Examples using G*Power software

We will work through 3 simple examples

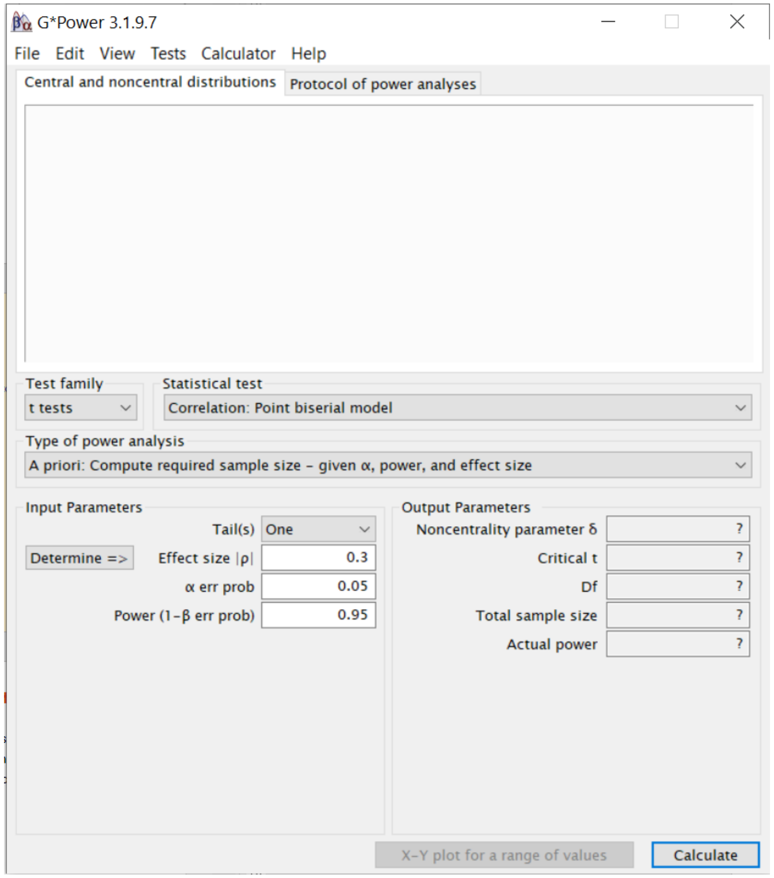
1. Difference between 2 means (continuous response)
2. Difference between 2 means (survey response)
3. Difference between 2 proportions

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



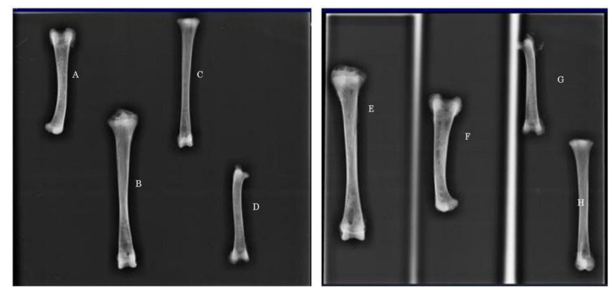
1. Difference between 2 means

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

Treatment Group = high mineral diet
Control Group = 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)

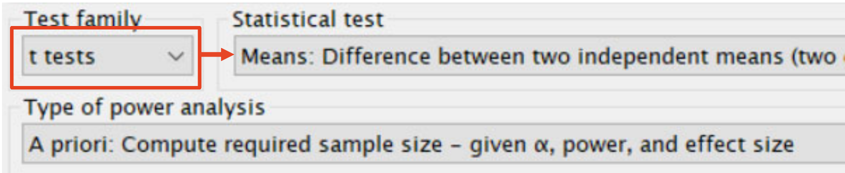
d = standardised effect size (Cohen's d)

1. Difference between 2 means

Step 5: G*Power

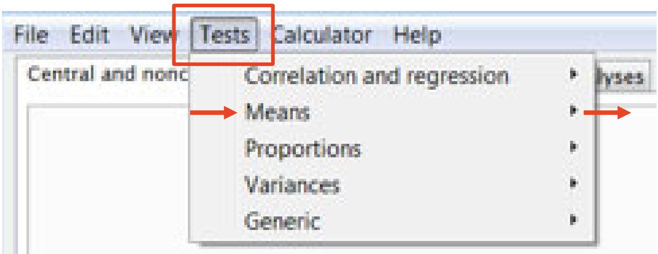
There are two ways to find the correct test

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



– Design based approach: Select the test parameter class (eg means), then the study design

– Select Tests/Means/Two independent groups



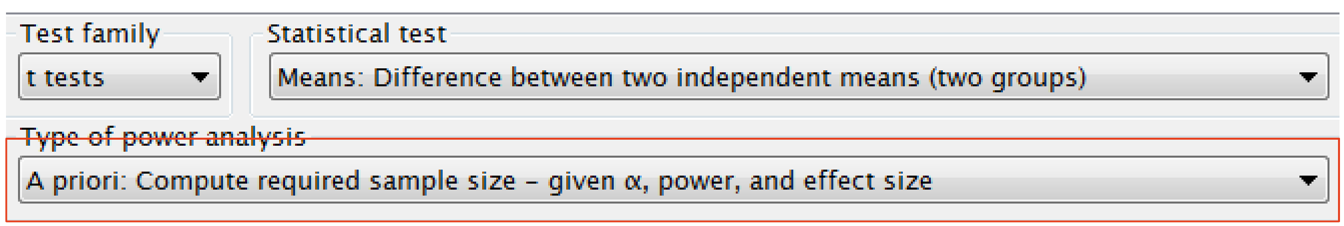
1. Difference between 2 means

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

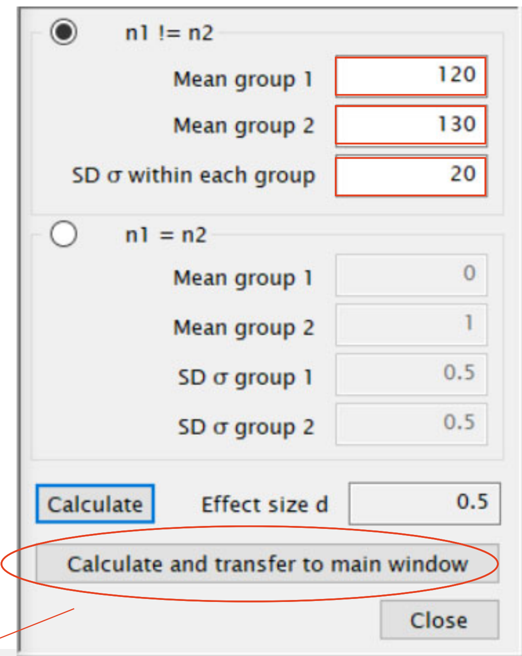


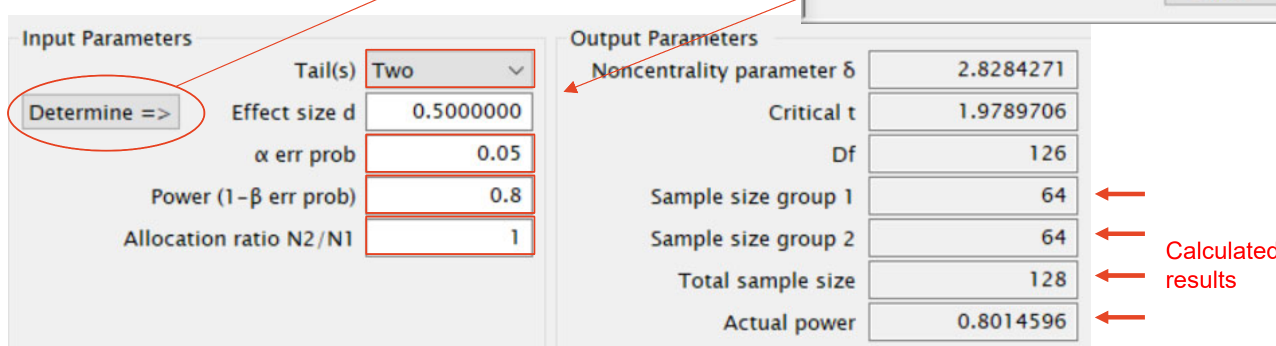
1. Difference between 2 means

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)





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.

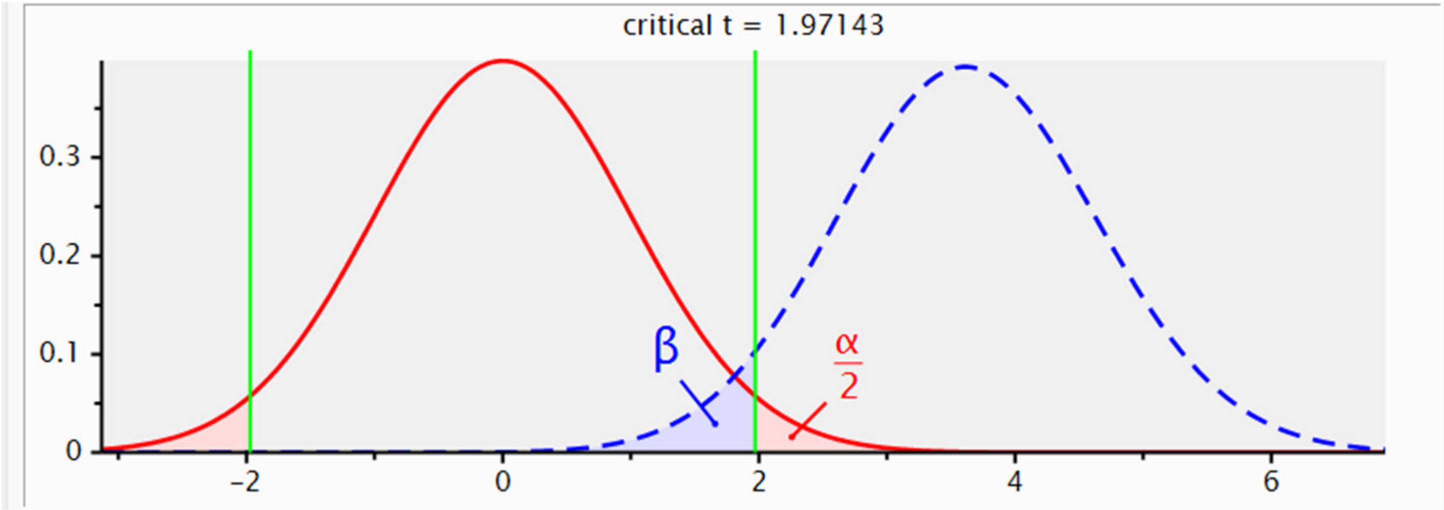
Protocol of the power analysis

- 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.

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

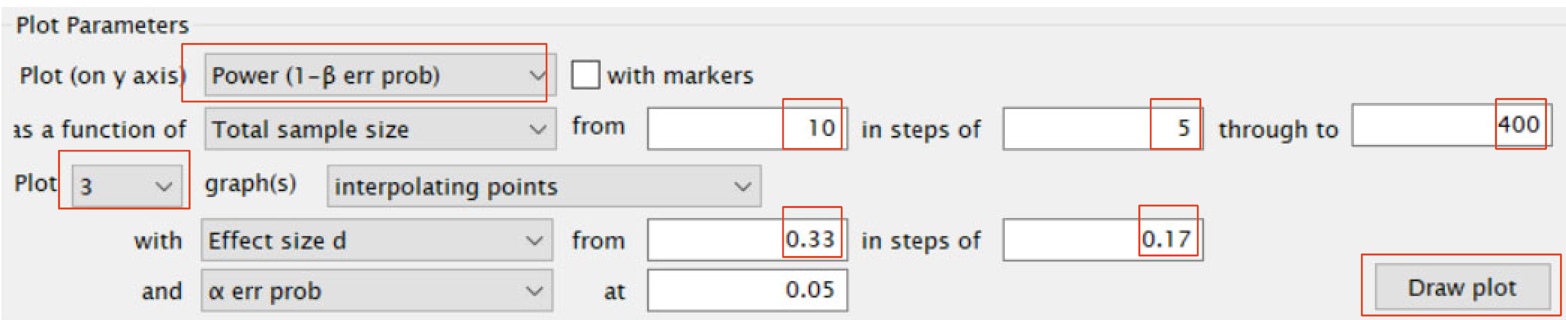
Example: Chicken Welfare – Bone density

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

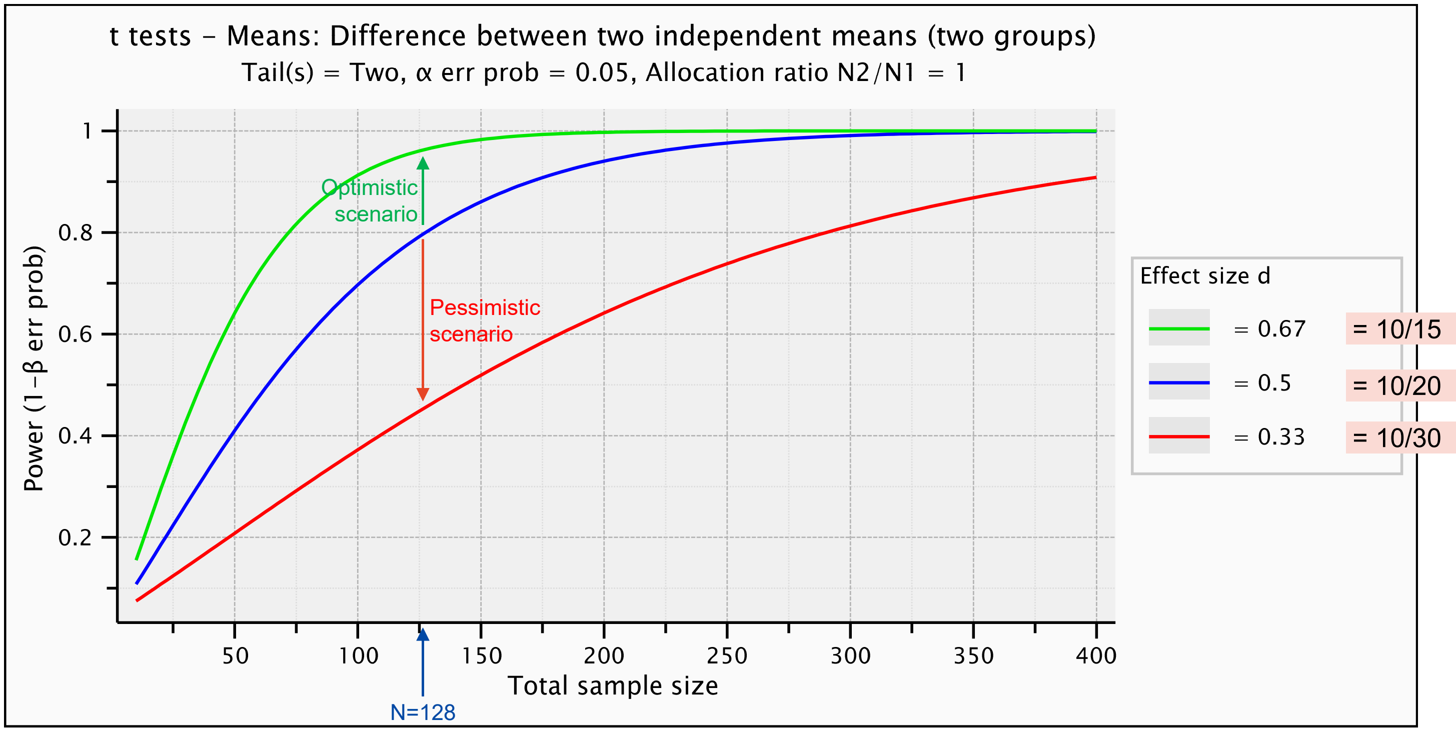
– Plot “3” graphs with d = 0.33 in steps of 0.17 (gets us to 0.5 and 0.67)



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.

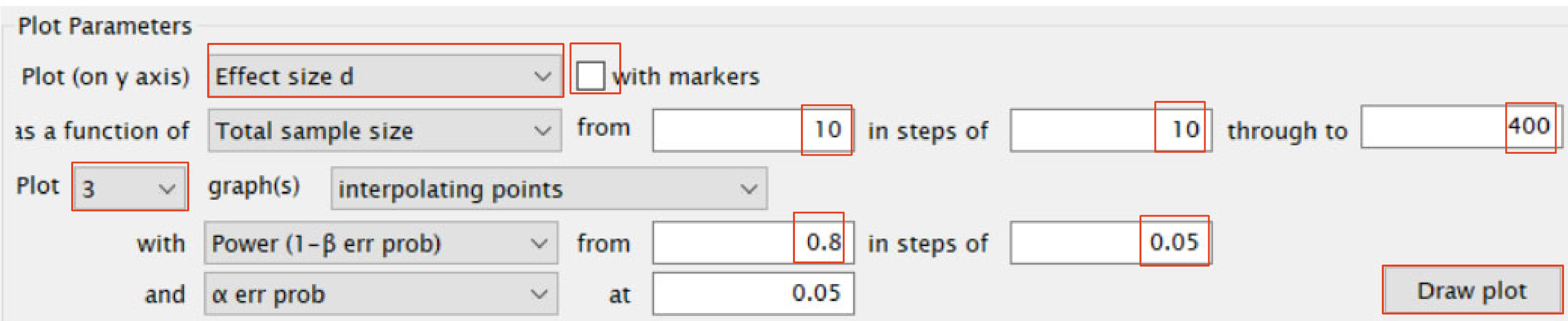
The sensitivity plot is another visualisation we can use in our power analysis. This plots effect size vs sample size.

1. Difference between 2 means

Example: Chicken Welfare – Bone density

Sensitivity Plot:

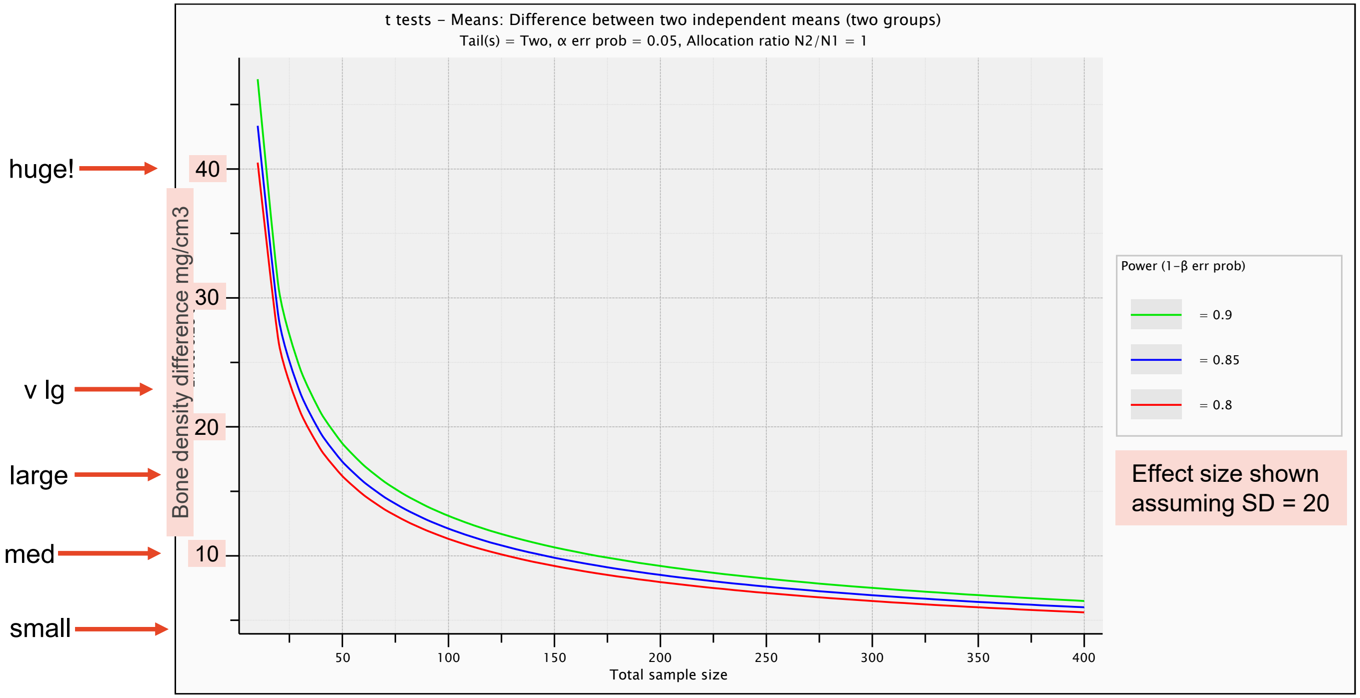
We want to look at a wide range of effect sizes. To do this, we will plot a sample size range from 10 up to 400 (as before) with 3 power curves for power = 0.8, 0.85, 0.90.



1. Difference between 2 means

Example: Chicken Welfare – Bone density

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



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)

<u>Customise plot in EXCEL</u>

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).

GPower - Plot

File Edit View

Graph Table

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 size 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

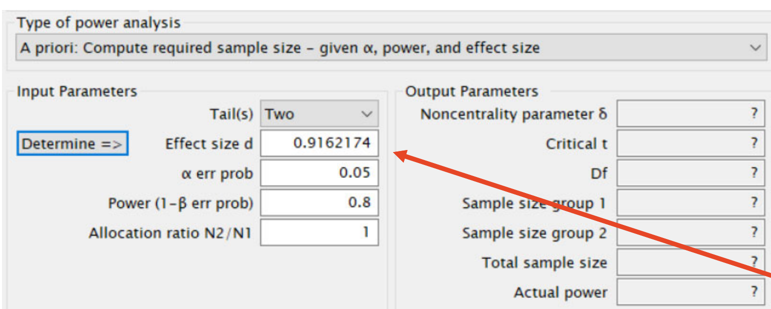
Heuristic method

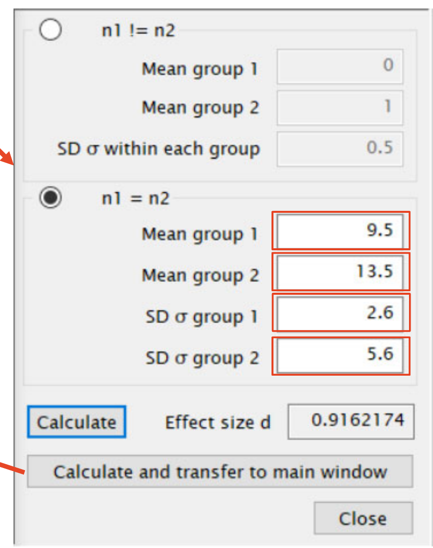
“Do the calculations as if performing the corresponding parametric test (i.e. the t-test), then add 15% to the sample size.

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)

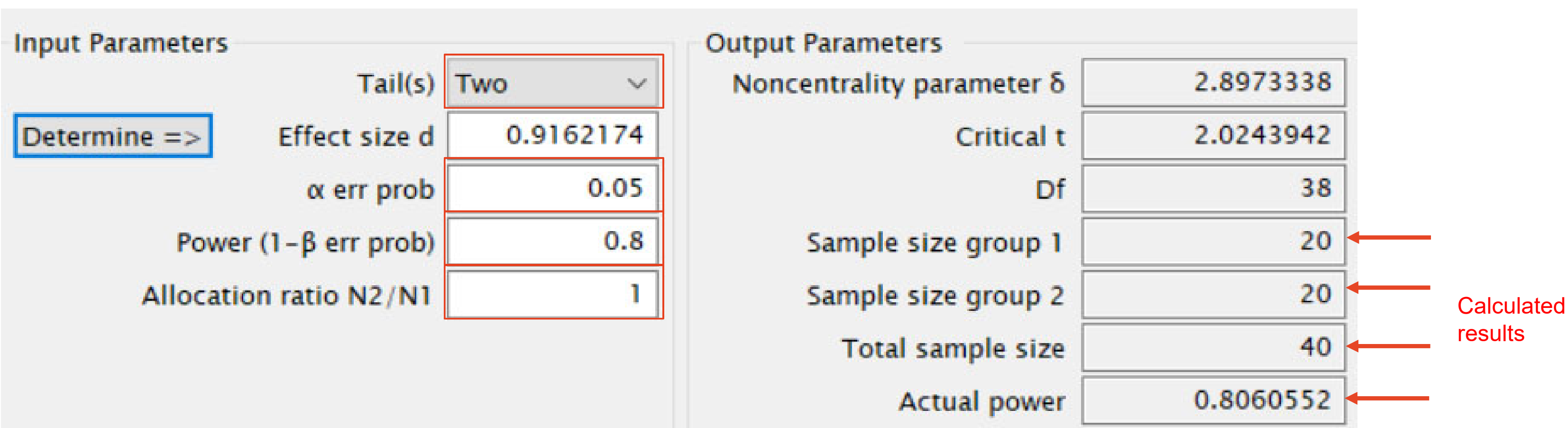




2. Difference between 2 means (Mann-Whitney)

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



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.

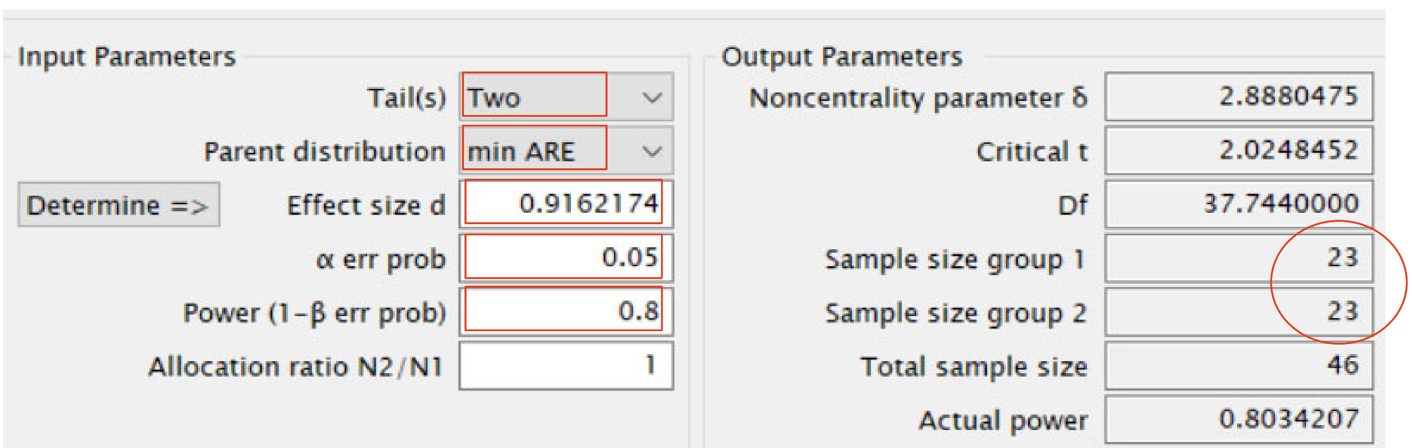
The Asymptotic Relative Efficiency (ARE) is the limit of the relative efficiencies as the sample size increases. It can be calculated or set and is used in the sample size calculation, along with the effect size.

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



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.

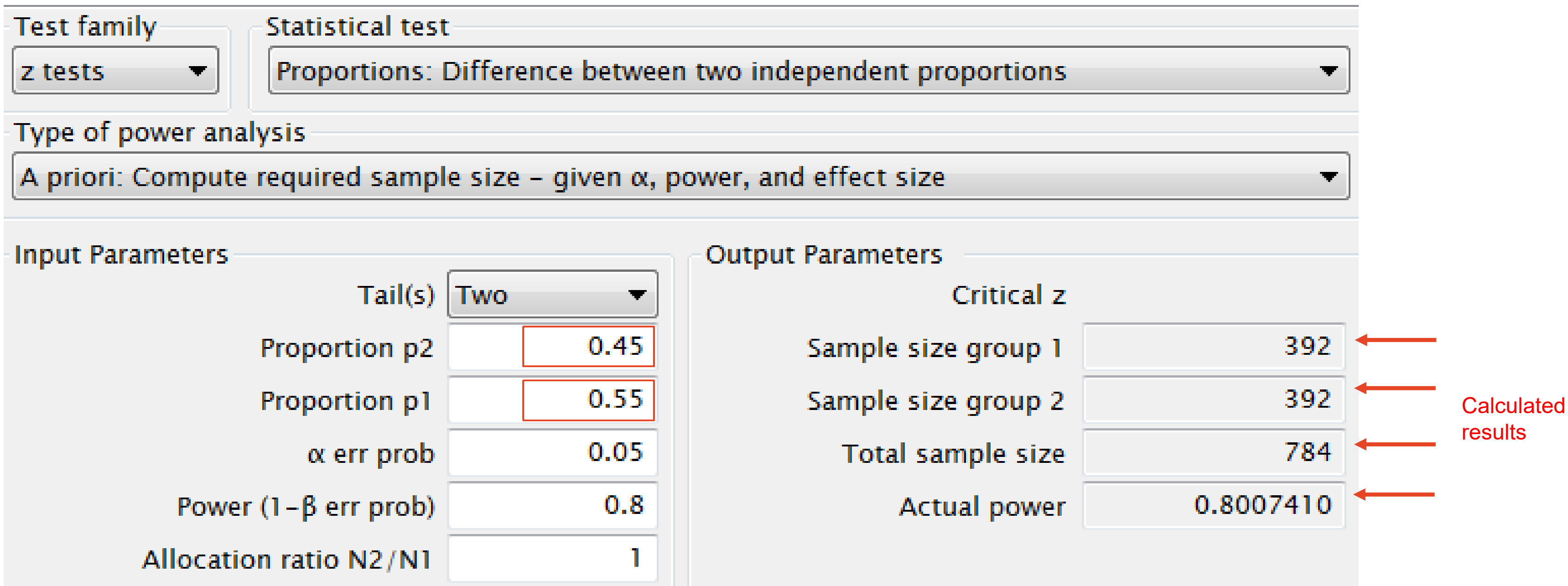
Variance = p(1-p). You do not need to calculate the variance just input the proportions into G*Power

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!

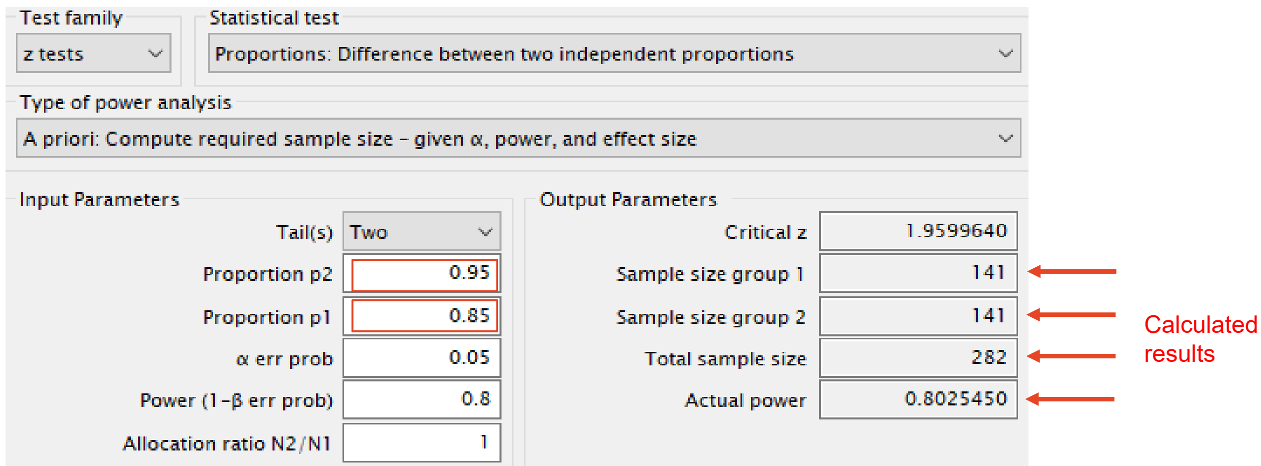


3. Difference between 2 proportions

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



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 $n_1p_1 \leq 5 \; or \; n_2p_2 \leq 5$)

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

3. Difference between 2 proportions

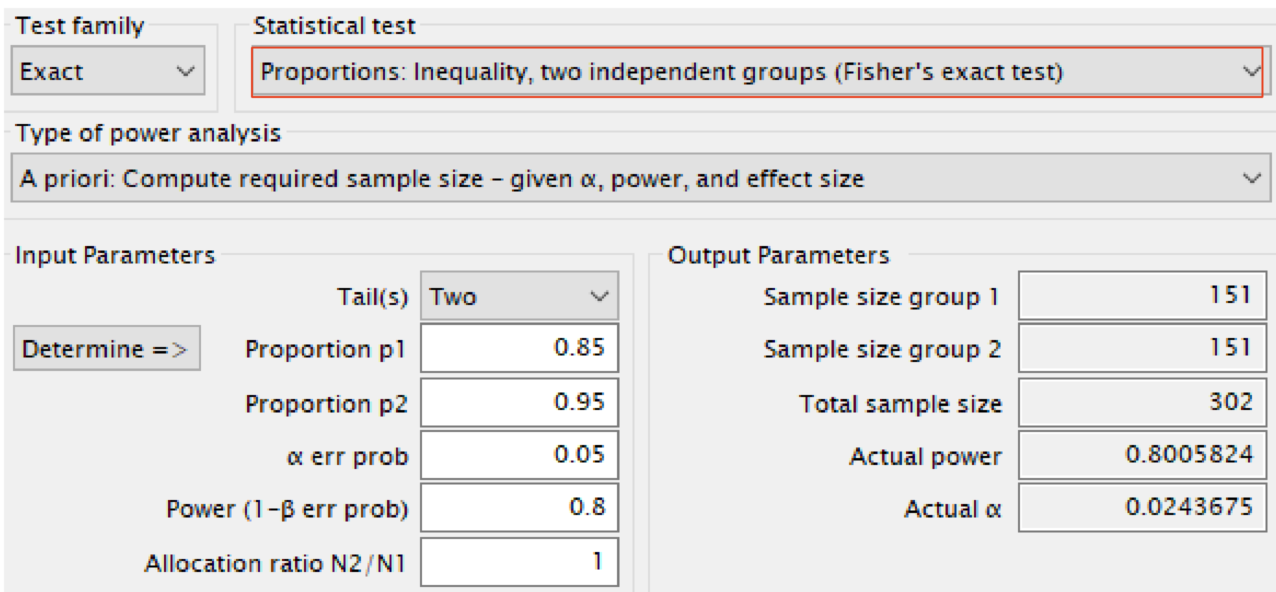
Example: Happiness survey

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



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).