calculation and adjusted the sample size for t-distribution. You may change the options by clicking [here](#) or the 'Options' button and the adjustments by clicking [here](#) or the 'Adjust' button.

A. Difference between two means

- Choose t-test from the top navigation bar and then 'Independent'



- Under 'What do you want to know?' choose 'Sample size'
- Enter the same inputs as you did in Statulator (including Statulator defaults) and click Calculate

PS Power and Sample Size

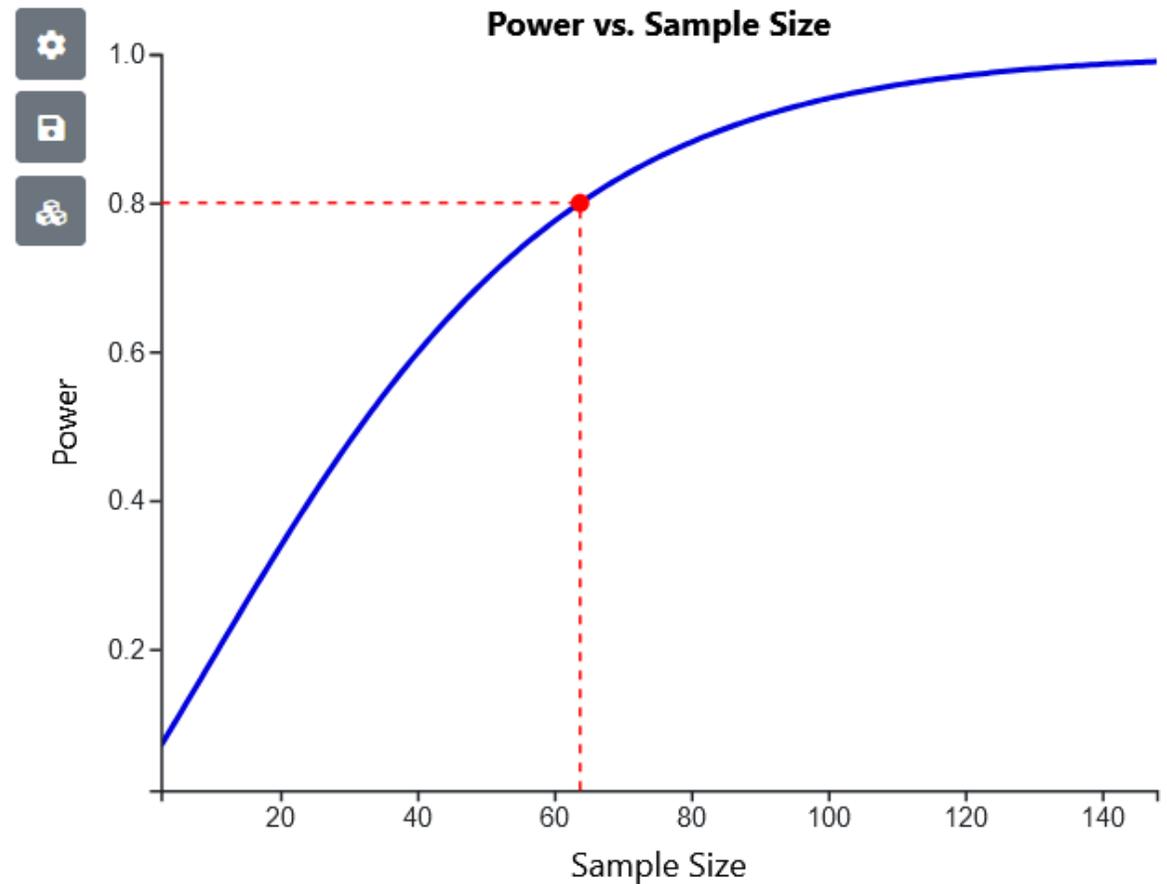
The screenshot shows the 'PS Power and Sample Size' calculator interface. At the top, there are tabs for 'Start' and 'Ind. t-test #1', and an 'Overview ?' button. The main form contains several input fields, each with a red arrow and the number '3' pointing to it: 'What do you want to know?' (dropdown menu set to 'Sample size'), 'Type I Error (α)' (dropdown menu set to '0.05'), 'Standard deviation (σ)' (dropdown menu set to '20'), 'Difference in population means (δ)' (dropdown menu set to '10'), and 'Power' (dropdown menu set to '0.8'). Below these is a text input field for 'Ratio of control/experimental subjects' containing the value '1'. At the bottom of the form is a green 'Calculate' button, with a red arrow and the number '4' pointing to it.

A. Difference between two means

PS Power and Sample Size

- When we click calculate, we get $n_1 = n_2 = 64$ per group, and an interpretation paragraph
- We also see a series of **power analysis** plots that are customisable
- Change the 'output' drop down to Power to see plots with Power on the y-axis
- The top-left plot shows the amount of power for a range of sample sizes
- The sample size needed to achieve our desired power of 80% is highlighted with the red point

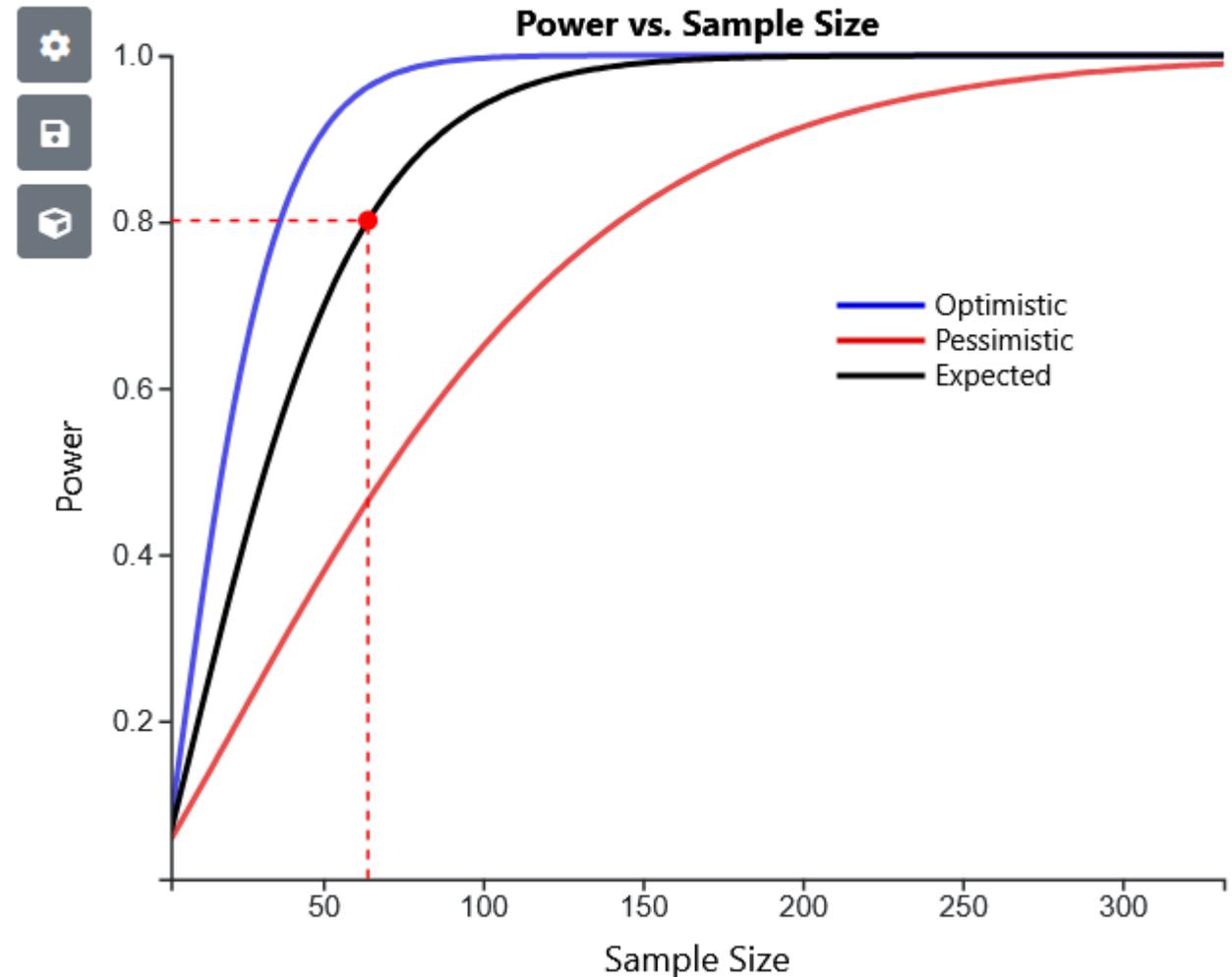
We are planning a study with 64 experimental subjects and 64 control subjects. In a previous study the response within each subject group was normally distributed with standard deviation 20.00. If the true difference in the experimental and control means is 10.00, we will be able to reject the null hypothesis that the population means of the experimental and control groups are equal with probability (power) 0.80. The Type I error probability associated with this test of this null hypothesis is 0.05.



A. Difference between two means

- To this plot we can add two more lines to evaluate our other scenarios for variance
- We evaluate a 'pessimistic' scenario (30 mg/cm³) and an 'optimistic' scenario (15 mg/cm³)
- If we read across at 80% (desired) power, we get an idea of how large our sample size would need to be in each scenario

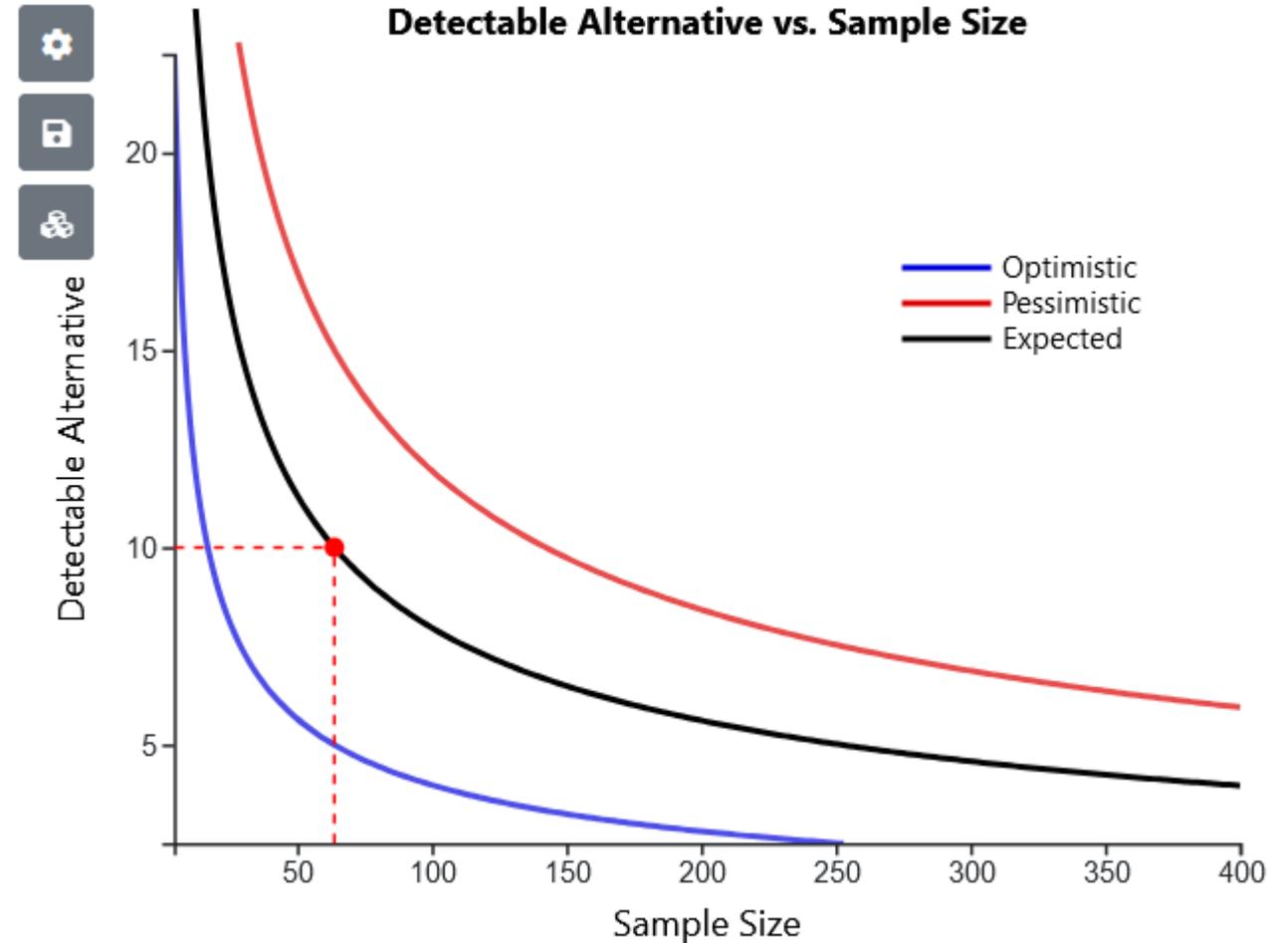
PS Power and Sample Size



A. Difference between two means

PS Power and Sample Size

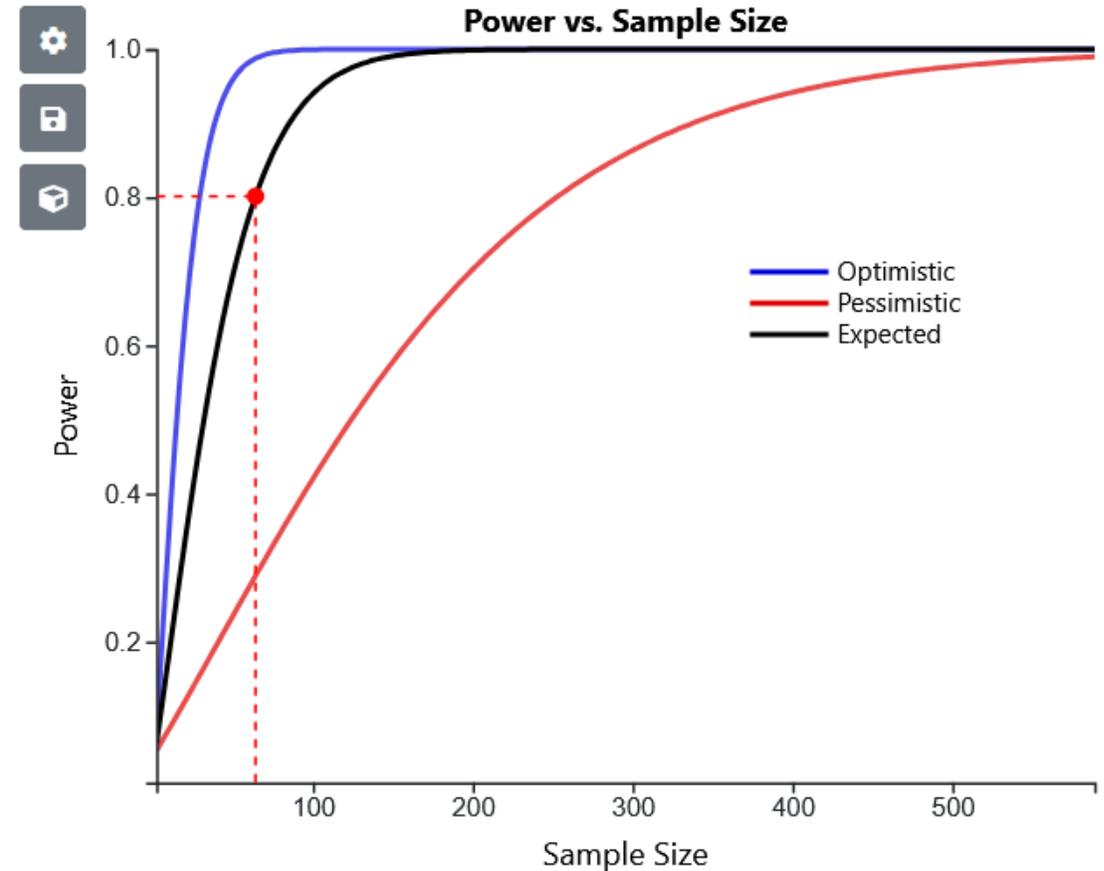
- Another very useful plot is a ‘sensitivity’ plot, showing the sample size vs. the minimum detectable effect size for a given power (in this case 80%)
- This plot shows that for our chosen sample size (of 64 per group):
 - A group difference of 10 will achieve 80% power in the ‘Expected’ scenario
 - A group difference of only 7.5 in the ‘Optimistic’ scenario
 - A group difference of 15 would be needed to achieve 80% power for the pessimistic scenario



We can also evaluate our effect size scenarios

- We can also investigate our optimistic and pessimistic scenarios for effect size
- 3 scenarios for variance x 3 scenarios for effect size = 9 different scenarios, which is too many lines for this graph (only five different colours are supported)
- Instead, we will fix the variance at our expected 20 mg/cm^3 , and evaluate the three effect-size scenarios

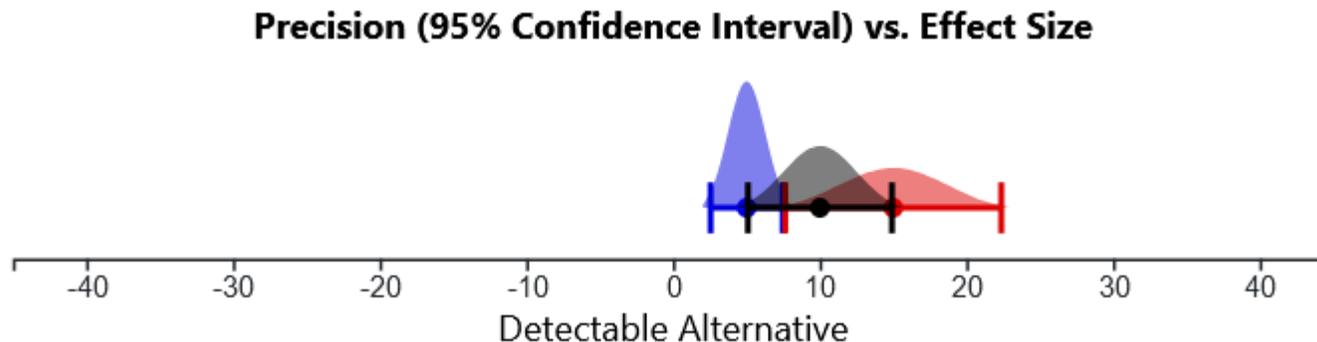
PS Power and Sample Size





A. Difference between two means

- We also see an interesting plot that shows the point estimate, confidence interval, and distribution of sample mean differences we should expect if the true mean difference(s) in the population are exactly as we predict
- While this plot isn't as useful for exploring different scenarios, it is a good reminder of the connection between precision of estimates and hypothesis testing we will eventually perform (recall that the effect size is half the width of the confidence interval)



B. Difference between two means (Mann-Whitney)

The Mann-Whitney U test is a non-parametric version of the t-test for a difference in means. It is based on ranks (also called Wilcoxon rank sum)

This is used when the data are not approximately normally distributed (could be highly skewed), or the underlying distribution is not normal (could be ordinal).

Often used for ordinal data from surveys using Likert response items.

The values of the two groups are combined and ranked. The values are then divided back into the groups and the mean of the assigned ranks for each group is calculated and compared.

The test doesn't use the information about the size of the effect.

B. Difference between two means (Mann-Whitney)

Example: Happiness Survey



2

3



5

6



You want to measure happiness using the Lyubomirsky & Lepper scale. Each item response ranges from 1 (unhappy) to 7 (happy). The score is the sum of 4 items, so the range is 4~28.

A pilot study on two groups produced the following results that can be used for the power calculation (Mean and SD)

	Values		Ranks	
	Single	Married	Single	Married
	12	20	3	1
	11	15	4	2
	10	9	5	6
	6	8	8	7
Mean	9.8	13.0	5	4
SD	2.6	5.6		

B. Difference between two means (Mann-Whitney)

Example: Happiness Survey

You want to apply it to different groups of people (e.g. single vs married) to see if there is a difference in scores.

What is a meaningful difference?

Let's suppose that a minimum difference of 4 points (average of 1 pt difference per item) is the smallest effect size of interest.

B. Difference between two means (Mann-Whitney)

Example: Happiness Survey

So, what are our first 4 steps?

Step 1:	Determine experiment type and statistical test	Two group comparison using Mann-Whitney Group 1: Single Group 2: Married
Step 2:	Set α and $1 - \beta$	0.05 and 0.8
Step 3:	Set the smallest effect size of interest	4 points
Step 4:	Estimate the variance	SD1=2.6, SD2=5.6

B. Difference between two means (Mann-Whitney)

Example: Happiness Survey

Sample size calculation

Heuristic method

“Do the calculations as if performing the corresponding parametric test (i.e. the t-test), then add 15% to the sample size.”

B. Difference between two means (Mann-Whitney)

Example: Happiness Survey

- t-test > Independent
- Enter Difference in population means = 4
- Enter the larger SD (Standard deviation = 5.6)
- Enter other inputs
- Click 'calculate'

The screenshot shows a web-based calculator titled "Power and Sample Size". It has a "Start" button and a tab labeled "Ind. t-test #1". A "Calculate" button is at the bottom. The form contains several input fields, each with a red arrow pointing to it from the left. The fields and their values are: "Sample size" (empty), "Type I Error (α)" (0.05), "Standard deviation (σ)" (5.6), "Difference in population means (δ)" (4), "Power" (0.8), and "Ratio of control/experimental subjects" (1). A "Calculate" button is at the bottom.

Field	Value
Sample size	
Type I Error (α)	0.05
Standard deviation (σ)	5.6
Difference in population means (δ)	4
Power	0.8
Ratio of control/experimental subjects	1

B. Difference between two means (Mann-Whitney)

Example: Happiness Survey

- N=32 per group
- Add 15% for non-parametric. $N=32 \times 1.15 = 36.8$ round up to 37

How and why to calculate within-group standard deviation

- In this example we had two groups with quite different variance/SDs: SD1=2.6, SD2=5.6
- The most conservative choice is to choose the within group standard deviation as the larger SD group (SD2 in this example)
- Unequal variances do not usually cause a problem when group sizes that are equal
- If the group sizes are equal, we can calculate a pooled standard deviation easily using Cohen's formula:

$$\sigma' = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}$$

- In this example: $\sqrt{\frac{2.6^2 + 5.6^2}{2}} = 4.37$, when input results in a sample size of 20 per group
- Add 15% for non-parametric: $20 \times 1.15 = 23$ per group

C. Difference between two proportions

Example: Happiness survey

The survey scores could also be analysed as proportions by considering how many report a value above a threshold (say >14 means “happy”)

Singles group $P1 =$ proportion of subjects respond “happy”

Married group $P2 =$ proportion of subjects respond “happy”

Effect size: Say we want to find a minimum difference in proportions of $P1 - P2 = 0.1$ What sample size is required?

- We also need to estimate the two proportions. Let’s first assume that there will be maximum variance ($p=0.50$)
- Try using $P1=0.55$ and $P2=0.45$. As $p = (p1 + p2)/2$.

C. Difference between two proportions

Example: Happiness survey

What are our first 4 steps this time?

Step 1:	Determine experiment type and statistical test	z-test for proportions
Step 2:	Set α and $1 - \beta$	0.05 and 0.8
Step 3:	Set the smallest effect size of interest	0.10
Step 4:	Estimate the variance	$p_1=0.55, p_2=0.45$

Note: The variance estimate comes from the proportion estimates and is calculated for you.

$$\text{Variance} = p(1-p).$$

C. Difference between two proportions

Example: Happiness survey

Step 5: Statulator

Select Sample Size -> Compare Two Independent Proportions

Calculate Visualise Tabulate

Input Values
Specify input values and click Calculate. Hover over the ? sign to obtain help.

Expected Outcome Proportion in the Reference Group ?

1 → 0.45

Expected Outcome Proportion in the Test Group ?
(Alternatively, specify a [Measure of Association](#))

2 → 0.55

Note: You may change the default options for Power, Significance, Alternate Hypothesis and Group Sizes by clicking the 'Options' button. Click the 'Adjust' button below to adjust sample sizes for continuity (option applied by default), clustering and response rate.

3 →

C. Difference between two proportions

- The same outcome (happiness) that required 32 per group, with a binary variable requires 409 per group
- That's a lot of people
- Tip from the consulting room:
 - With a binary outcome each subject provides only one [computer] bit of information: 0 or 1
 - If you do have an option to use the underlying scale as your outcome, your required sample size for a given power will often be drastically smaller

Results and Live Interpretation

 Download

Assuming that 45% of the subjects in the reference population have the factor of interest, and after applying continuity correction, the study would require a sample size of:

409

for each group (i.e. a total sample size of 818, assuming equal group sizes), to achieve a power of 80% for detecting a difference in proportions of 0.10 between the two groups (test - reference group) at a two sided p-value of 0.05 .

In other words, if you select a random sample of 409 from each population, and determine that 45% and 55% of subjects in the two groups have the factor of interest, you would have 80% power to declare that the two groups have significantly different proportions at a 5% level of significance.

Reference: Dhand, N. K., & Khatkar, M. S. (2014). Statulator: An online statistical calculator. Sample Size Calculator for Comparing Two Independent Proportions. Accessed 6 February 2025 at <http://statulator.com/SampleSize/ss2P.html>

Note: Statulator used the input values of a power of 80%, a two sided level of significance of 5% and equal group sizes for sample size calculation and adjusted the sample size for continuity. You may change the options by clicking [here](#) or the 'Options' button and the adjustments by clicking [here](#) or the 'Adjust' button.

C. Difference between two proportions

- We deliberately chose the proportions that would produce the maximum variance in the outcome for a given group difference (of 10%)
- What if we collected more pilot data, and saw that our expected proportions were closer to: $p_1 = 95\%$ and $p_2 = 85\%$
- The required sample size shrinks to 157 per group, which a lot less but still far more than when using the scale outcome directly

Results and Live Interpretation

 Download

Assuming that 95% of the subjects in the reference population have the factor of interest, and after applying continuity correction, the study would require a sample size of:

157

for each group (i.e. a total sample size of 314, assuming equal group sizes), to achieve a power of 80% for detecting a difference in proportions of -0.10 between the two groups (test - reference group) at a two sided p-value of 0.05 .

In other words, if you select a random sample of 157 from each population, and determine that 95% and 85% of subjects in the two groups have the factor of interest, you would have 80% power to declare that the two groups have significantly different proportions at a 5% level of significance.



C. Difference between two proportions

There are multiple options when comparing independent proportions

G*Power provides a total of 4 options for power calculations for proportions with independent groups:

- Inequality, z-test (used in Happiness intervention example)
- Inequality, Fisher's Exact test
- Inequality, Unconditional exact
- Inequality with offset, Unconditional exact

The Fisher's Exact test should be used when sample sizes are going to be small (say $n_1p_1 \leq 5$ or $n_2p_2 \leq 5$)

- The Fisher's Exact result for the Happiness example is shown on the next slide for your reference



C. Difference between two proportions

Example: Happiness survey

Step 6: Use the Fisher's Exact test to get the sample size with

$P_1=0.85$ and $P_2=0.95$

Fisher's Exact suggests 159 subjects per group.

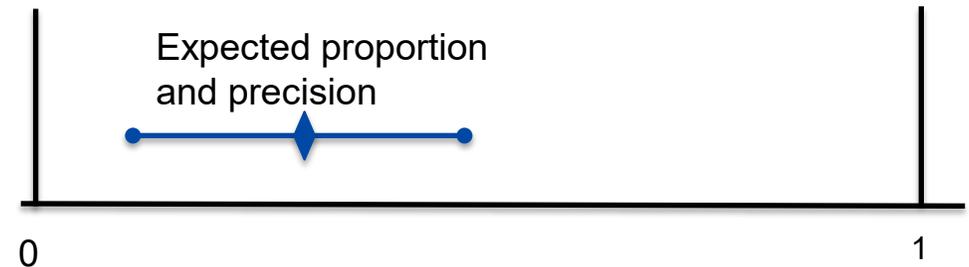
Not quite the same result as the z-test, but very close

The screenshot shows a software interface for calculating sample size for Fisher's exact test. The parameters are as follows:

Parameter	Value
Output	Sample size
Type I Error (α)	0.05
Sample size (Computed value)	159
Power	0.8
Probability of the outcome for a control patient (p_0)	0.85
Probability of the outcome for an experimental patient (p_1)	0.95
Ratio of control/experimental subjects (m)	1

D. Estimation of a single proportion

- A common sample size calculation that does not involve a hypothesis test
- Proportions are encountered often:
 - Prevalence of disease in a population
 - Proportion of people failing a screening test
 - Sensitivity and specificity of a diagnostic test
- The sample size calculation is to estimate a proportion to a given level of precision
- Precision in this context is the width of the confidence interval (usually 95% confidence interval)



D. Estimation of a single proportion

- Lets continue with the happiness example, but estimate the prevalence of happiness in the general population
- We need to generalise our workflow for sample size to a given precision
 1. Determine experiment type and statistical test **or population property to estimate**
 2. Set α [~~and $1 - \beta$~~]
 3. Set the smallest effect size of interest **or required precision**
 4. Estimate the variance
 5. Calculate the minimum sample size
 6. Explore scenarios

D. Estimation of a single proportion

Example: Happiness survey

Step 5: Statulator

Select Sample Size -> Estimate a Single Proportion

Step 1:	Determine population property	Proportion of happy people in general population
Step 2:	Set α	0.05 (95% CI)
Step 3:	Set the required precision	0.10 (10% CI width)
Step 4:	Estimate the variance	Maximal at $p=0.5$

Calculator Visualisation Tabulate

Input Values

Specify input values and click Calculate. Hover over the ? sign to obtain help.

Level of Confidence ?

1 → 0.95

Expected Proportion ?

2 → 0.5

Precision or Margin of Error ?

Absolute value 0.05 ← 3

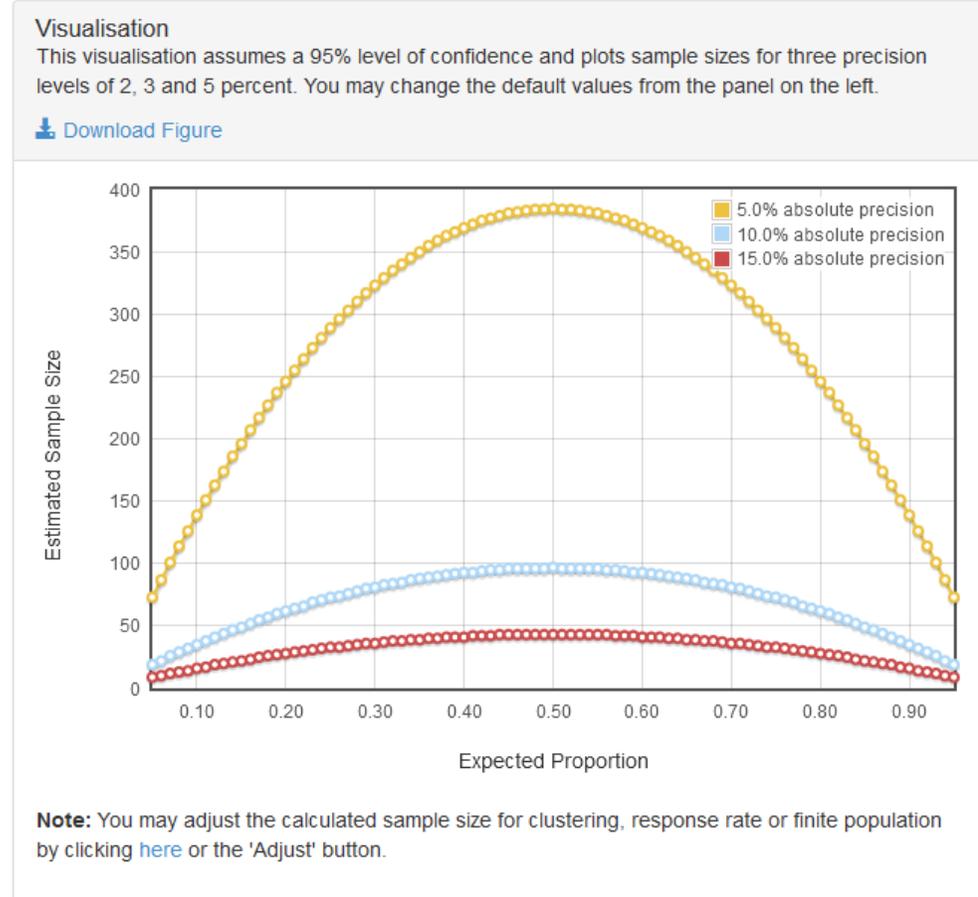
Note: You may adjust sample sizes for finite population, clustering and response rate by clicking the 'Adjust' button below.

4 → Calculate Adjust Reset

The precision, specified here as an 'absolute value' is half the width of the desired confidence interval

D. Estimation of a single proportion

- The sample size required is 385 for a 10% wide confidence interval
- Again, the maximal variance is when the $p=0.5$, so if we have a better idea of our expected p , the required sample size may be less than this



Common Pitfalls

Tips from the consulting room:

- Underestimating variance → underpowered studies
- Ignoring dropout → fewer completers than planned
- Choosing unrealistic SESOI → waste resources or miss meaningful effects
- Blindly using defaults ($\alpha=0.05$, power=0.8)
- Multiple testing not accounted for → inflated Type I error
- Over-reliance on software (or GenAI) without understanding assumptions
- Not exploring scenarios → sensitivity to variance/effect size assumptions





Handling dropout

Tip from the consulting room:

- It is common, especially in clinical trials to inflate the required sample size to account for an expected rate of dropout
- One pitfall is to inflate the minimum sample number (N) by the expected dropout rate (W). After dropouts this would not achieve the required sample size of completers
- Instead we need to use the formula for N** for the inflated sample size that will restore the required sample size of completers after dropouts

$$N^{**} = \frac{N}{1 - W}$$

Worked example for N=100, and W=15%

Wrong:

$100 \times 115\% = 115 \times 15\% = 17.25$ so 18 dropouts and 97 completers

Right:

$100 / 85\% = 117.64$, so 118 required
 $118 \times 15\% = 17.7$ dropouts and 100 completers

Handling multiple testing

- Multiple testing corrections, must be incorporated into your power analysis
- The easiest way to incorporate these is to change the alpha in the sample size calculation to match the reduced 'per test' alpha needed to achieve the Family Wise Error Rate (FWER) alpha
- For Bonferroni correction, the 'per test' alpha is simply the FWER alpha divided by the number of tests
- For more complex multiple testing corrections, you may need to chose a different alpha to input
- You also need to consider your study goals, and your analysis plan. Feel free to have a consult to discuss further

Power Analysis for other designs

Statulator scope:

- Two group [in]dependent means, [in]dependent proportions, single proportions

Power and Sample Size (PS) scope:

- As above but also OR under prospective or retrospective designs, precision of confidence intervals

G*Power scope:

- Most of above but also ANOVA, Correlation, Linear Regression, Logistic Regression

Refer to the manual for details. Search for “Gpower manual” online or find it [here](#)



Effect Sizes for other designs

Effect size for ANOVA

G*Power uses the standardised effect size; Cohen's f is related to the partial eta squared

$$\eta^2 = \frac{f^2}{(1 + f^2)}$$

Partial eta squared is often reported in the ANOVA table output

Recall that this F test is rarely used to power a study. Instead choose the two groups that are most similar and power a t-test to detect the difference between them (after accounting for multiple testing in post-hoc).

Effect size for other designs: Use a wide variety of effect size measures

Power Analysis for other designs

From simple designs to complex designs

So far we have considered power analysis for simple designs where the mathematical calculations are tractable and rely on a limited set of assumptions regarding the data to be obtained.

As design complexity increases, it becomes more difficult or perhaps impossible to find an analytical solution to calculate power.

When no formula exists:

- First option – determine sample size for a simplified version of the study design and extrapolate this to the more complex design
- Second option – Use a simulation method (that does not rely on formulae)

Power Analysis by simplification

A couple of examples

ANOVA: Choose the two groups that are most similar and power the post-hoc test to detect the expected difference between them. All other post-hocs and the F test should have sufficient power.

Multiple regression:

- For a categorical factor of interest choose the factor of interest and power to detect the difference as a t-test (the same as a univariate model). The addition of covariates should only increase the power of your actual analysis
- If you have an idea of the variance explained by your factor of interest and the residual variance you can use the linear multiple regression module of G Power

Linear mixed models:

- If you have a repeated measures experiment, power the study as if you only had one measurement. The addition of repeated measures should only increase the power of your actual analysis

Switch to simulation methods for complex study designs where analysis of a simplified design is not sufficiently rigorous.

Power Analysis – by simulation

Simulation based power estimation

- Simulate (many) data sets
- Analyse each data set and test for statistical significance
- Calculate the proportion of simulations with significant p values

$$Power = \frac{\textit{significant simulations}}{\textit{all simulations}}$$

- The ‘trick’ is to set the parameters of the simulation in a sensible, realistic way

<https://link.springer.com/article/10.3758/s13428-021-01546-0>

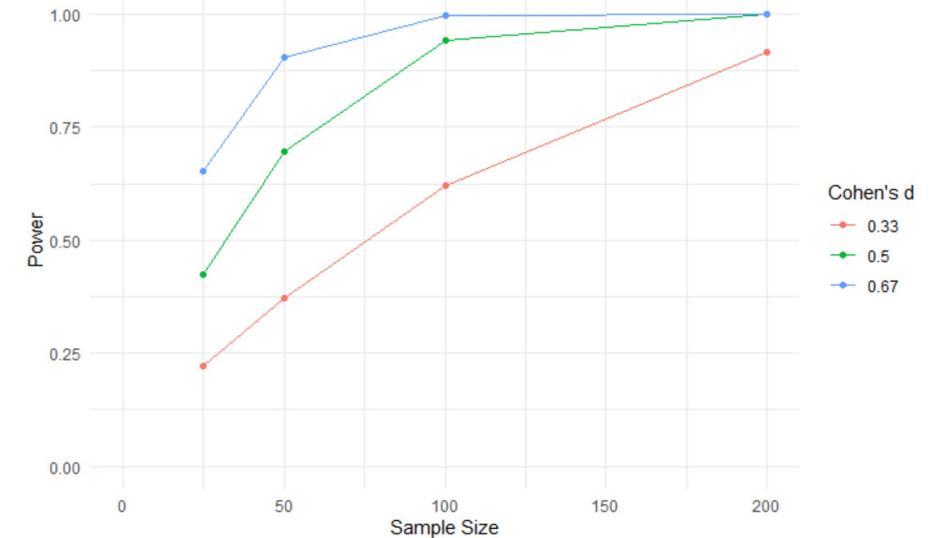
Power Analysis – by simulation

Example A: Chicken Welfare - bone density (difference between two means)

- Simulation in R using package “paramtest”
- Results for this simple simulation will be very similar to those obtained from G*Power.
- See R Markdown files for details

Generalised linear models

- You can use the package **simr**
- Specify the model as you would for analysis
- **Simr** then simulates data from that model



Software for Power Analysis

Free and Open Source software

- R /R Studio:
 - Base R has functions covering basic proportions, t-tests, etc.
 - Package “pwr” has 9 functions covering proportions, t-tests, ANOVA, chi-square and correlations
 - Package “epiR” has 23 functions covering many statistics including AUC, sensitivity and specificity
 - Package “paramtest” basic power calculations by simulation
 - Package “mixedpower” for generalised linear mixed models
 - Package “simr” simulation based power calculations for mixed models
- Statulator (from Sydney Uni) <https://www.statulator.com/SampleSize/>
- PS (from Vanderbilt University) <https://cqsclinical.app.vumc.org/ps/>
- G*Power is a (free) program to download (from Dusseldorf Uni) <https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower>
- Online calculators such as www.powerandsamplesize.com and <https://sample-size.net/>

Software for Power Analysis

Proprietary \$\$ software

- Packages such as STATA, SPSS and SAS include a calculator
- GraphPad have “StatMate” separate to Prism
- PASS by NCSS dedicated software esp. for medical research



Further Assistance at Sydney University

SIH

- [Statistical Consulting website](#): containing our workshop slides and our favourite external resources (including links for learning R and SPSS)
- [Hacky Hour](#) an informal monthly meetup for getting help with coding or using statistics software
- 1on1 Consults can be requested [on our website](#) (click on the big red ‘contact us’ link)

SIH Workshops

- Create your own custom programmes tailored to your research needs by attending more of our Statistical Consulting workshops. Look for the statistics workshops on [our training page](#).
- [Other SIH workshops](#)
- [Sign up to our mailing list](#) to be notified of upcoming training

Other

- Open Learning Environment (OLE) courses
- [Linkedin Learning](#)



A reminder: Acknowledging SIH

All University of Sydney resources are available to Sydney researchers **free of charge**. The use of the SIH services including the Artemis HPC and associated support and training warrants acknowledgement in any publications, conference proceedings or posters describing work facilitated by these services.

The continued acknowledgment of the use of SIH facilities ensures the sustainability of our services.

Suggested wording for use of workshops and workflows:

“The authors acknowledge the Statistical workshops and workflows provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney.”