THE UNIVERSITY OF NEWCASTLE AUSTRALIA

Statistical Support Service

Statistics refresher seminar series

How much data should I collect?

13-Jun-2019

Kim Colyvas

statsupport@newcastle.edu.au

Our website, search for StatSS on university web site

How To Analyse My Data

3- 5 July 2019

Outlines

- Exploratory data analysis and visualising data
- Formulating research questions
- Data types and related statistical tests
- How to interpret statistical results

♦ **Explanation of common statistical tests**

♦ **Workbook with worked examples then hands on practice**

♦ **Use statistical software to create output (SPSS)**

♦ **SPSS software guide provided**

♦ **Focus on understanding, concepts and interpretation of results**

Instructors
Nic Croce, Fran Baker

Statistical Support Service

Notes for all seminars can be downloaded from the Courses, Seminars and Workshops section at
http://www.newcastle.edu.au/about-uon/governance-and-leadership/faculties-and-schools/faculty-of-science-and-information-technology/resources/statistical-support-services

Easier however is to type **StatSS**
into the university web site's search box.
Our site is the first result in the list – choose the heading
Courses, seminars and workshops heading.

Intent of this session

- What information is needed to determine sample size for a study.
- How this information is used.
- Interpreting the results of a sample size determination.
- Understanding effect sizes.
- Not how to do sample size and power calculation (but will get some idea).

Variable types

- **Numeric** - values that “mean numbers”
  - **Continuous**: temperature, weight,, speed, distance
  - **Discrete**: #defects, result of die toss, product count
- **Categorical** – values based on categories
  - **Nominal**
    gender – male/female   colour - blue/green/yellow
  - **Ordinal**
    Grades - FF, P, C, D, HD,
    Temperature - Low, Medium, High

Response	Explanatory	Specific question(s)	Displays	Statistical method
Categorical	Categorical	How do proportions in response depend on **the levels of the explanatory variable?**	Tables	Chi-squared statistic
Categorical	Numeric (Continuous)	How does the proportion in response depend on the explanatory variable?	Tables (X groups)	*Logistic regression* *Correlation (for a binary response only)*
Numeric (Continuous)	Categorical	How does mean level in response change with **the levels of the explanatory variable? If so how does it vary?**	Box plots Mean plots CI plots	t test (2 groups) ANOVA (3 or more groups)
Numeric (Continuous)	Numeric (Continuous)	How does mean level of response change with the explanatory variable	Scatter plots	Correlation Regression
Categorical				Kappa (2x2 - Agreement) McNemar's test (2x2 - bias in agreement)
Numeric (Continuous)	Categorical	How does mean level in response change with the levels of the explanatory variable WITHIN e.g. subject	Box plots Mean plots Within CI plots	Paired t test Repeated measures ANOVA

Todays focus will be on the 2 research questions in red

Differences between 2 groups

Purpose: **Test differences between 2** treatments, genders etc

- Outcome is categorical

  Increase awareness of service following training intervention from 35% to 54%.

- Outcome is numeric

  Improvement in pain index after treatment was 17.3.

Statistical significance & Practical significance

- **Statistical significance**
  A statistical test is carried out and we find the difference is significant based say on a p value.
- **Practical significance**
  Whether the difference is meaningful within our field of study.
- It is easy with large sample sizes to obtain statistically significant differences that are not meaningful.
- This session is concerned with designing studies to find **practically significant** effects, i.e. important clinically, biologically, environmentally, socially etc.

My Study – most important variables

1) **Response** Type:
2) **Explanatory** Type:
   Numbers of levels for each if categorical
3) **Dependent/Independent:**
4) **Practical significance**
   What size change is important?
   - Previous experience or research
   - Don't know, use Cohen's effect size (see later)
5) **How large is the variability?**
   Prior information, prior research, guess,
   Don't know, use Cohen's effect size (see later)

Break in lecture
for ~10 mins
for class exercise

Sample size estimation

- We will not be using formulae.
- Free software is available.
- The demonstrations in this talk use the **Power and Sample Size** program.
- See first reference on last slide for link to download.
- Also G*Power 3 is an alternative free program.

Sample Size and Power analysis

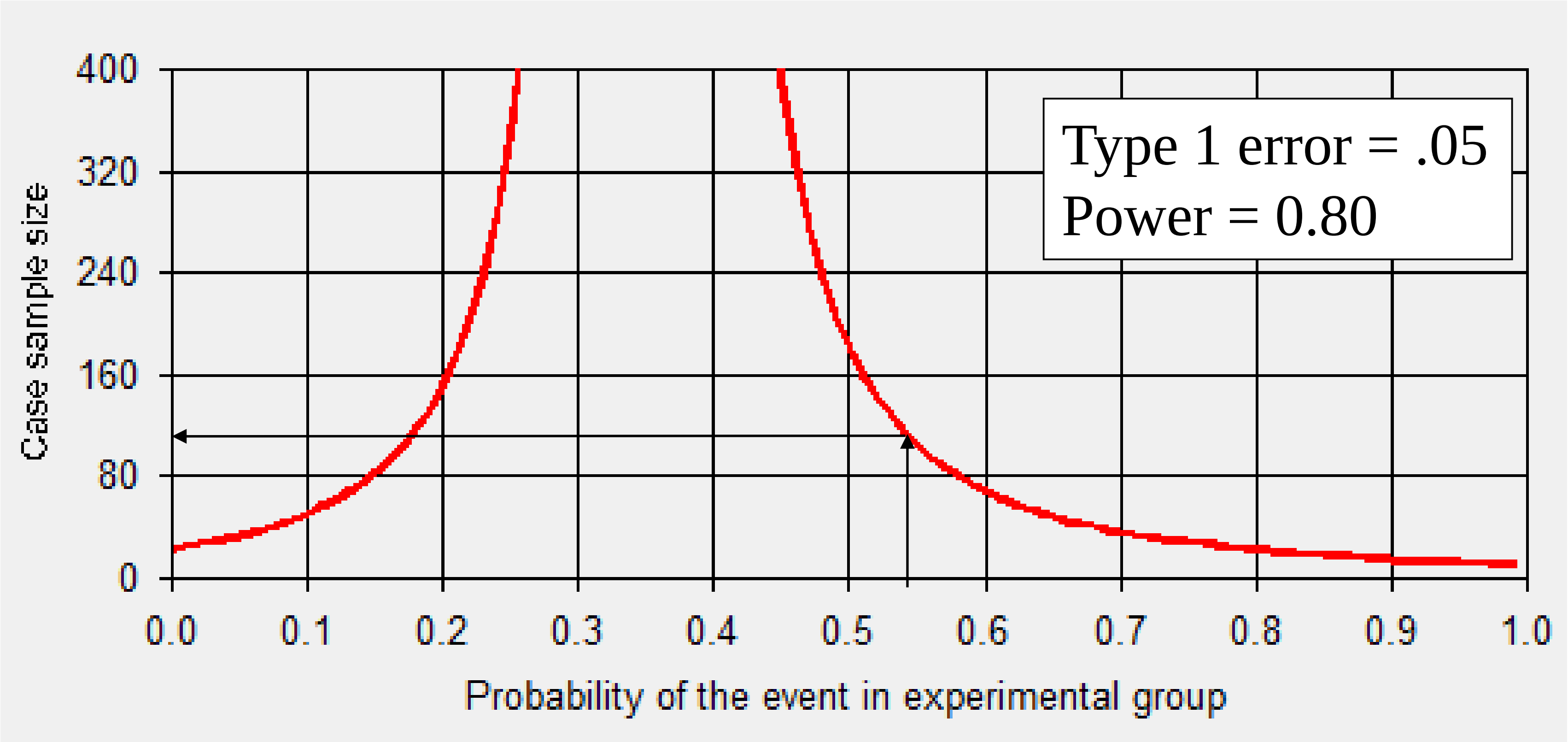
- Key idea is to know before a study what is the **chance** you will discover something.
- Sample size is a key driver of this.
- Can save wasted effort and disappointment if proper planning is carried out prior to the study.

Statistical Concepts for sample size

- **Type I Error** – alpha (typically $\alpha = 0.05$)
  Chance that an effect will be declared as real when in reality there is no effect.
- **Type II Error** – beta (typically $\beta=0.20$)
  Chance a real effect **WILL NOT** be detected.
- **Power (1-β) (typically 0.8)**
  Chance a real effect WILL be detected.
  Power of 0.8 means we have an
  8 in 10 chance of detecting the effect.
  2 in 10 chance of NOT detecting the effect.

Categorical outcome – difference of 2 proportions

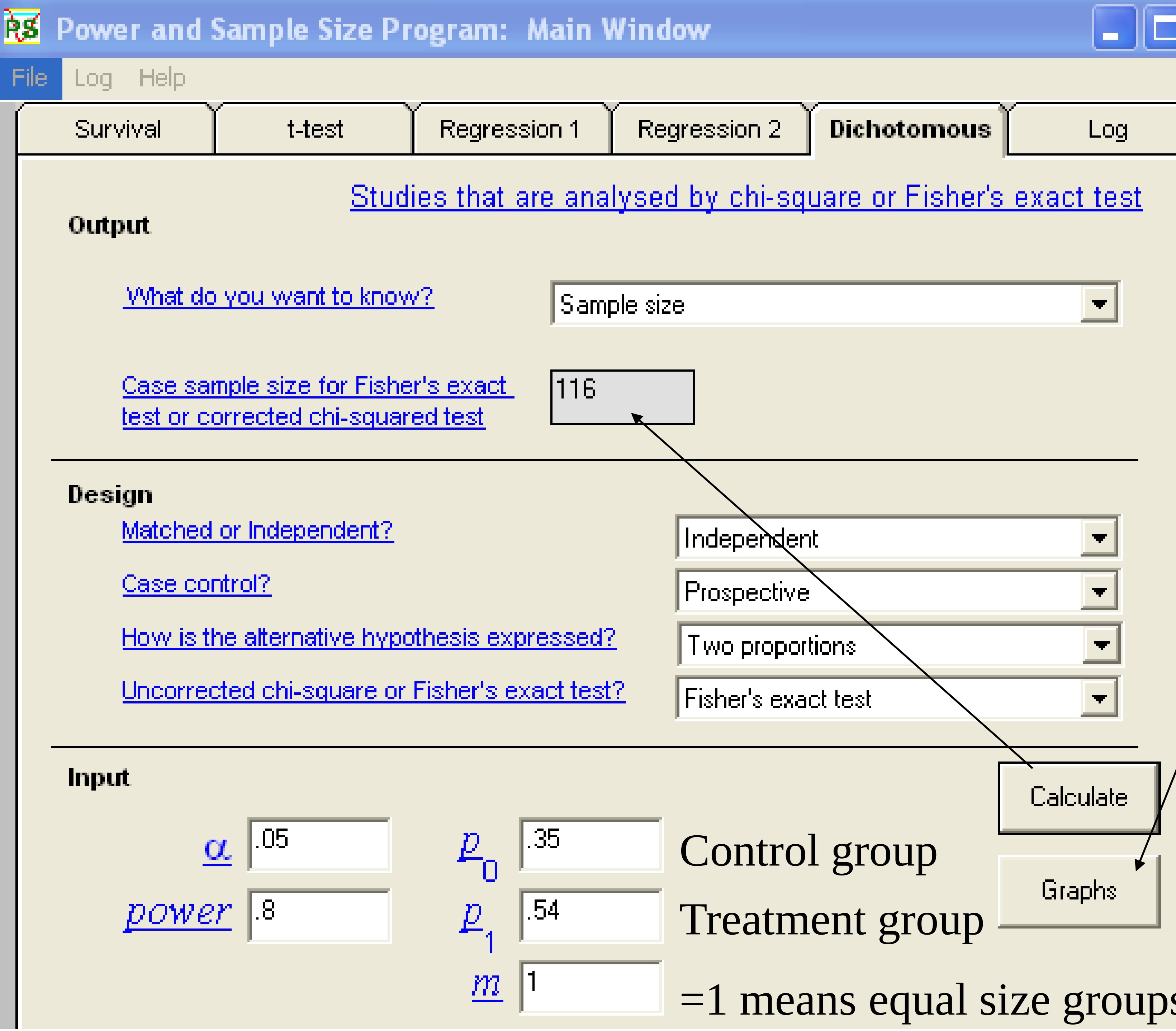
- Difference between 2 groups, propose that Control Group = 35%, Treatment group =54%



Sample size (n) required in each group ~ 115, ie total N = 230

What is n for 35% vs70%?

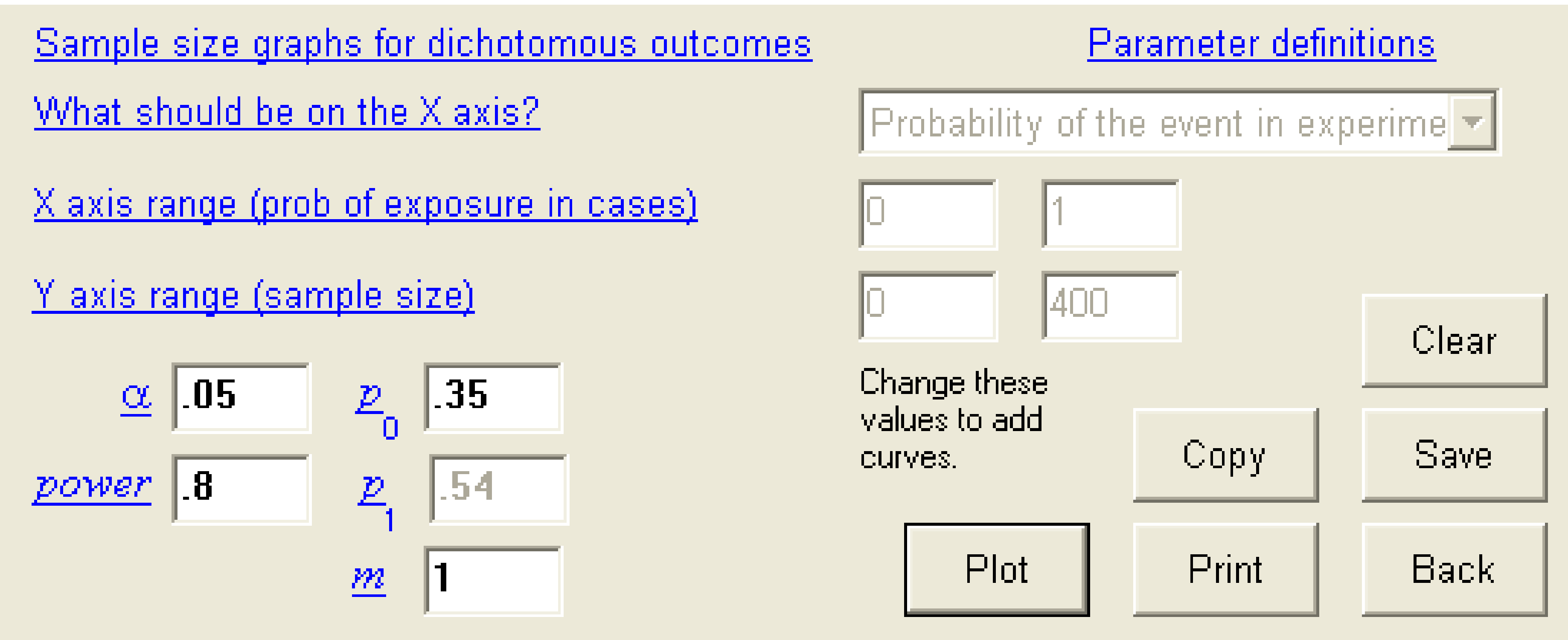
What is n for 35% vs 18%?



<u>Program setup</u>
First screen for previous slide then click Graphs button

Program setup - second screen

This is used to create the graphical output two slides before

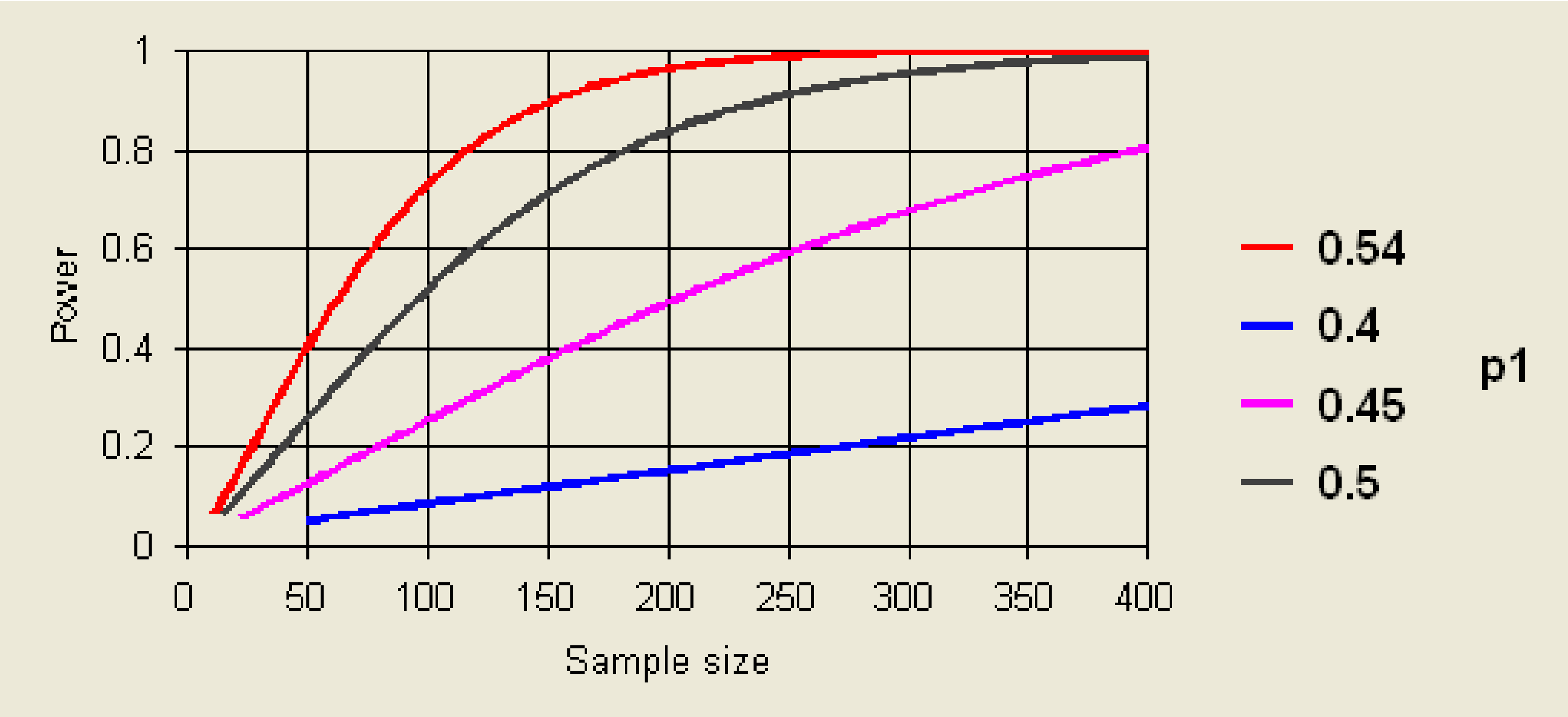


Sensitivity analysis

- Very useful not to do just a single sample size calculation.
- Try a range of options to get a feel for how they might impact the effectiveness of your planned study.
- How would the power of your study be affected by different sample sizes?
  For example loss to follow-up.
- How would sample size change for other sizes of practical significance?

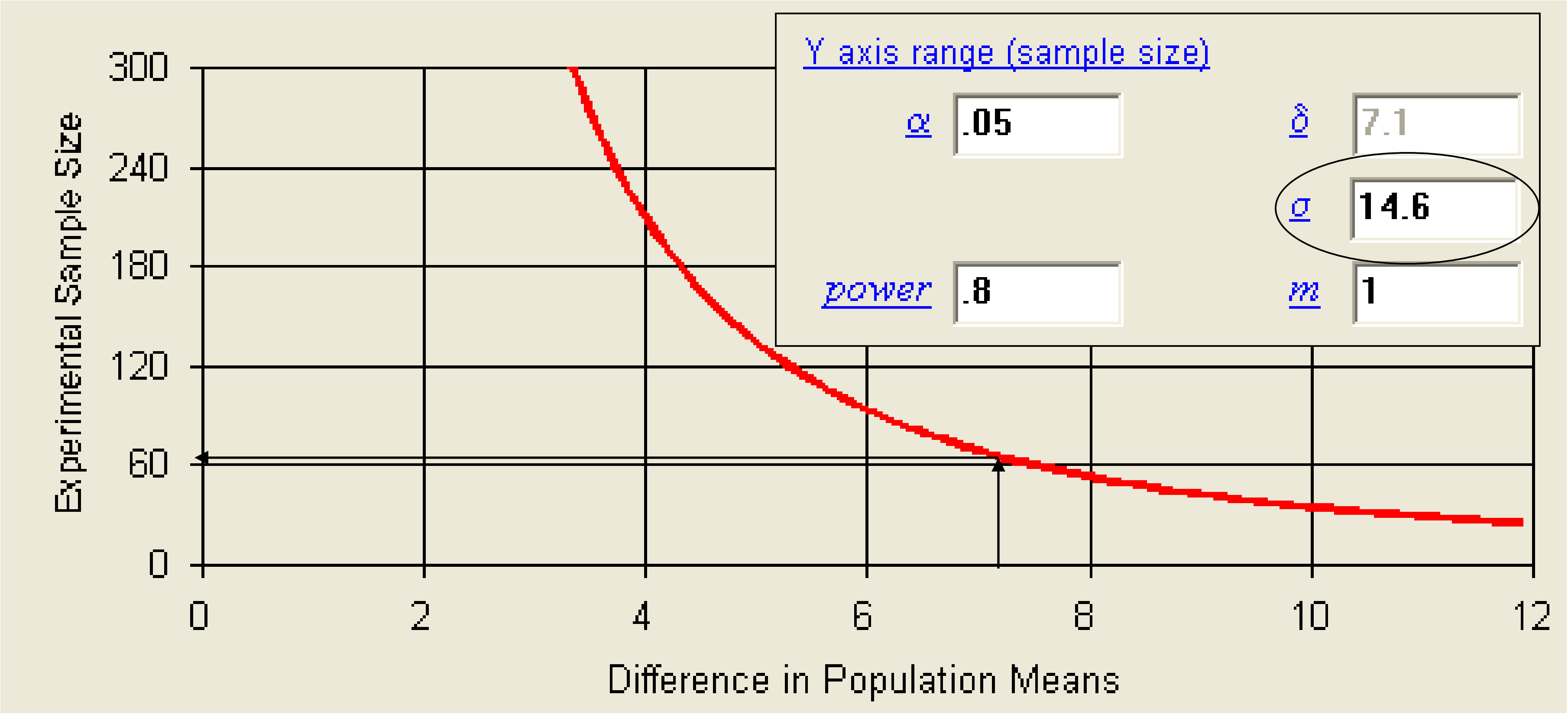
Varying sample size

- As the size of the **difference** changes the effect on the power of the test is shown below.
- Or for a constant sample size the power changes



Numeric outcome difference between 2 means

- Mean difference between 2 independent groups 103.6 Low group, 96.4 in High group, Diff = 7.1



Sample size (n) required in each group ~ 65 , ie total N = 130

What is n for difference = 4?

Effect on sample size if variability (σ) was larger?

Effect size (ratio of signal/noise)

- See Cohen references

Change in means.
What is the **practical/clinical significance** of this?

$$d = \frac{(mean_1 - mean_2)}{SD}$$

Signal / Noise

Variability within each group controls (or limits) the ability to detect a difference.

Sample sizes for means

For a single (1) sample compared to a reference value.
For two (2) samples, between two groups for example.
(Howell 2002)

Total Sample size (N)

Number of samples		**1**	**2**
Small effect	d = 0.2	196	784
Medium effect	d = 0.5	32	126
Large effect	d = 0.8	13	49

$\alpha = 0.05$, power = 0.8

Other effect size measures

Table 1
ES Indexes and Their Values for Small, Medium, and Large Effects

Test	ES index	Effect size: Small	Medium	Large
1. m_A vs. m_B for independent means	$d = \frac{m_A - m_B}{\sigma}$	.20	.50	.80
2. Significance of product–moment r	r	.10	.30	.50
3. r_A vs. r_B for independent rs	$q = z_A - z_B$ where z = Fisher's z	.10	.30	.50
4. P = .5 and the sign test	$g = P - .50$	.05	.15	.25
5. P_A vs. P_B for independent proportions	$h = \phi_A - \phi_B$ where ϕ = arcsine transformation	.20	.50	.80
6. Chi-square for goodness of fit and contingency	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	.10	.30	.50
7. One-way analysis of variance	$f = \frac{\sigma_m}{\sigma}$	.10	.25	.40
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

Note. ES = population effect size.

From Cohen, J., A power primer, Psychological Bulletin, 1992, 112(1), 155-159

Numeric variables for best power

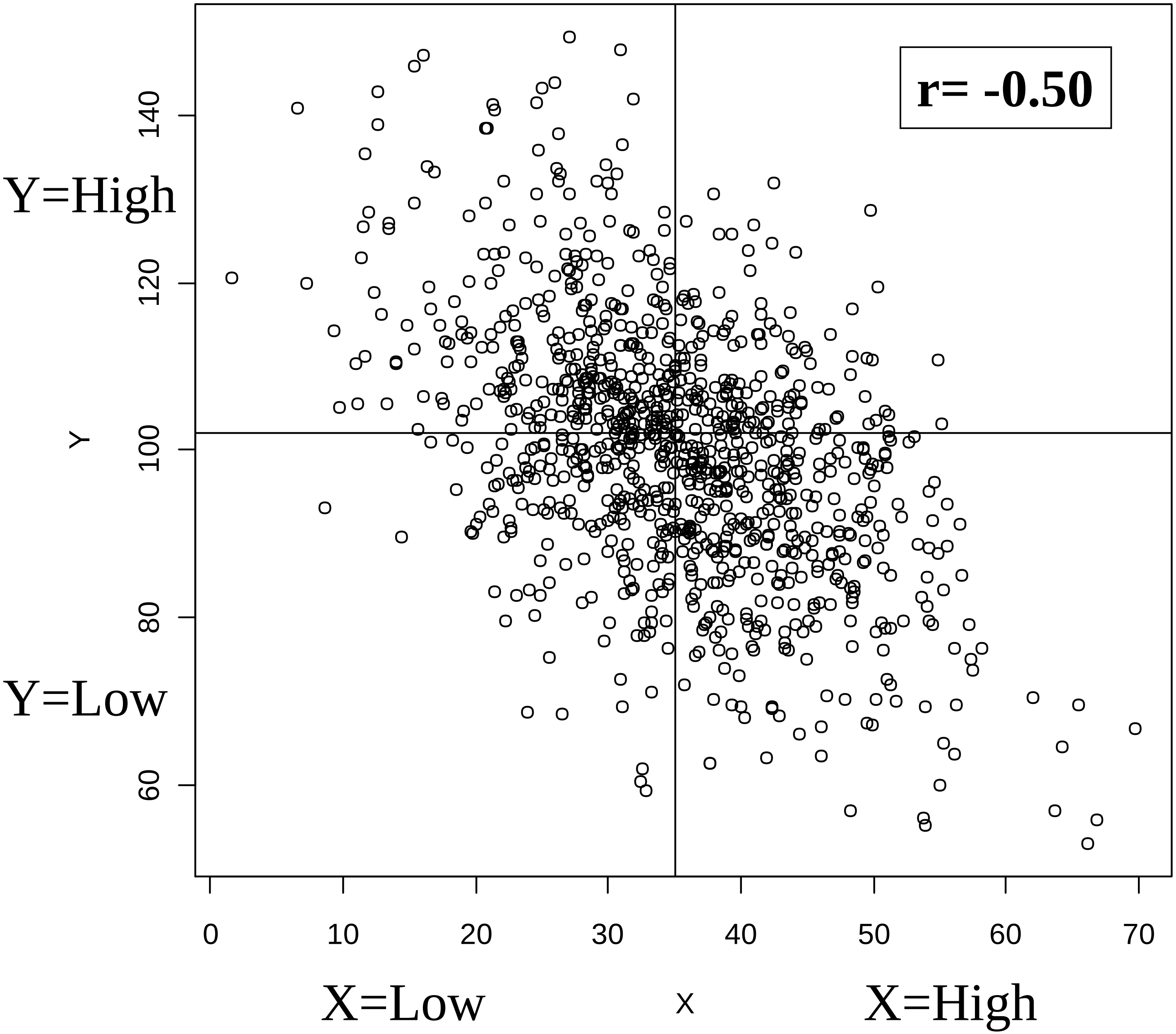
- If you can collect data on a variable in the numeric form rather than as categorical you will have greater power.

- There might be other considerations that require a categorical form, but if possible use the numeric.

Numeric variables for best power (2)

- **Apparently** easier interpretation of results is the wrong reason for making categories.
- The results that follow illustrate the extent of the loss in power if continuous variables are converted to categorical.

Numeric vs categorical variables



- Can treat both variables as numeric
- Y as numeric, X as categorical
- Both X and Y as categorical

Numeric vs categorical variables (2)

- Converting the numeric variables to categorical variables leads to the following tables.

		Means for Y			
Effect size	r	X=Low	X=High	Diff	SD(Y)
Large	-0.5	106.0	94.0	12.0	13.8
Medium	-0.3	**103.6**	**96.4**	**7.1**	14.6
Small	-0.1	101.2	98.8	2.3	14.95

Y=Numeric
X=Categorical

This detail is provided for the interested reader and can be skipped

		Percentages for Y=High		
Effect size	r	X=Low	X=High	Diff
Large	-0.5	61.3%	28.1%	33.2%
Medium	-0.3	**54.3%**	**35.1%**	**19.2%**
Small	-0.1	47.7%	41.7%	6.1%

Both categorical

Numeric vs categorical variables (3)

		Sample size (N) total of both groups		
Effect size	**r**	**Both Numeric**	**Y - numeric X Categorical**	**Both Categorical**
Large	-0.50	30	44	68
Medium	-0.30	85	134	208
Small	-0.10	784	1328	2154
Statistical test		correlation coef.	t-test	chi-squared

- **Converting one numeric variable to categorical sample size increases range from 45% to 70%** **(Large to small effect size)**
- **Converting both numeric variables to categorical sample size increases range from 125% to 175%.**

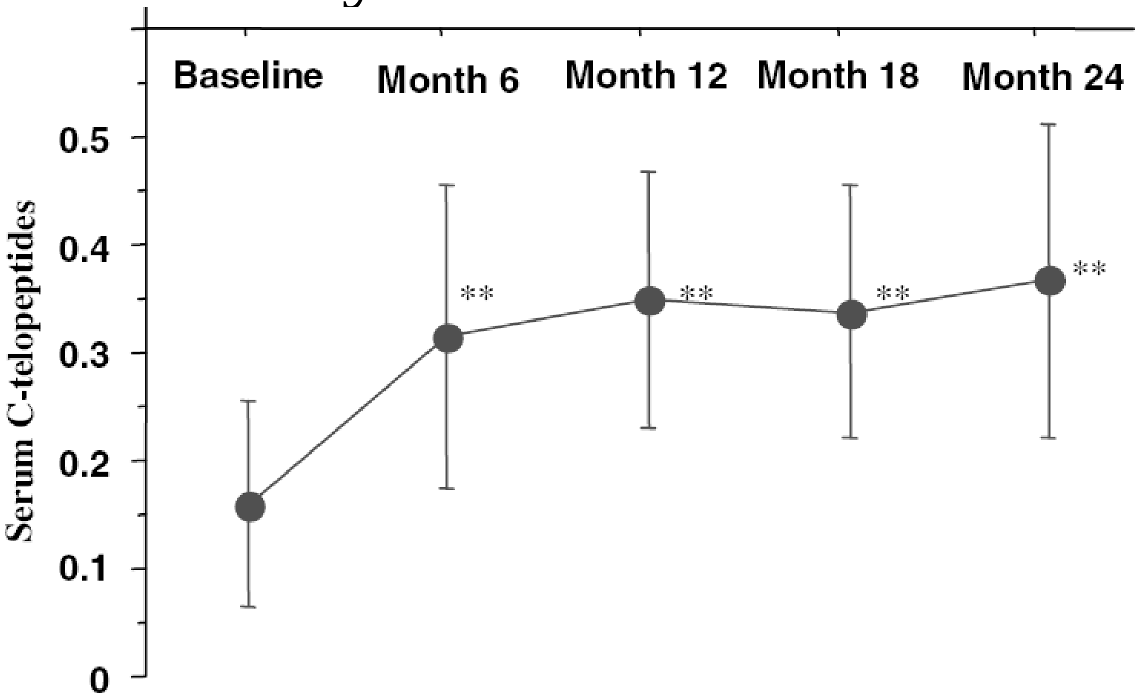
Effect size labels, small, medium and large using Cohen's criteria – earlier slides
r is Pearson correlation coefficient

Relationship between 3 variables

1 Response
2 Explanatory

Case study: Proposed design Pre-Post/Control-Treatment

- This is a common high quality study design
- Some literature data available for variability but only for no intervention case



SD's from Table 2 at each time period range from 0.12 to 0.14

Giusti, V. et al, International Journal of Obesity (2005) 29, 1429-1435
Re effect of Gastric banding on bone growth

Pre, post data are dependent Control, Treatment are independent

- Response: Serum concentration of enzyme
- Treatment: Fortified dairy products.
- Control: Normal dairy products.
- What is a clinically significant improvement?
- Do better than control group by at least 20%
- Assume control does not worsen.
- Treatment reduced by 20%.
- Need to know correlation between pre and post scores, or SD of differences.

Scenario1

- **20% reduction for treatment compared to control**

	Pre	Post	(Pre-Post)
Control	0.16	0.16	0
Treatment	0.16	0.128	0.032
Difference (Treatment - Control)			0.032

- SD = 0.12
- Correlation between pre & post scores = 0.50
  (guess, conservative, not available from paper)
- n=220 in each group, **too much work!**

Scenario2

- **Fixed sample size, n=50, all that can be handled**

	Pre	Post	(Pre-Post)
Control	0.16	0.16	0
Treatment	0.16	0.085	0.075
Difference (Treatment - Control)			0.075

- SD = 0.14
  correlation between pre & post scores = 0.50
- Difference of .075/.16 =
  Only 47% reduction is achievable.
- Both cases were with power = 0.80, type 1 error = .05

Result of sample size analysis

- Desired practical significance to detect 20% change requires more resources than we can afford (n=220 in each group, total N=440).
- Alternative based on resources we can afford (n=50 in each group, total N=100).
  Can only detect a large change 47%.
- What do you do? Only you and your supervisor can answer that question.
- But at least you know you have an issue to solve!

Which test should I use?

Test	Response	Explanatory	This lecture	P &S software
m_a vs. m_b for independent samples	Continuous	Categorical 2 levels	✓	✓
Correlation (r) = 0	Continuous			
Independent r_a vs. r_b	Continuous	Continuous		
Sign test (P = 0.5)	Categorical			
Independent P_a vs. P_b	2 categories	2 categories	✓	✓
χ^2 test	2 or more categories	2 or more categories		
One-way ANOVA (Between Subjects)	Continuous	3 or more categories		
Regression	Continuous	Continuous		✓

Pre, Post measures with Treatment & Control groups ✓

Sample Size and Power Calculation

References

- Power and Sample Size program, Dupont WD and Plummer WD: PS power and sample size program available for free on the Internet. Controlled Clin Trials,1997;18:274 https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=882761
- G*Power 3 – free on the Internet – wider range of calculations than Power and Sample size, most suited to social sciences, strong Psychology basis http://www.gpower.hhu.de
- Howell, D.C, Statistical Methods for Psychology, 5th ed, 2002, pp 223-239. (Continuous variables only)
- Cohen, J. (1992). A power primer, *Psychological Bulletin*, 112, 155-159. (Overview of a range of methods – full text available on-line as PDF)
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd Ed., Earlbaum, NJ. (Comprehensive reference)
- Cumming, G. and Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and non-central distributions, Educational and Psychological Measurement, 61(4), 523-574.
  Confidence intervals – the better direction.
- Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis.* Book web site with demo videos and Excel spreadsheets available (ESCI) http://www.latrobe.edu.au/psychology/research/research-areas/cognitive-and-developmental-