Power and Sample Size Calculation

Presented by
Alex Shaw
Sydney Informatics Hub
Core Research Facilities
The University of Sydney

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.”

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 can 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 (e.g. experimental design, power calculation, model building, analysis using linear models/survival/multivariate/survey methods)

Statistical Workflows

- Our statistical workflows can be found within our workshop slides

- Statistical workflows are software agnostic, in that they can be applied using any statistical software

- There may also be accompanying software workflows that show you how to perform the statistical workflow using particular software packages (e.g. R or SPSS). We won't be going through these in detail during the workshop. If you are having trouble using them, we suggest you attend our monthly Hacky Hour where SIH staff can help you.

During the workshop

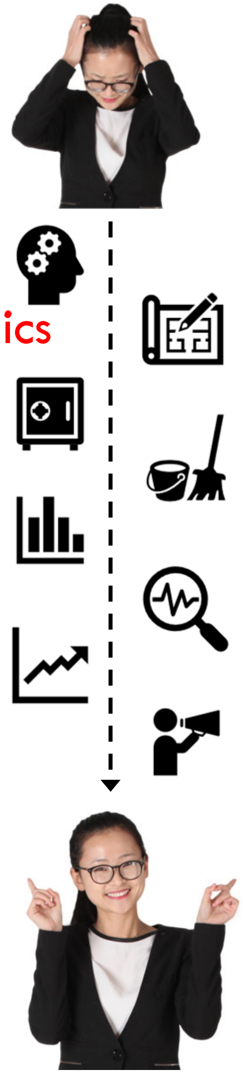
Ask short questions or clarifications during the workshop (either by Zoom chat or verbally). There will be breaks during the workshop for longer questions.

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
- Software tools - G*Power
- Example 1: Difference between 2 means (t-test)
- Example 2: Difference between 2 means (Mann-Whitney)
- Example 3: Difference between 2 proportions (z-test)
- Power calculation for other designs
- References

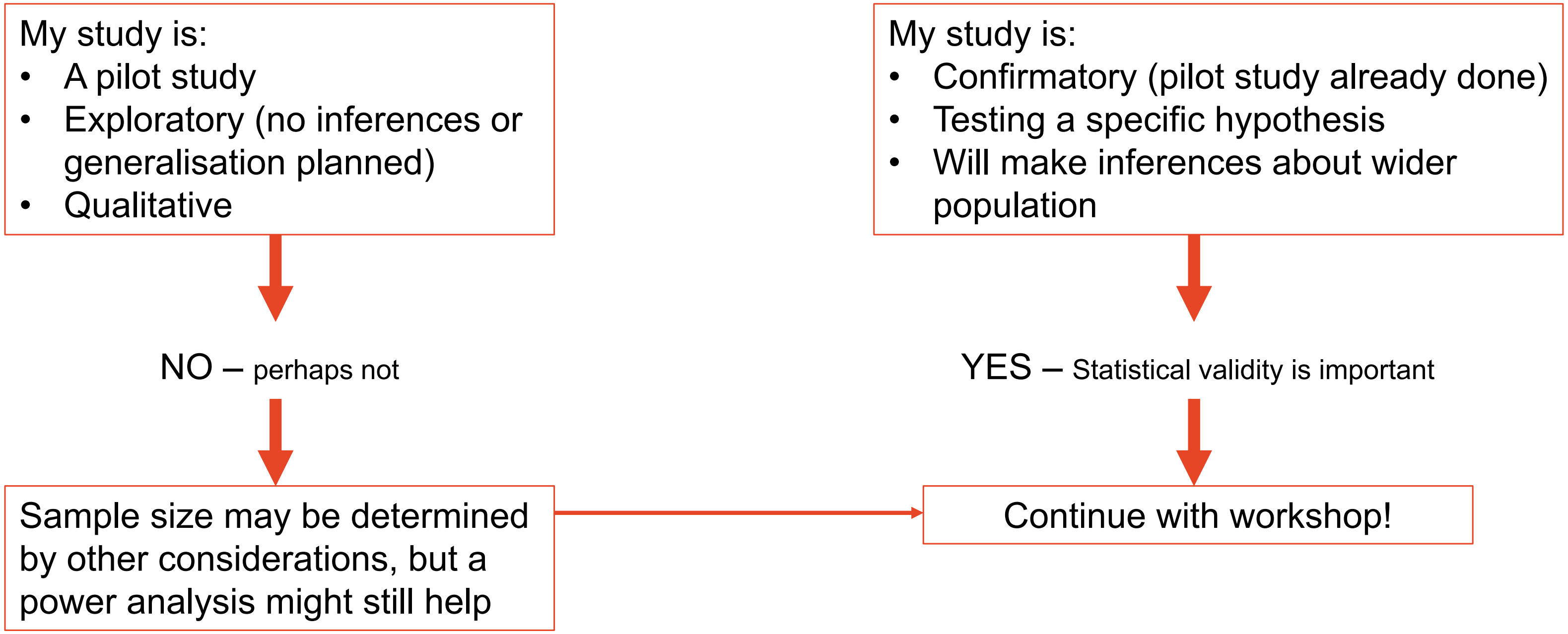
Why do we need to calculate power and sample size?

Why do we want to estimate the power of an experiment?

- To know if it is worth doing the experiment
- To plan the time and resources necessary
- To make sure we are not wasting our time
- To get a grant application approved
- To make sure the study design is ethically acceptable

But do I <u>really</u> need to calculate power?

What type of study are you planning?



My study is:

- A pilot study
- Exploratory (no inferences or generalisation planned)
- Qualitative

NO – perhaps not

Sample size may be determined by other considerations, but a power analysis might still help

My study is:

- Confirmatory (pilot study already done)
- Testing a specific hypothesis
- Will make inferences about wider population

YES – Statistical validity is important

Continue with workshop!

What is the power of an experimental design?

The power to know…

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

"The means of two groups are different"

In statistics, this is referred to as the alternative hypothesis H_1.
Classically, we test the veracity of the null hypothesis:

H_0: There is no difference between the means of the two groups

A statistical test of the null hypothesis is always subject to **uncertainty.** When we make a decision based on a hypothesis test, we may make an error. There are two main types of error.

Types of statistical error

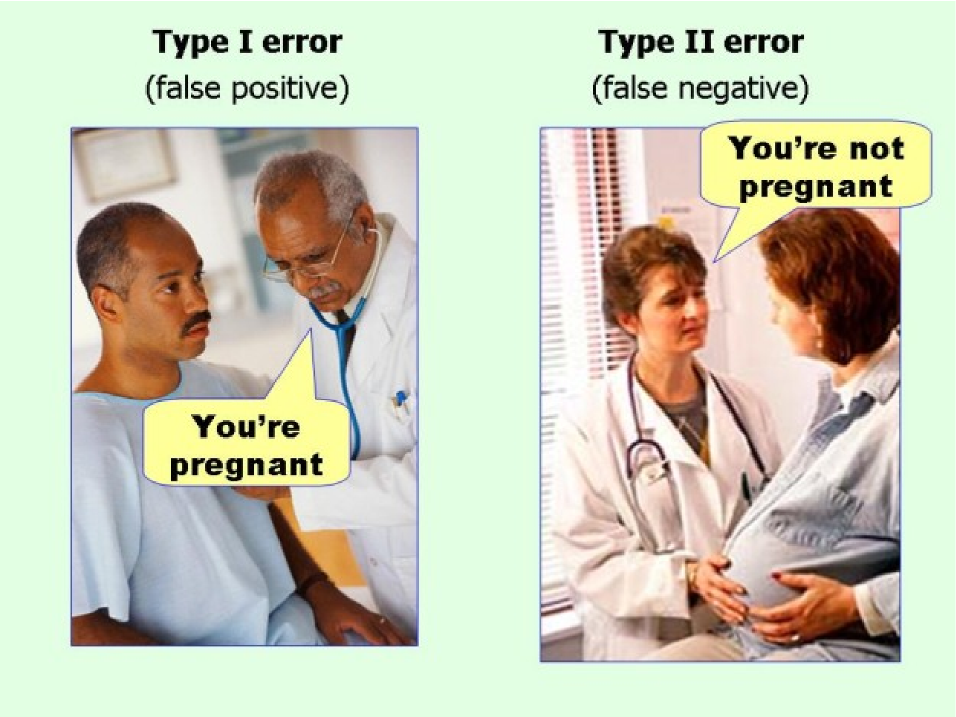
Type I error

- Incorrectly rejecting the null hypothesis
- Also called a false positive
- Rate of false positives designated by the signficiance level, α
- 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 an experiment

Type II error

- Incorrectly accepting the null hypothesis
- Also called a false negative
- Rate of false negatives designated by β
- Power is the complement of β, denoted by $1 - \beta$
- We want Power to be as high as possible, typically $1 - \beta > 0.8$, at 0.8, even when the null hypothesis is false, it will not be rejected one in every five runs of an experiment

Types of statistical error



Types of statistical error

When we perform a null hypothesis test, we are setting up a binary choice that can result in these types of error.

Types of statistical error

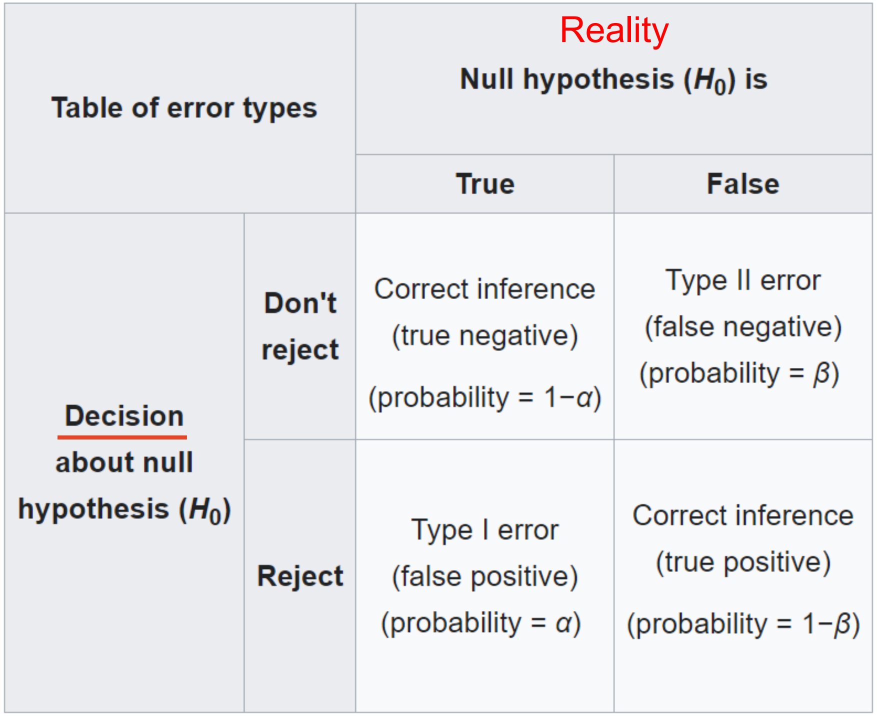
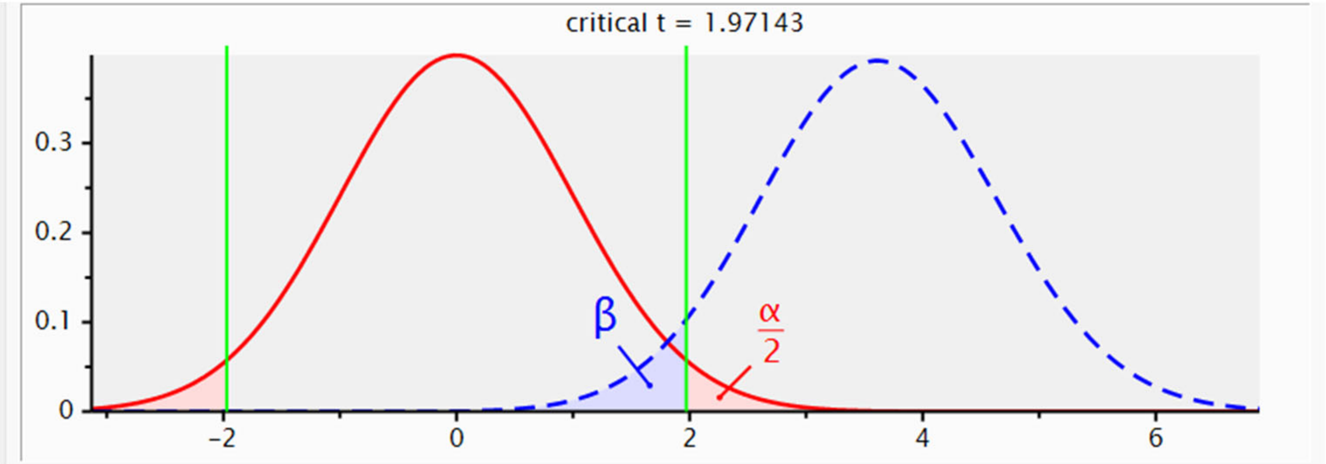


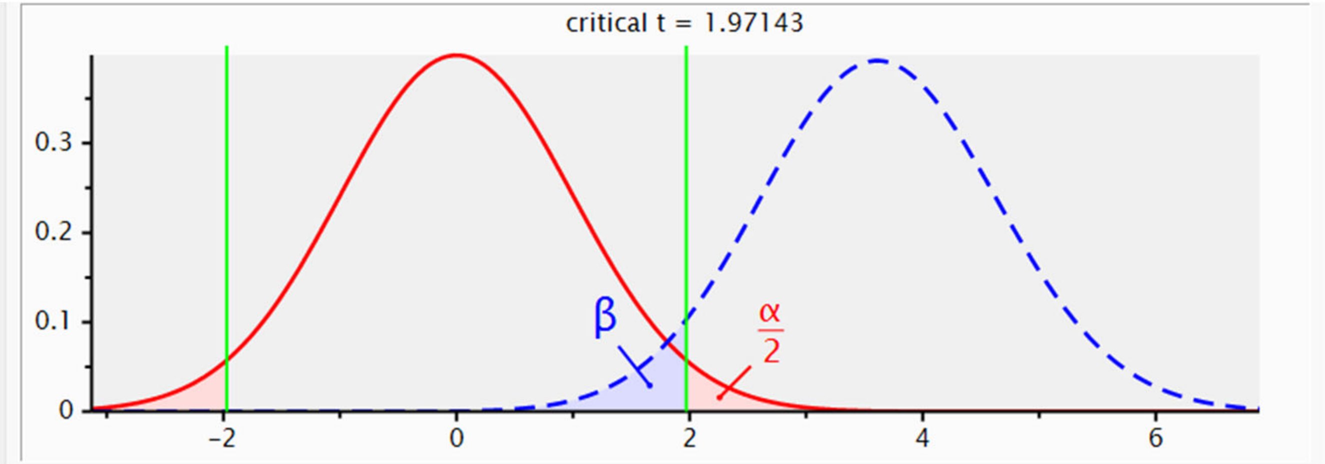
Table of error types		Reality Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference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

Looking ahead



- The above plot is from G*Power and shows two distributions.
- The red distribution is how we expect our test statistic (in this case t statistic) to be distributed (over many experimental runs) when the null hypothesis is true
- Notice the green vertical lines, when the t statistic is beyond the green lines, the null hypothesis is rejected. We see the red shaded area represents alpha (false positives).

Looking ahead



- Now have a look at the blue, dotted distribution. This represents a particular scenario where the null hypothesis is false. The blue shaded area represents false negatives.
- Think about the asymmetry of the blue and red shaded areas. We are usually more accepting of false negatives than false positives when it comes to confirmatory research.

Types of hypothesis

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

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)

- Equivalence trials
- Non-inferiority trials
- As-good-as-or-better trials
- Bioequivalence trials
- Trials to a given precision

The hypothesis tests that apply will vary depending on the study objective.

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

Hypothesis test or estimation of effect size?

What if we don’t want to perform a hypothesis test?
What if we just want to estimate group means for example?

The same power calculation process can be applied.
We will consider why later.

Power calculation

How do we estimate the power of an experiment?

- It will depend on:
  - Sample size (more samples = more power)
  - Chosen significance level (typically $\alpha = 0.05$)
  - Minimum effect size to detect (larger minimum effect = more power)
  - Variance within groups (larger variance = less power)
  - Experimental design and type of statistical hypothesis test

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

Sample size calculation - workflow

Often we want to calculate a sample size given a required minimum power

Sample size calculation workflow steps

1. Determine experimental design 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

Understanding the types of outcome variables and explanatory variables and their distribution is covered in the “**Research Essentials**” workshop.

For example:

Experimental Design	assumptions	proposed statistical test
Comparison of 2 means	independent groups, normally distributed outcome	Student’s t-test
Comparison of 2 means	independent groups, no assumption of normality	Mann-Whitney U test
Comparison of 2 proportions	independent groups	z-test
Comparison of means, more than 2 groups	independent groups normally distributed outcome	ANOVA, F-test

2. Set α and $1 - \beta$

Setting values of parameters

- Typically choose $\alpha = 0.05$ (or lower)
- Sometimes α is set to 0.01
- Typically choose $1 - \beta = 0.8$ (or higher)
- Sometimes power $(1 - \beta)$ is required at 0.90 or 0.95

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 replicated.
  - *Consider increasing 1- 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 by repeating the study in a new and larger sample.
  - Consider increasing α to *minimise the chance of a false negative. A false positive is less of a problem, because the finding will be replicated*

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, as all differences detected on the high throughput platform will be verified using a lower throughput technology.

2. Set α and $1 - \beta$

It’s 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 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 of .05.*** *The 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.

3. Set the smallest effect size of interest

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*.

- 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

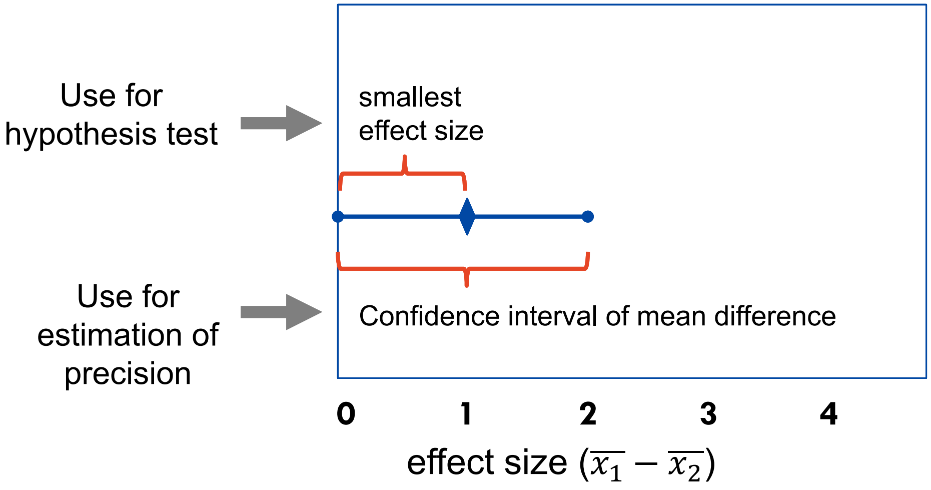
Effect size chosen is larger than necessary

- The sample size is too small
- May detect large effects only
- Not able to achieve statistical significance for small effect sizes of interest
- Could be a waste of resources
- Will lead to a higher Type I error rate over the long run (poor reproducibility)

goldilocks

3. Set the smallest effect size of interest

Effect size – what it means for the hypothesis test and for the estimation of effect size. In the plot below for a mean difference, the null hypothesis is rejected when the CI does not cross 0. This depends on the estimated mean difference, and the CI width.



Use for hypothesis test

smallest effect size

Use for estimation of precision

Confidence interval of mean difference

effect size $(\overline{x_1} - \overline{x_2})$

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**

The minimum confidence interval width is twice the smallest effect size

4. Estimate the variance

Within study variance may be the big unknown in this calculation

How to estimate it?

- Estimate standard deviation (or proportions) from previous experiments?
- Consider theoretical bounds (eg for 5pt scales, proportions)
- Simulate some data and evaluate possible scenarios
- 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

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

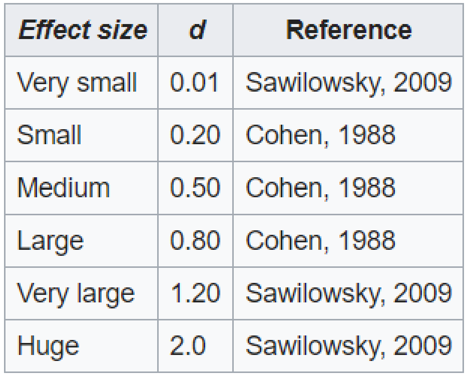
Cohen's d is therefore analogous to the number of standard deviations difference, or the z-score difference. Also called the standardised mean difference (SMD).

Cohen's d will be calculated within G*Power software we will use later in scenarios where you have an expected effect size and SD.

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.



Effect size	*d*	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.

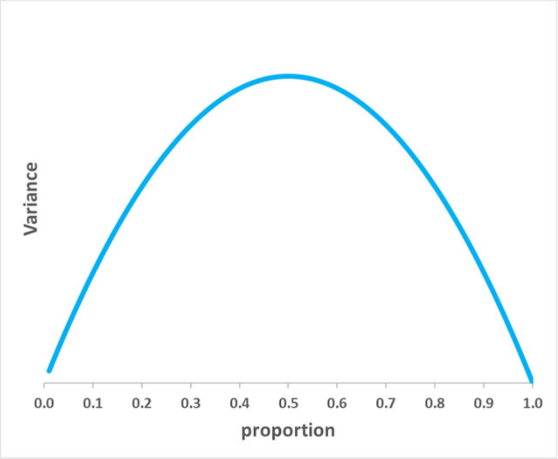
https://en.wikipedia.org/wiki/Effect_size#Cohen's_d

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.

$$Variance(p) = p(1-p)$$



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, the precision is smaller 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

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

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

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 (we will use **G*Power** in this workshop)
- 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).

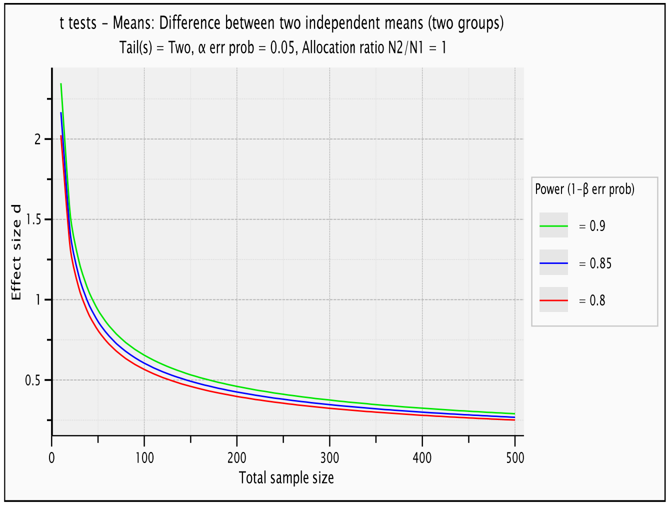
$$n = \frac{2\,(Z_{\alpha/2} + Z_{1-\beta})^2}{\left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2}$$

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 $1 - \beta$ or vice-versa.

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

- Don’t just calculate a single sample size n!
- Use the software to calculate n for a range of scenarios in order to explore the consequences of uncertainty in the values used in the calculation
- This is called a **Power Analysis**
- *Consider also the shape of the cost curve for sample data collection*



For the above example note:
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.

Recap

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

Examples using G*Power software

We will work through 3 simple examples

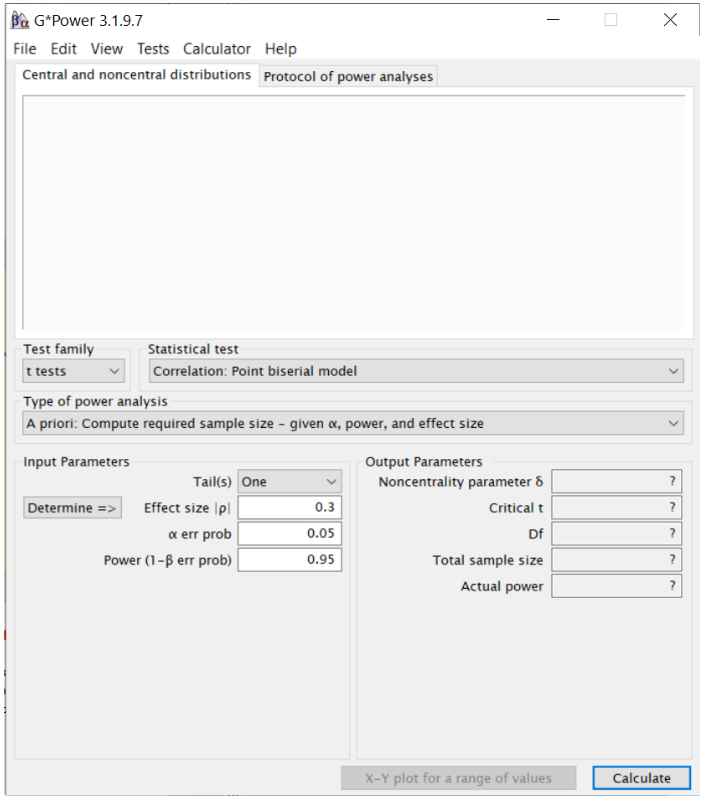
1. Difference between 2 means (continuous response)
2. Difference between 2 means (survey response)
3. Difference between 2 proportions

Followed by a discussion of what to do when your study is more complicated than this

Power calculation software

G*Power

- Download from website:
- http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html
- Current release 3.1.9.7 (Windows) 17 March 2020 (and 3.1.9.6 for Mac)
- Program has a simple user interface
- There is also a manual available online: http://www.psychologie.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf


G*Power 3.1.9.7
File
Edit
View
Tests
Calculator
Help
Central and noncentral distributions
Protocol of power analyses
Determine =>
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Correlation: Point biserial model

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

Input Parameters

Tail(s)	One
Effect size |ρ|	0.3
α err prob	0.05
Power (1-β err prob)	0.95

Output Parameters

Noncentrality parameter δ	?
Critical t	?
Df	?
Total sample size	?
Actual power	?

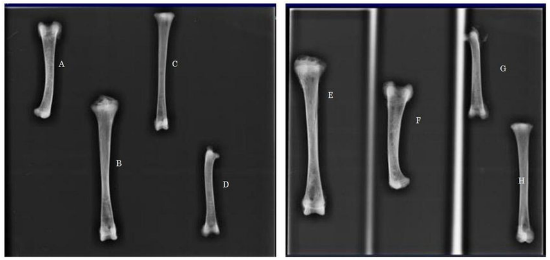
1. Difference between 2 means

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/cm^3

<u>Treatment Group</u> = high mineral diet
<u>Control Group</u> = normal 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/cm^3?

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 - Mbajiorgu, 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 -

1. Difference between 2 means

Example: Chicken Welfare – Bone density

- Step 1: We will use a t-test (assume normality)
- Step 2: α=0.05 and $1 - \beta$=0.8
- 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 mg/cm^3
- Assume we will have equal size groups, n1=n2

1. Difference between 2 means

Step 5: Calculate the minimum sample size

- Put all the information into G*Power
- Note: G*Power will convert the difference in means with the estimated SD to a standardized effect size called Cohen's d.
- G*Power always works with standardised effect sizes, but has additional pop-out dialogue boxes for you to calculate standardised effect sizes from the original scale of your outcome
- The *disadvantage* of this approach is that the effect size and the variance are effectively combined in your power analysis outputs*

* There are workarounds you can use, but if this is a deal-breaker for you, have a look into alternative software that is not based on standardised effect sizes (some of these are listed at the end of the presentation).

1. Difference between 2 means

Step 5: G*Power

G*Power will use this formula to calculate the sample size:

$$n = 2\frac{\delta^2}{d^2}$$

where:

n = sample size per group (when n1=n2)

δ = non-centrality parameter (of the t statistic, based on α, β and group difference)

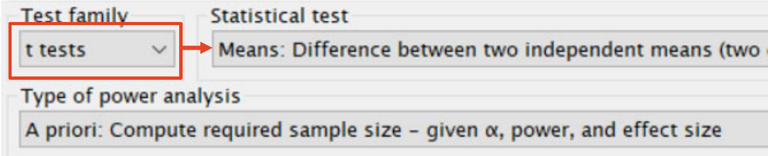
d = standardised effect size (Cohen’s d)

1. Difference between 2 means

Step 5: G*Power

There are two ways to find the correct test

- Distribution approach: Select the test family (eg t tests), then the statistical test

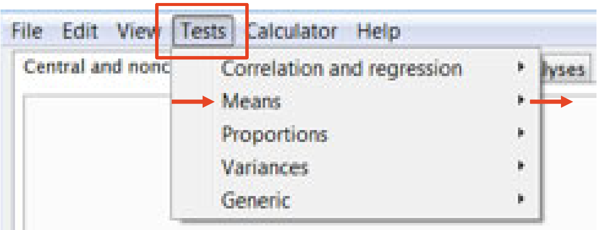


Test family: t tests

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

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

- Design based approach: Select the test parameter class (eg means), then the study design
- Select Tests/Means/Two independent groups



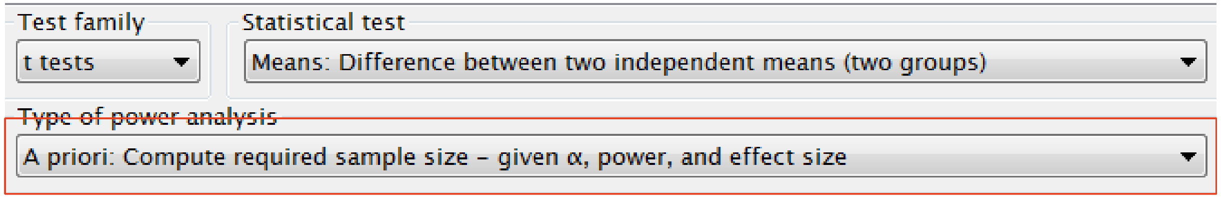
1. Difference between 2 means

G*Power

There are five different types of power analysis

- A priori
- Compromise
- Criterion
- Post Hoc
- Sensitivity

The “A priori” type is suitable for sample size calculation



Test family: t tests

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

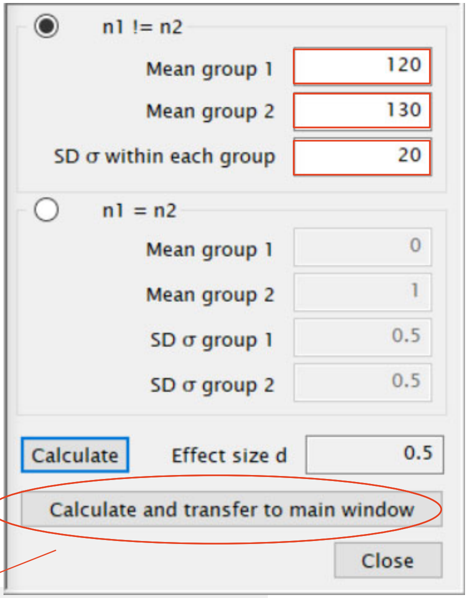
Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

1. Difference between 2 means

Example: Chicken Welfare – Bone density

Enter the values for the chick experiment

- Use α=0.05 and $1 - \beta$=0.8
- Allocation ratio N2/N1=1
- Open the "determine" window to calculate the effect size d. Use mean group 1=120, mean group 2=130, SD1=SD2=20, "calculate and transfer" (on popout window)
- Effect size is now shown d=0.5, select "two" tails, "Calculate" (on main window)



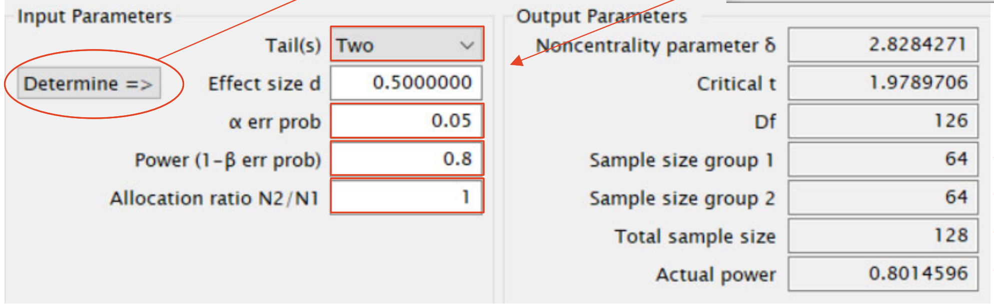
● n1 != n2

Mean group 1	120
Mean group 2	130
SD σ within each group	20

○ n1 = n2

Mean group 1	0
Mean group 2	1
SD σ group 1	0.5
SD σ group 2	0.5

Effect size d 0.5



Input Parameters

Tail(s)	Two
Effect size d	0.5000000
α err prob	0.05
Power (1-β err prob)	0.8
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	2.8284271
Critical t	1.9789706
Df	126
Sample size group 1	64
Sample size group 2	64
Total sample size	128
Actual power	0.8014596

Calculated results

1. Difference between 2 means

Example: Chicken Welfare – Bone density

- Group sample sizes are N1=64, N2=64
- Actual power = 0.8015
- G*Power rounds up the sample size to the nearest integer, so actual power is slightly higher than the minimum requested.

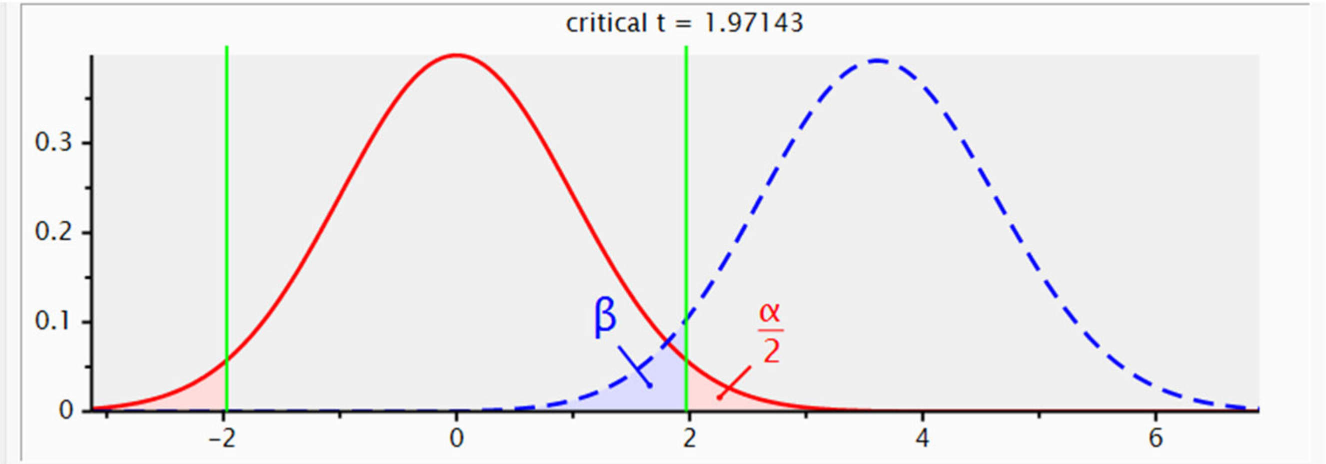
Protocol of the power analysis

- You may want to save a copy of the calculation from this window (at the top right)

Central and non central distributions

- You may be interested to check the visual display of the test statistics in this window (at the top left)

1. Difference between 2 means
Central and non central test statistic distribution



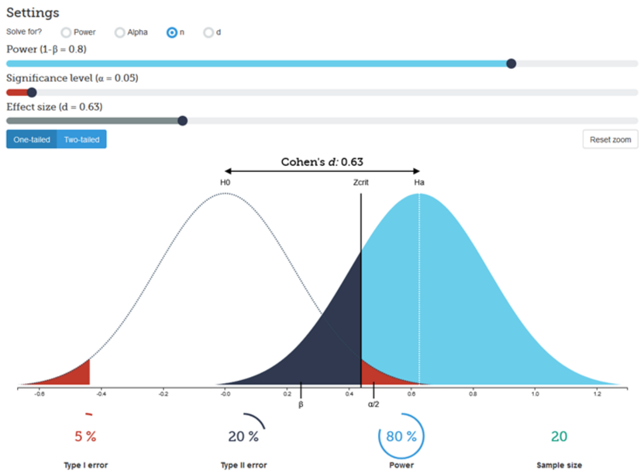
The central distribution of a test statistic (in red) describes how a test statistic is distributed when the null hypothesis is true.

The non central distribution (blue dashed line) describes how the test statistic is distributed when the null hypothesis is false (alternate hypothesis is true).

Shows the distributions with the minimum effect size threshold (green lines). Notice that the alpha is distributed across two tails (alpha/2). We almost always choose two-tailed, because it is *possible* the effect could be positive or negative.

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.



Settings

Solve for? ○ Power ○ Alpha ● n ○ d

Power (1-β = 0.8)

Significance level (α = 0.05)

Effect size (d = 0.63)

One-tailed | Two-tailed

5 % Type I error

20 % Type II error

80 % Power

20 Sample size

1. Difference between 2 means

Example: Chicken Welfare – Bone density

Step 6: Explore scenarios

Power Analysis

- It is advisable to explore some different scenarios for different experimental settings.
- Consider how much your within study standard deviation could vary from your point estimate
  - Our estimate is SD = 20
  - Possible min value = 15 (optimistic)
  - Possible max value = 30 (pessimistic, conservative)

1. Difference between 2 means

Example: Chicken Welfare – Bone density

For G*Power we will use Cohen’s d values to match the possible range of SD values

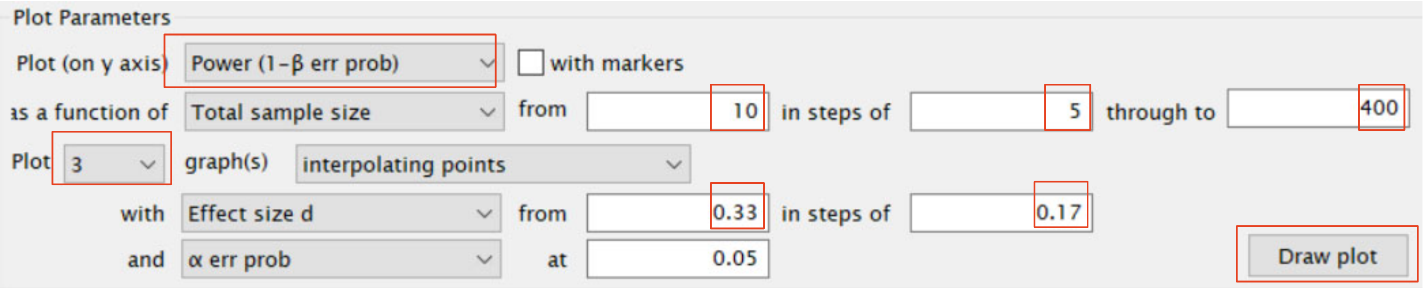
Min	SD = 15	d = 10/15 = 0.67
Expected	SD = 20	d = 10/20 = 0.5
Max	SD = 30	d = 10/30 = 0.33

1. Difference between 2 means

Example: Chicken Welfare – Bone density

X-Y Plot for a range of values

- Plot (on y axis) change to "power"
- Sample size from 10 to 400 in steps of 5
- Plot "3" graphs with d = 0.33 in steps of 0.17 (gets us to 0.5 and 0.67)



Plot Parameters

Plot (on y axis) Power (1−β err prob) □ with markers

as a function of Total sample size from 10 in steps of 5 through to 400

Plot 3 graph(s) interpolating points

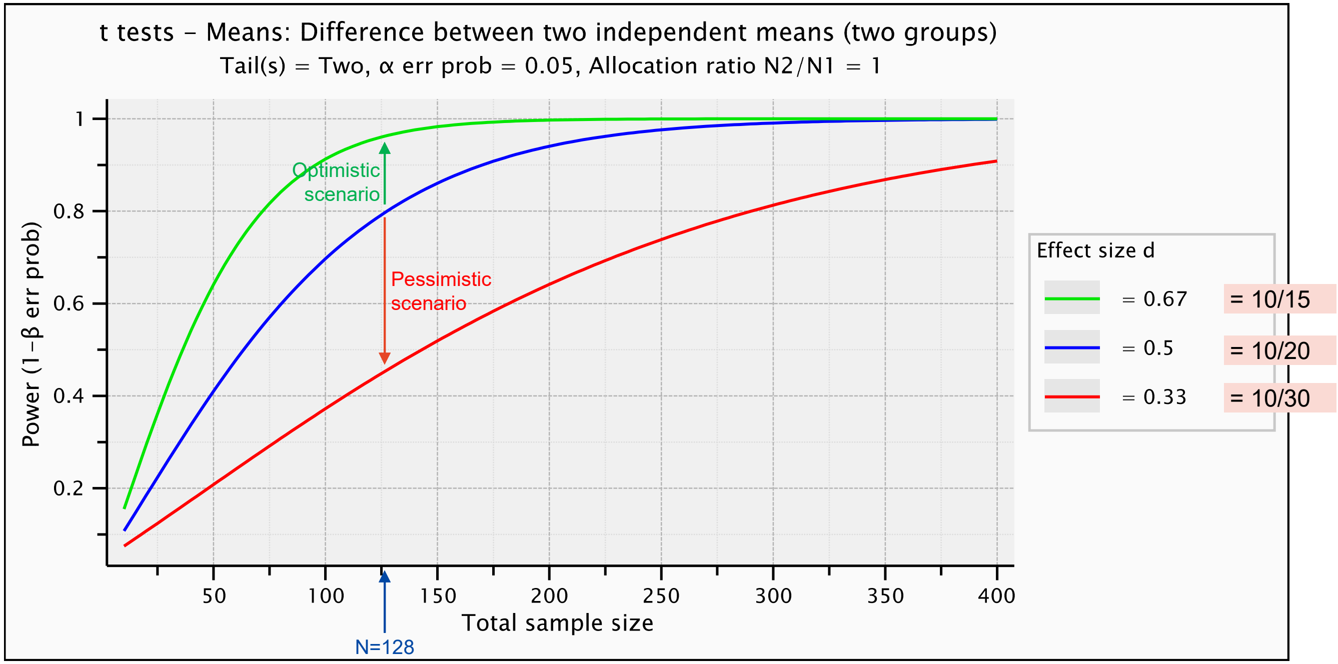
with Effect size d from 0.33 in steps of 0.17

and α err prob at 0.05

1. Difference between 2 means

Example: Chicken Welfare – Bone density

X-Y Plot: sample size vs power



1. Difference between 2 means

Example: Chicken Welfare – Bone density

Remember: the accepted meaning of d=0.5 is that this is a "medium" standardised effect size, so our value of d is roughly in the right ballpark for our planned study.

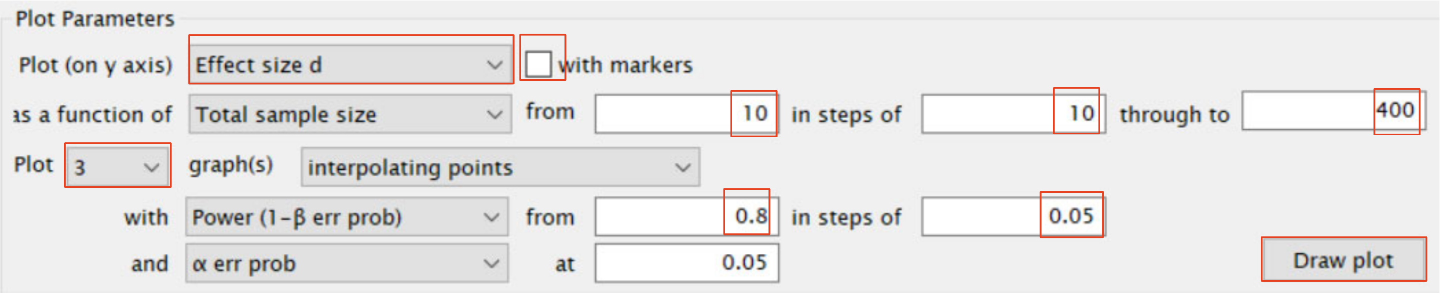
The sensitivity plot is another visualisation we can use in our power analysis. This plots effect size vs sample size.

1. Difference between 2 means

Example: Chicken Welfare – Bone density

Sensitivity Plot:

We want to look at a wide range of effect sizes. To do this, we will plot a sample size range from 10 up to 400 (as before) with 3 power curves for power = 0.8, 0.85, 0.90.



Plot Parameters

Plot (on y axis) Effect size d □ with markers

as a function of Total sample size from 10 in steps of 10 through to 400

Plot 3 graph(s) interpolating points

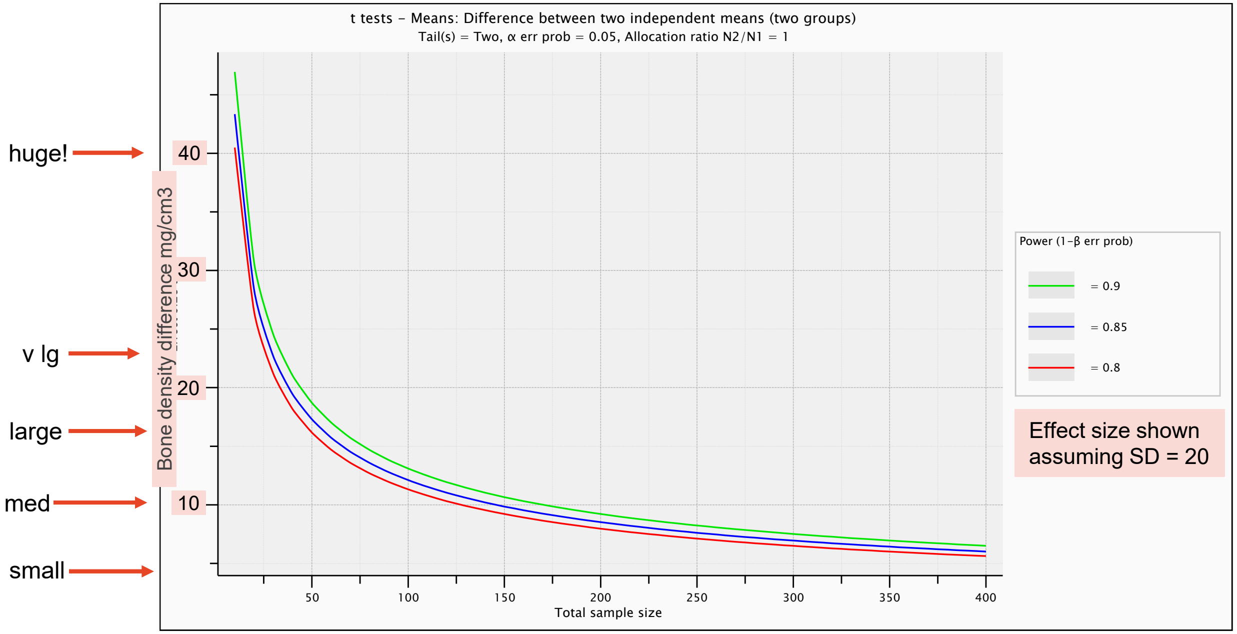
with Power (1−β err prob) from 0.8 in steps of 0.05

and α err prob at 0.05

1. Difference between 2 means

Example: Chicken Welfare – Bone density

X-Y Plot: sample size vs effect size (sensitivity)



Effect size shown assuming SD = 20

G*Power doesn’t’ provide axis format options, so you will have to do it manually if you want use your original outcome scale

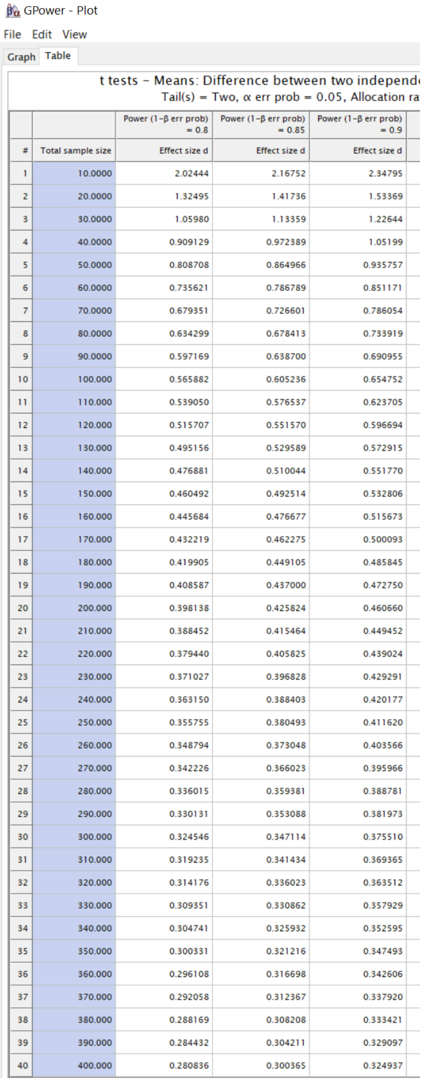
1. Difference between 2 means

Example: Chicken Welfare – Bone density

X-Y Plot: sample size vs effect size (sensitivity)

Customise plot in EXCEL

If you aren't happy with the G*Power plot, select the data from the Table tab and paste it into Excel (or your favourite plotting program).



t tests – Means: Difference between two independ
Tail(s) = Two, α err prob = 0.05, Allocation ra

		Power (1−β err prob) = 0.8	Power (1−β err prob) = 0.85	Power (1−β err prob) = 0.9
#	Total sample size	Effect size d	Effect size d	Effect size d
1	10.0000	2.02444	2.16752	2.34795
2	20.0000	1.32495	1.41736	1.53369
3	30.0000	1.05980	1.13359	1.22644
4	40.0000	0.909129	0.972389	1.05199
5	50.0000	0.808708	0.864966	0.935757
6	60.0000	0.735621	0.786789	0.851171
7	70.0000	0.679351	0.726601	0.786054
8	80.0000	0.634299	0.678413	0.733919
9	90.0000	0.597169	0.638700	0.690955
10	100.000	0.565882	0.605236	0.654752
11	110.000	0.539050	0.576537	0.623705
12	120.000	0.515707	0.551570	0.596694
13	130.000	0.495156	0.529589	0.572915
14	140.000	0.476881	0.510044	0.551770
15	150.000	0.460492	0.492514	0.532806
16	160.000	0.445684	0.476677	0.515673
17	170.000	0.432219	0.462275	0.500093
18	180.000	0.419905	0.449105	0.485845
19	190.000	0.408587	0.437000	0.472750
20	200.000	0.398138	0.425824	0.460660
21	210.000	0.388452	0.415464	0.449452
22	220.000	0.379440	0.405825	0.439024
23	230.000	0.371027	0.396828	0.429291
24	240.000	0.363150	0.388403	0.420177
25	250.000	0.355755	0.380493	0.411620
26	260.000	0.348794	0.373048	0.403566
27	270.000	0.342226	0.366023	0.395966
28	280.000	0.336015	0.359381	0.388781
29	290.000	0.330131	0.353088	0.381973
30	300.000	0.324546	0.347114	0.375510
31	310.000	0.319235	0.341434	0.369365
32	320.000	0.314176	0.336023	0.363512
33	330.000	0.309351	0.330862	0.357929
34	340.000	0.304741	0.325932	0.352595
35	350.000	0.300331	0.321216	0.347493
36	360.000	0.296108	0.316698	0.342606
37	370.000	0.292058	0.312367	0.337920
38	380.000	0.288169	0.308208	0.333421
39	390.000	0.284432	0.304211	0.329097
40	400.000	0.280836	0.300365	0.324937

2. Difference between 2 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, or the underlying distribution is not normal (could be categorical or continuous and highly skewed).

Often used for ordinal data from surveys.

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 <u>size</u> of the effect.

2. Difference between 2 means (Mann-Whitney)

Example: Happiness Survey

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.

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

2. Difference between 2 means (Mann-Whitney)

Example: Happiness Survey

You want to apply it to different groups of people (eg 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 1pt difference per item) is the smallest effect size of interest.

2. Difference between 2 means (Mann-Whitney)

Example: Happiness Survey

So, what are our first 4 steps?

Step 1:	Determine experiment type and statistical test	Mann-Whitney
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

2. Difference between 2 means (Mann-Whitney)

Example: Happiness Survey

Sample size calculation

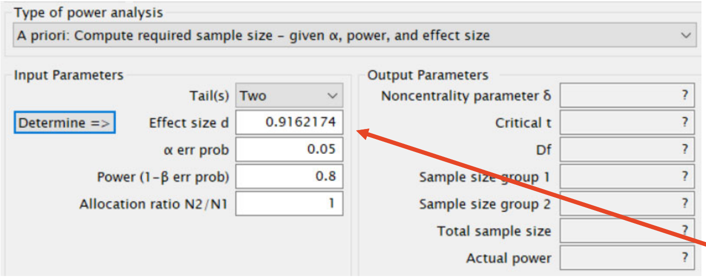
Heuristic method

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

2. Difference between 2 means (Mann-Whitney)

Example: Happiness Survey

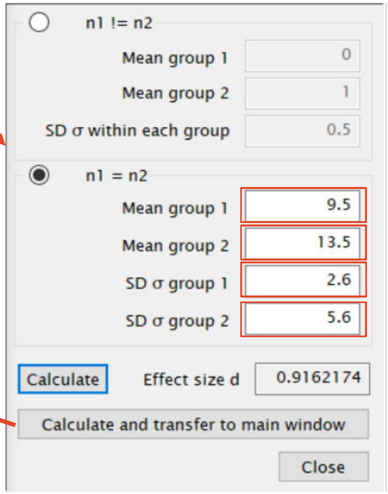
- Tests>Means>Two Independent Groups
- Click “Determine” (different SDs so use n1=n2)
- Enter expected means (use 9.5 for singles (group 1) and 13.5 for married (group 2) equates to 4pt diff)
- Enter SDs from pilot study (SD1=2.6, SD2=5.6)



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

Input Parameters
Tail(s): Two
Effect size d: 0.9162174
α err prob: 0.05
Power (1−β err prob): 0.8
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

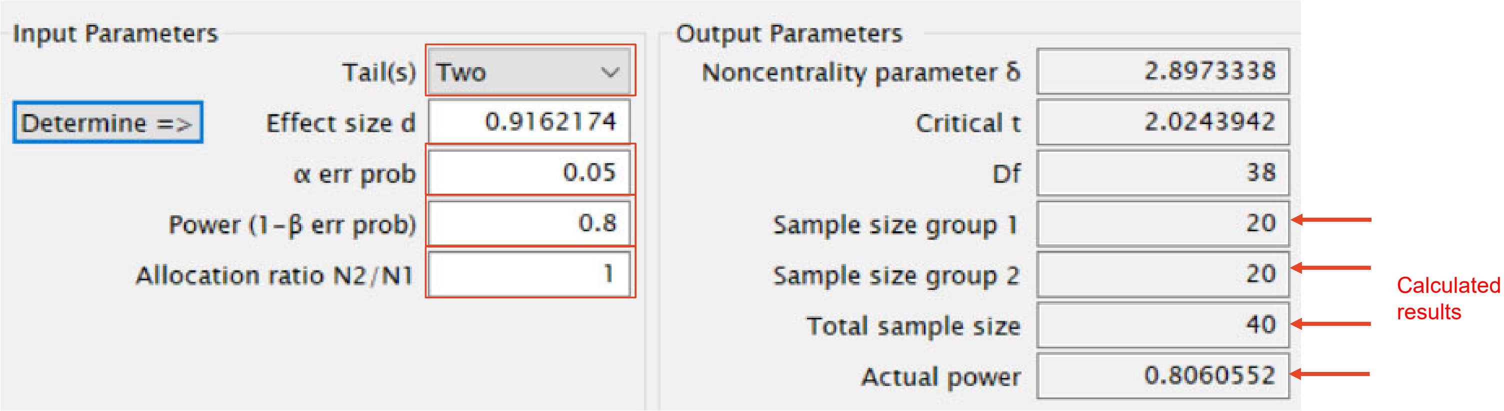
● n1 = n2
Mean group 1: 9.5
Mean group 2: 13.5
SD σ group 1: 2.6
SD σ group 2: 5.6

Effect size d: 0.9162174

2. Difference between 2 means (Mann-Whitney)

Example: Happiness Survey

- Check α, $1-\beta$, two tails, allocation ratio=1.
- Calculate sample size. N=20 per group
- Add 15% for non-parametric. N=20x1.15 = 23



Input Parameters

Tail(s)	Two
Effect size d	0.9162174
α err prob	0.05
Power (1-β err prob)	0.8
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	2.8973338
Critical t	2.0243942
Df	38
Sample size group 1	20
Sample size group 2	20
Total sample size	40
Actual power	0.8060552

Calculated results

2. Difference between 2 means (Mann-Whitney)

Theoretical approach
Statistical procedures can be compared according to their efficiency.

One test is more efficient than another if it requires fewer observations to obtain a given result.

The relative efficiency of two tests is the ratio of their efficiencies.

With smaller sample numbers, parametric tests are often more efficient than non-parametric tests although they approach equal efficiency with larger sample sizes.

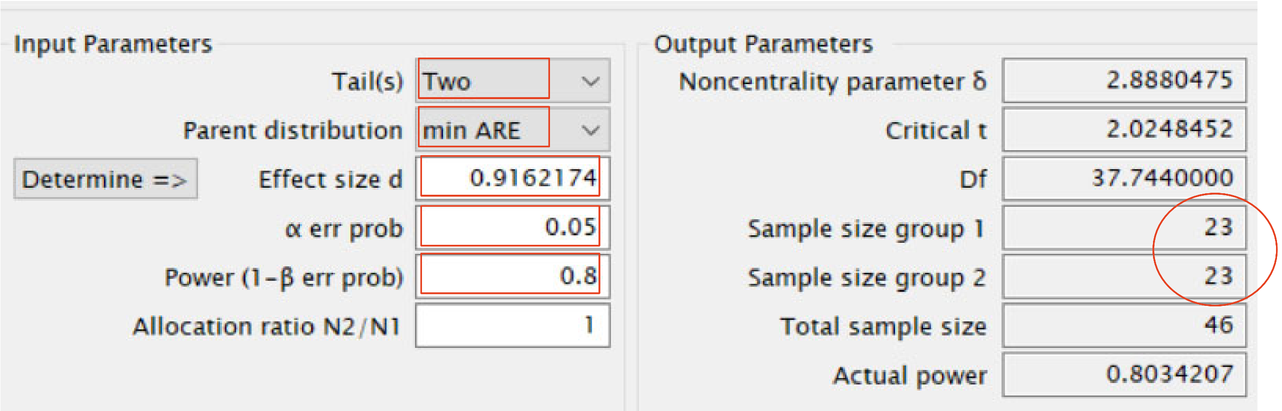
The Asymptotic Relative Efficiency (ARE) is the limit of the relative efficiencies as the sample size increases. It can be calculated or set and is used in the sample size calculation, along with the effect size.

It can be shown that the minimum ARE for these two tests is 0.864.

2. Difference between 2 means (Mann-Whitney)

Example: Happiness Survey

- Under “Tests” select “Means” and then the option:
- “Two independent groups: Wilcoxon (non-parametric)
- Use the same values as before:
- Two tails, α=0.05 and Power=0.80, group means and SDs.
- Select Parent distribution = “min ARE”
- Calculate sample size >> N=23 per group



Input Parameters

Tail(s)	Two
Parent distribution	min ARE
Effect size d	0.9162174
α err prob	0.05
Power (1-β err prob)	0.8
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	2.8880475
Critical t	2.0248452
Df	37.7440000
Sample size group 1	23
Sample size group 2	23
Total sample size	46
Actual power	0.8034207

3. Difference between 2 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?

- Set α=0.05 and $1-\beta$=0.8, two tails
- Allocation ratio N2/N1 = 1
- 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

3. Difference between 2 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	P1=0.55, P2=0.45

Note: The variance estimate comes from the proportion estimates.

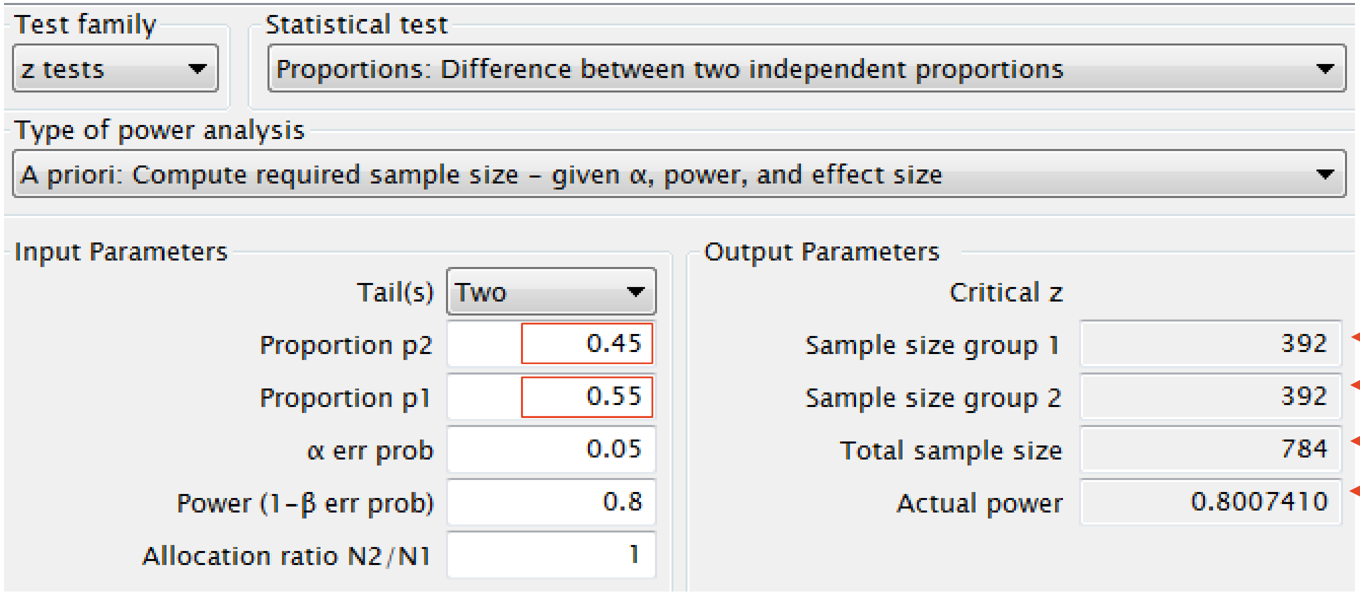
Variance = p(1-p). You do not need to calculate the variance just input the proportions into G*Power

3. Difference between 2 proportions

Example: Happiness survey

We need 392 subjects per group to achieve Power=0.80

That's a lot of happy/unhappy people!



Test family: z tests

Statistical test: Proportions: Difference between two independent proportions

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

Input Parameters

Tail(s)	Two
Proportion p2	0.45
Proportion p1	0.55
α err prob	0.05
Power (1-β err prob)	0.8
Allocation ratio N2/N1	1

Output Parameters

Critical z	
Sample size group 1	392
Sample size group 2	392
Total sample size	784
Actual power	0.8007410

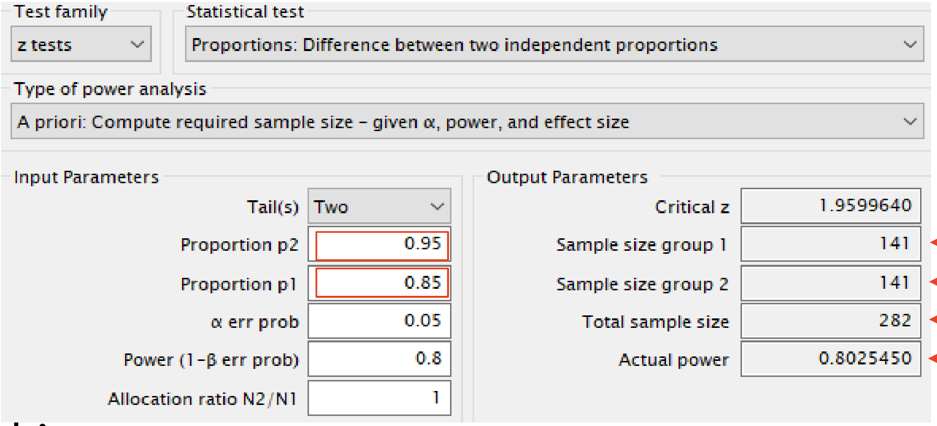
Calculated results

3. Difference between 2 proportions

Example: Happiness survey

Step 6: Suppose the proportion of subjects responding "happy" is expected to be higher, around 90%.

Try using P1=0.85 and P2=0.95



Test family: z tests

Statistical test: Proportions: Difference between two independent proportions

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

Input Parameters

Tail(s)	Two
Proportion p2	0.95
Proportion p1	0.85
α err prob	0.05
Power (1-β err prob)	0.8
Allocation ratio N2/N1	1

Output Parameters

Critical z	1.9599640
Sample size group 1	141
Sample size group 2	141
Total sample size	282
Actual power	0.8025450

Calculated results

Now we only need 141 subjects per group

Note the difference in sample sizes corresponding to the different proportion estimates. Remember the variance of the proportion parameter [var=p(1-p)] is at a maximum at 0.5 and gets smaller close to zero and one.

3. Difference between 2 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 $n_1 p_1 \leq 5 \; or \; n_2 p_2 \leq 5$)

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

3. Difference between 2 proportions

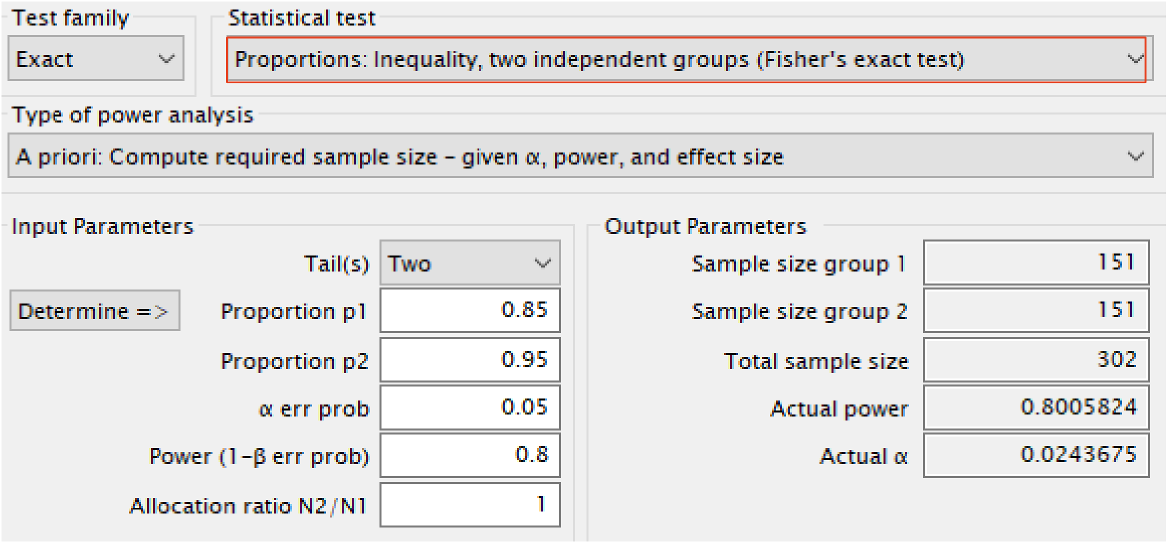
Example: Happiness survey

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

P1=0.85 and P2=0.95

Fisher’s Exact suggests 151 subjects per group.

Not quite the same result as the z-test, but note that the actual alpha is 0.024 rather than 0.05. This is a result of using an exact test rather than a [normal] approximation



Test family: Exact

Statistical test: Proportions: Inequality, two independent groups (Fisher's exact test)

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

Input Parameters

Tail(s)	Two
Proportion p1	0.85
Proportion p2	0.95
α err prob	0.05
Power (1-β err prob)	0.8
Allocation ratio N2/N1	1

Output Parameters

Sample size group 1	151
Sample size group 2	151
Total sample size	302
Actual power	0.8005824
Actual α	0.0243675

3. Difference between 2 proportions

Example: Happiness survey

Step 6: Explore scenarios

- When considering various scenarios, look for value estimates that provide a conservative power estimate.
- In this example proportions centred around 0.5 represent the most conservative estimate. This gives the largest sample size estimate.
- This principle may also be applied to the study design as well. For example powering your study for a non-parametric test is conservative (Mann-Whitney instead of t-test).

Power Analysis for other designs

G*Power scope

G*Power includes methods to calculate power and sample size for a wide variety of design scenarios, eg

- ANOVA
- Correlation
- Linear Regression
- Logistic Regression

Refer to the manual for details. Note that the manual cannot be reached from within G*Power itself. Instead 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

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

Note that this F test is rarely used to power your 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 significant p values

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

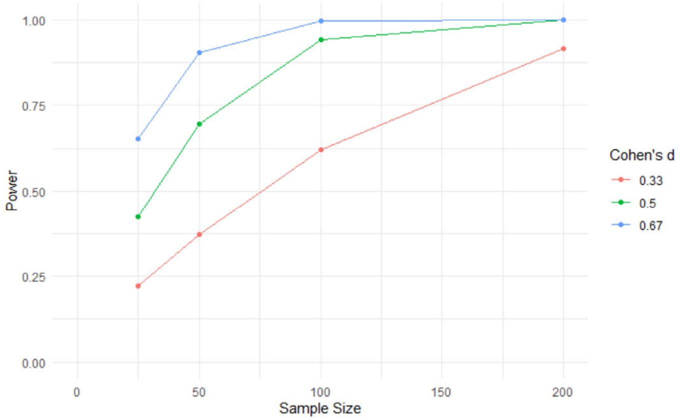
- The ‘trick’ is to set the parameters of the simulation in a sensible, realistic way

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

Power Analysis – by simulation

Example 1: Chicken Welfare - bone density (difference between 2 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

- Online calculators such as www.powerandsamplesize.com and https://sample-size.net/ and https://www.statulator.com/SampleSize/
- G*Power is a dedicated (free) program
- Make your own in Excel! (for example see Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4:863. doi:10.3389/fpsyg.2013.00863)

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

Challenge Questions

1. Power calculation

Your power calculation gives a sample size of n=20 per group, with α =0.05 and power = 0.8.

The Type I error rate is:

a. 80%

b. 20%

c. 5%

d. 2.5%

JM6

Challenge Questions

1. Power calculation

Your power calculation gives a sample size of n=20 per group, with α =0.05 and power = 0.8.

The Type I error rate is:

a. 80%

b. 20%

[✓ c. 5%]

d. 2.5%

Slide 81

JM6 Darya recommended the use of three types of challenge questions:
1. MCQ - multiple choice questions
2. Parson's problems - list of items out of order
3. Faded problem - each stage of problem contains fewer pieces of the puzzle.

Jim Matthews, 19/02/2020

Challenge Questions

2. Effect Size

a)The standardised effect size for group means can be expressed as the ratio of $\frac{?}{?}$ ______________________________

Challenge Questions

2. Effect Size

a)The standardised effect size for group means can be expressed as the ratio of

$$\frac{\textbf{difference in means}}{\textbf{within group std deviation}}$$

Challenge Questions

2. Effect Size

b)

If you halve the smallest effect size of interest in your power calculation…

a. you will need twice the sample size
b. you will need half the sample size
c. you will need 4 times the sample size
d. you will increase the Type I error

Challenge Questions

2. Effect Size

b) If you halve the smallest effect size of interest in your power calculation…

a. you will need twice the sample size

b. you will need half the sample size

c. [✓ you will need 4 times the sample size]

d. you will increase the Type I error

remember the inverse square relationship between n and d

$$n = 2\frac{\delta^2}{d^2}$$

Challenge Questions

3. Sample Size calculation

You are planning an experiment involving measuring the weight gain of 2 groups of foals (horses). One group is assigned Treatment A and the other Treatment B.

What hypothesis test would you use?

* _______________ or maybe * __________________

What pieces of information do you need to determine the sample size?

1	
2	
3	group balance (eg n1=n2)
4	
5	
6	one-tailed or two-tailed test

Challenge Questions

3. Sample Size calculation

You are planning an experiment involving measuring the weight gain of 2 groups of foals (horses). One group is assigned Treatment A and the other Treatment B.

What hypothesis test would you use?

* t-test or maybe * Mann Whitney (non-parametric)

What pieces of information do you need to determine the sample size?

1	alpha level (Type I error rate)
2	power level (1- Type II error rate)
3	group balance (eg n1=n2)
4	standard deviation (within group variance)
5	minimum weight increase (min. effect size of interest)
6	one-tailed or two-tailed test

Challenge Questions

4. Error Types

The types of error that may result from a hypothesis test are analogous with the errors that a jury might make when deciding on guilt or innocence of a defendant.

If the jury wrongly convicts, what type of error has occurred?

____

If the jury acquits the defendant, but she was actually guilty, what type of error has occurred?

____

"Beyond reasonable doubt" is a high standard of proof. It should result in a low ______________error rate

Challenge Questions

4. Error Types

The types of error that may result from a hypothesis test are analogous with the errors that a jury might make when deciding on guilt or innocence of a defendant.

If the jury wrongly convicts, what type of error has occurred?

________Type I__________

If the jury acquits the defendant, but she was actually guilty, what type of error has occurred?

________Type II__________

“Beyond reasonable doubt” is a high standard of proof. It should result in a low ______Type I______error rate

Questions?

Your turn…

Power calculation references

- **G*Power** http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html
- **NCSS PASS Statistical software** https://www.ncss.com/software/pass/
- **Causal Evaluation** https://www.causalevaluation.org/power-analysis.html
- **Epi Tools for disease prevalence (by AUSVET)** http://epitools.ausvet.com.au/content.php?page=SampleSize
- **Demidenko (Dartmouth) for logistic regression** https://www.dartmouth.edu/~eugened/power-samplesize.php
- **National Institutes of Health (NIH – USA) for cluster randomised trials** https://researchmethodsresources.nih.gov/SampleSizeCalculator.aspx
- **UCSF Clinical and Translational science institute** (Survival for clinical research) http://www.sample-size.net/sample-size-survival-analysis/
- **Lakens, D. Open Science Framework** https://osf.io/ixGcd/

Power Analysis – library references

- **Cohen, Jacob. Statistical Power Analysis for the Behavioral Sciences.**
  Burlington: Elsevier Science, 2013. Print.
  https://sydney.primo.exlibrisgroup.com/permalink/61USYD_INST/14vvljs/alma991005702359705106

- **Dattalo, Patrick. Determining Sample Size Balancing Power, Precision, and Practicality**
  Oxford: Oxford University Press, 2008. Print.
  https://sydney.primo.exlibrisgroup.com/permalink/61USYD_INST/14vvljs/alma991015395569705106

- **Julious, Steven A. Sample Sizes for Clinical Trials**
  Boca Raton: CRC Press/Taylor & Francis, 2010. Print.
  https://sydney.primo.exlibrisgroup.com/permalink/61USYD_INST/14vvljs/alma991000960739705106

- **Ryan, Thomas P., and Thomas P Ryan. Sample Size Determination and Power.**
  Somerset: John Wiley & Sons, Incorporated, 2013. Web.
  https://sydney.primo.exlibrisgroup.com/permalink/61USYD_INST/1367smt/cdi_askewsholts_vlebooks_9781118439203

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.”

We value your feedback

We want to hear about you and whether this workshop has helped you in your research. What **worked** and what **didn't work.**

We actively use the feedback to improve our workshops.

Completing this survey really does help us and we would appreciate your help! It only takes a few minutes to complete (*promise!*)

You will receive a link to the anonymous survey by email