The “Big 4” – Significance, Effect Size, Sample Size, and Power

18 Sep 2009
Dr. Sean Ho
CPSY501

cpsy501.seanho.com

Outline for today

- Stats review:
  - Correlation (Pearson, Spearman)
  - *t*-tests (indep, paired)
- Discussion of research article (Missirlian et al)
- The “Big 4”:
  - Statistical significance (*p*-value, α)
  - Effect size
  - Power
  - Finding needed sample size
- SPSS tips

Measuring correl: Pearson's *r*

- The most common way to measure correlation is Pearson's product-moment correlation coefficient, named *r*:
- Requires parametric data
  - Indep obs, scale level, normally distrib!
- Example: ExamAnxiety.sav
  - Measured anxiety before exam, time spent reviewing before exam, and exam performance (% score)

Pearson's correlation coeff

Name of Correlation Statistic

Significance Value (*p*)

Correlations

		Exam performance (%)	Exam Anxiety	Time spent revising
Exam performance (%)	Pearson Correlation	1	-.441**	.397**
	Sig. (1-tailed)		.000	.000
	N	103	103	103
Exam Anxiety	Pearson Correlation	-.441**	1	-.709**
	Sig. (1-tailed)	.000		.000
	N	103	103	103
Time spent revising	Pearson Correlation	.397**	-.709**	1
	Sig. (1-tailed)	.000	.000	
	N	103	103	103

**. Correlation is significant at the 0.01 level (1-tailed).

Each variable is perfectly correlated with itself!

Spearman's Rho (ρ or r_s)

- Another way of calculating correlation
- Non-parametric: can be used when data violate parametricity assumptions
- No free lunch: loses information about data
- Spearman's works by first ranking the data, then applying Pearson's to those ranks
- Example (grades.sav):
  - grade on a national math exam (GCSE)
  - grade in a univ. stats course (STATS)
  - coded by “letter” (A=1, B=2, C=3, ...)

Spearman's Rho (ρ or r_s): ex

Name of Correlation Statistic

Correlations

			Statistics Grade	GCSE Maths Grade
Spearman's rho	Statistics Grade	Correlation Coefficient	1.000	.455*
		Sig. (1-tailed)	.	.011
		N	25	25
	GCSE Maths Grade	Correlation Coefficient	.455*	1.000
		Sig. (1-tailed)	.011	.
		N	25	25

*. Correlation is significant at the 0.05 level (1-tailed).

Sample Size

The correlation is positive

TRINITY WESTERN UNIVERSITY

Chi-Square test (χ^2)

- Evaluates whether there is a relationship between 2 categorical variables
- The Pearson chi-square statistic tests whether the 2 variables are independent
- If the significance is small enough ($p<\alpha$, usually $\alpha=.05$), we reject the null hypothesis that the two variables are independent (unrelated)
  - i.e., we think that they are in some way related.

t-Tests: comparing two means

- Moving beyond correlational research…
- We often want to look at the effect of one variable on another by systematically changing some aspect of that variable
- That is, we want to manipulate one variable to observe its effect on another variable.
- *t*-tests are for comparing two means
- Two types of application of *t*-tests:
  - Related/dependent measures
  - Independent groups

Related/dependent *t*-tests

- A repeated measures experiment that has 2 conditions (levels of the IV)
- the <u>same subjects</u> participate in both conditions
- We expect that a person's behaviour will be the same in both conditions
  - external factors – age, gender, IQ, motivation, ...
    – should be same in both conditions
- Experimental Manipulation: we do something different in Condition 1 than what we do in Condition 2 (so the only difference between conditions is the manipulation the experimenter made)
  - e.g., Control vs. test

Trinity Western University

Independent samples *t*-tests

- We still have 2 conditions (levels of the IV), but different subjects in each condition.
- So, differences between the two group means can possibly reflect:
  - The manipulation (i.e., systematic variation)
  - Differences between characteristics of the people allotted to each group (i.e., unsystematic variation)
  - Question: what is one way we can try to keep the ‘noise’ in an experiment to a minimum?

t-Tests

- *t*-tests work by identifying sources of systematic and unsystematic variation, and then comparing them.
- The comparison lets us see whether the experiment created *considerably* more variation than we would have got if we had just tested the participants w/o the experimental manipulation.

Example: dependent samples

- "Paired" samples *t*-test
- 12 'spider phobes' exposed to a picture of a spider (**picture**), and on a separate occasion, a real live tarantula (**real**)
- Their **anxiety** was measured at each time (i.e., in each condition).

Paired samples *t*-test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Picture of Spider	40.0000	12	9.29320	2.68272
	Real Spider	47.0000	12	11.02889	3.18377
Pair 2	Picture of Spider	40.0000	12	9.29320	2.68272
	Real Spider	47.0000	12	11.02889	3.18377

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Picture of Spider & Real Spider	12	.545	.067
Pair 2	Picture of Spider & Real Spider	12	.545	.067

TRINITY WESTERN UNIVERSITY

Example: paired *t*-Tests

Degrees of Freedom (in a repeated measures design, it's N-1)

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Picture of Spider - Real Spider	-7.00000	9.80723	2.83110	-13.23122	-.76878	-2.473	11	.031
Pair 2	Picture of Spider - Real Spider	-7.00000	9.80723	2.83110	-13.23122	-.76878	-2.473	11	.031

Standard Deviation of the pairwise difference

Standard error of the differences b/w subjects' scores in each condition

SPSS uses df to calculate the exact probability that the value of the 't' obtained could occur by chance

The probability that 't' occurred by chance is reflected here

Example: indep samples *t*-test

Used in situations where there are 2 experimental conditions – and different participants are used in each condition

Example: SpiderBG.sav

- 12 spider phobes exposed to a picture of a spider (picture); 12 different spider phobes exposed to a real-life tarantula
- Anxiety was measured in each condition

Group Statistics

	Condition	N	Mean	Std. Deviation	Std. Error Mean
Anxiety	Picture	12	40.0000	9.29320	2.68272
	Real Spider	12	47.0000	11.02889	3.18377

Summary Statistics for the 2 experimental conditions

(N1 + N2) - 2 = 22

Independent Samples Test

		Levene's Test fo Equality of Varian		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-taile	Mean Difference	Std. Erro Difference	Lower	Upper
Anxiety	Equal variances assumed	.782	.386	-1.681	22	.107	-7.00000	4.16333	-15.6342	1.63422
	Equal variances not assumed			-1.681	21.385	.107	-7.00000	4.16333	-15.6486	1.64864

Parametric tests (e.g., *t*-tests) assume variances in the experimental conditions are ‘roughly’ equal.

If Levene’s test is sig., the assumption of homogeneity of variance has been violated

Significance (*p*-value): 0.107 > α=.05, so there is no significant difference between the means of the 2 samples

Outline for today

- Stats review:
  - Correlation (Pearson, Spearman)
  - *t*-tests (indep, paired)
- Discussion of research article (Missirlian et al)
- The “Big 4”:
  - Statistical significance (*p*-value, α)
  - Effect size
  - Power
  - Finding needed sample size
- SPSS tips

Practice reading article

- For practice, try reading this journal article, focusing on their statistical methods: see how much you can understand
- Missirlian, et al., "*Emotional Arousal, Client Perceptual Processing, and the Working Alliance in Experiential Psychotherapy for Depression*", Journal of Consulting and Clinical Psychology, Vol. 73, No. 5, pp. 861–871, 2005.
- Download from website, under today's lecture

For discussion:

- What research questions do the authors state that they are addressing?
- What analytical strategy was used, and how appropriate is it for addressing their questions?
- What were their main conclusions, and are these conclusions warranted from the actual results /statistics /analyses that were reported?
- What, if any, changes/additions need to be made to the methods to give a more complete picture of the phenomenon of interest (e.g., sampling, description of analysis process, effect sizes, dealing with multiple comparisons, etc.)?

Outline for today

- Stats review:
  - Correlation (Pearson, Spearman)
  - *t*-tests (indep, paired)
- Discussion of research article (Missirlian et al)
- The “Big 4”:
  - Statistical significance (*p*-value, α)
  - Effect size
  - Power
  - Finding needed sample size
- SPSS tips

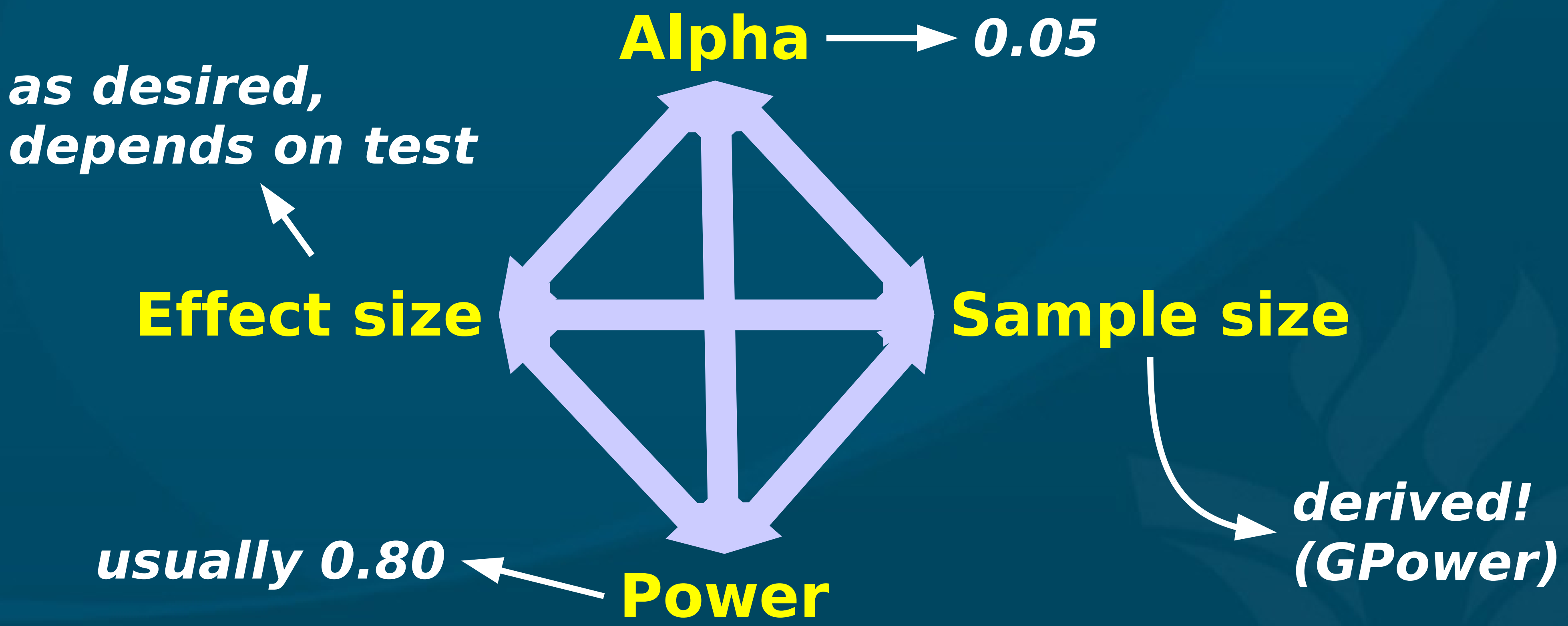
Central Themes of Statistics

- Is there a real effect/relationship amongst certain variables?
- How big is that effect?

- We evaluate these by looking at
  - Statistical significance (p-value) and
  - Effect size (r^2, R^2, η, $\mu_1 - \mu_2$, etc.)
- Along with sample size (n) and statistical power ($1-\beta$), these form the "Big 4" of any test

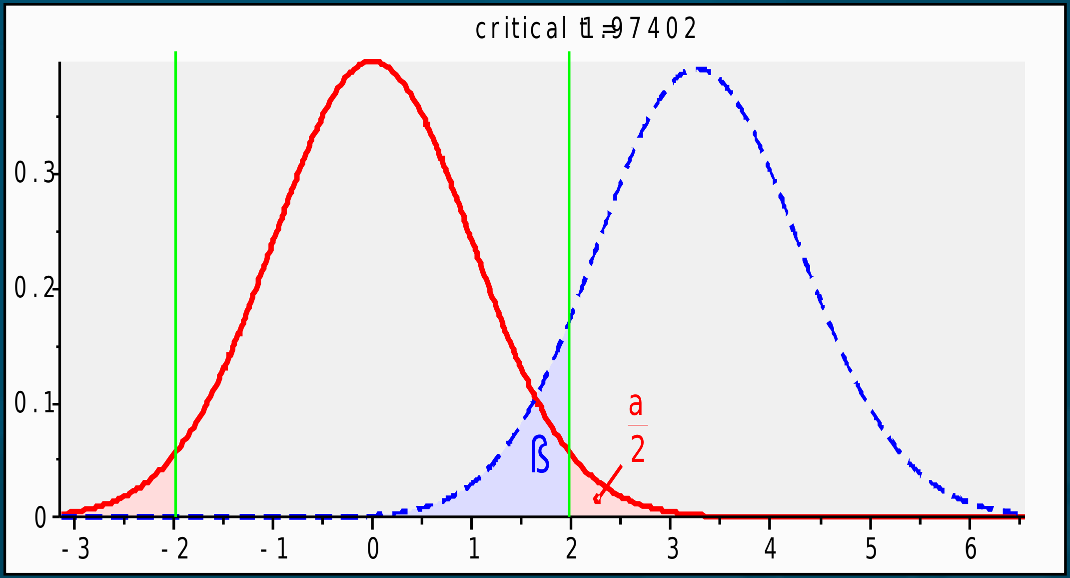
The “Big 4” of every test

- Any statistical test has these 4 facets
- Set 3 as desired → derive required level of 4th



Significance (α) vs. power (1-β)

- α is the chance of Type-I error: incorrectly rejecting the null hypothesis (H_0)
- β is the chance of Type-II error: failing to reject H_0 when should have rejected H_0.
- Power is 1-β
- e.g., parachute inspections
- α,β in independent groups *t*-test:



Statistical significance

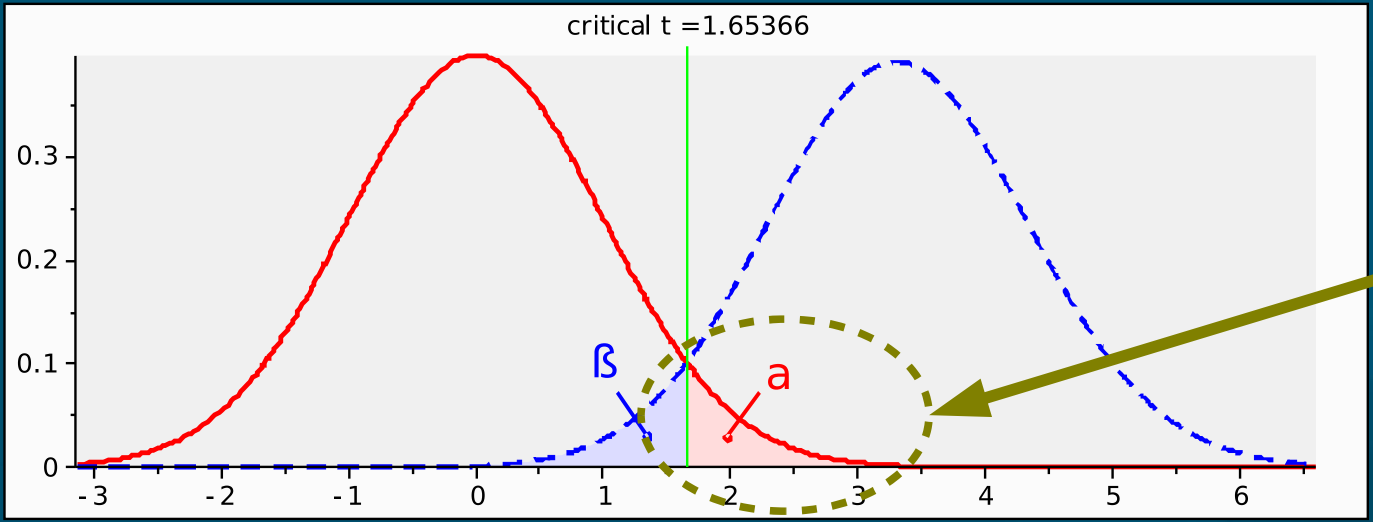
- An effect is “real” (statistically significant) if the probability that the observed result came about due to random variation is so small that we can reject random variation as an explanation.
  - This probability is the *p*-value
- Can never truly “rule out” random variation
- Set the level of significance (α) as our threshold tolerance for Type-I error
- If $p < \alpha$, we confidently say there is a real effect
- Usually choose α=0.05 *(what does this mean?)*

Myths about significance

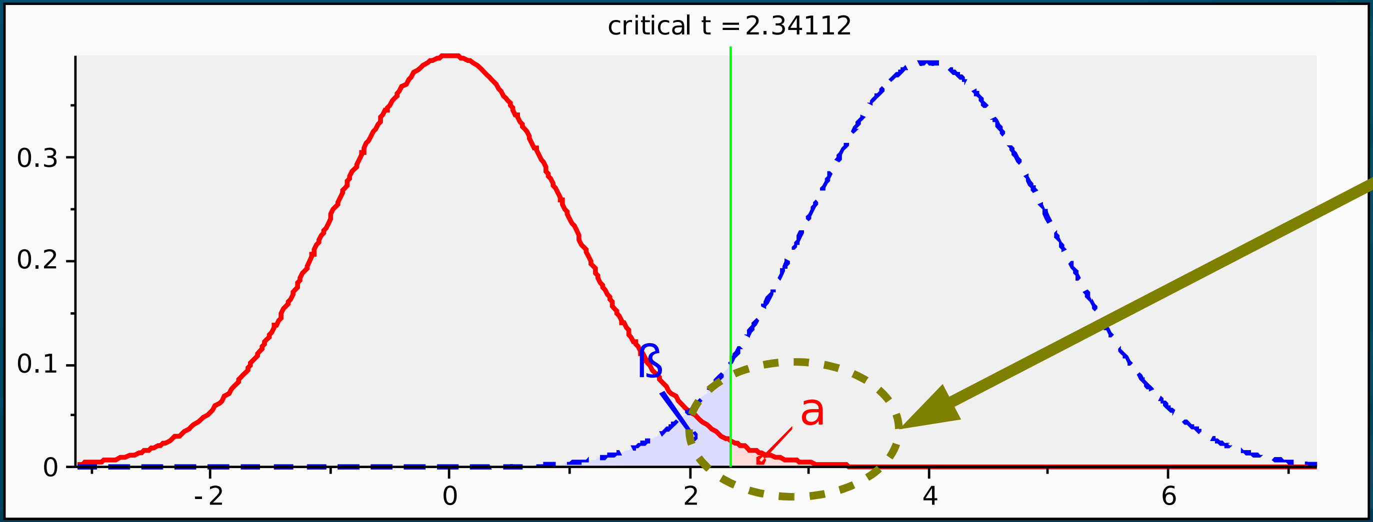
- *(why are these all myths?)*

- Myth 1: “If a result is not significant, it proves there is no effect.”
- Myth 2: “The obtained significance level indicates the reliability of the research finding.”
- Myth 3: “The significance level tells you how big or important an effect is.”
- Myth 4: “If an effect is statistically significant, it must be clinically significant.”

Impact of changing α



α=0.05



α=0.01

Effect size

- Historically, researchers only looked at significance, but what about the effect size?
- A small study might yield non-significance but a strong effect size
  - Could be spurious, but could also be real
  - Motivates meta-analysis – repeat the experiment, combine results
- Current research standards require reporting both significance as well as effect size

Measures of effect size

- For *t*-test and any between-groups comparison:
  - Difference of means: d = $(\mu_1 - \mu_2)/\sigma$
- For ANOVA: η^2 (eta-squared):
  - Overall effect of IV on DV
- For bivariate correlation (Pearson, Spearman):
  - r and r^2: r^2 is fraction of variability in one var explained by the other var
- For regression: R^2 and ΔR^2 ("R^2-change")
  - Fraction of variability in DV explained by overall model (R^2) or each predictor (ΔR^2)

Interpreting effect size

- What constitutes a "big" effect size?
- Consult literature for the phenomenon of study
- Cohen '92, *"A Power Primer"*: rules of thumb
  - Somewhat arbitrary, though!
- For correlation-type *r* measures:
  - 0.10 → small effect (1% of var. explained)
  - 0.30 → medium effect (9% of variability)
  - 0.50 → large effect (25%)

Example: dependent *t*-test

- Dataset: SpiderRM.sav
- 12 individuals, first shown picture of spider, then shown real spider → measured anxiety
- Compare Means → Paired-Samples T Test
- SPSS results: $t(11)=-2.473$, $p<0.05$
- Calculate effect size: see text, p.332 (§9.4.6)
  - $r \sim 0.5978$ (big? Small?)
- Report sample size (*df*), test statistic (*t*), *p*-value, and effect size (*r*)

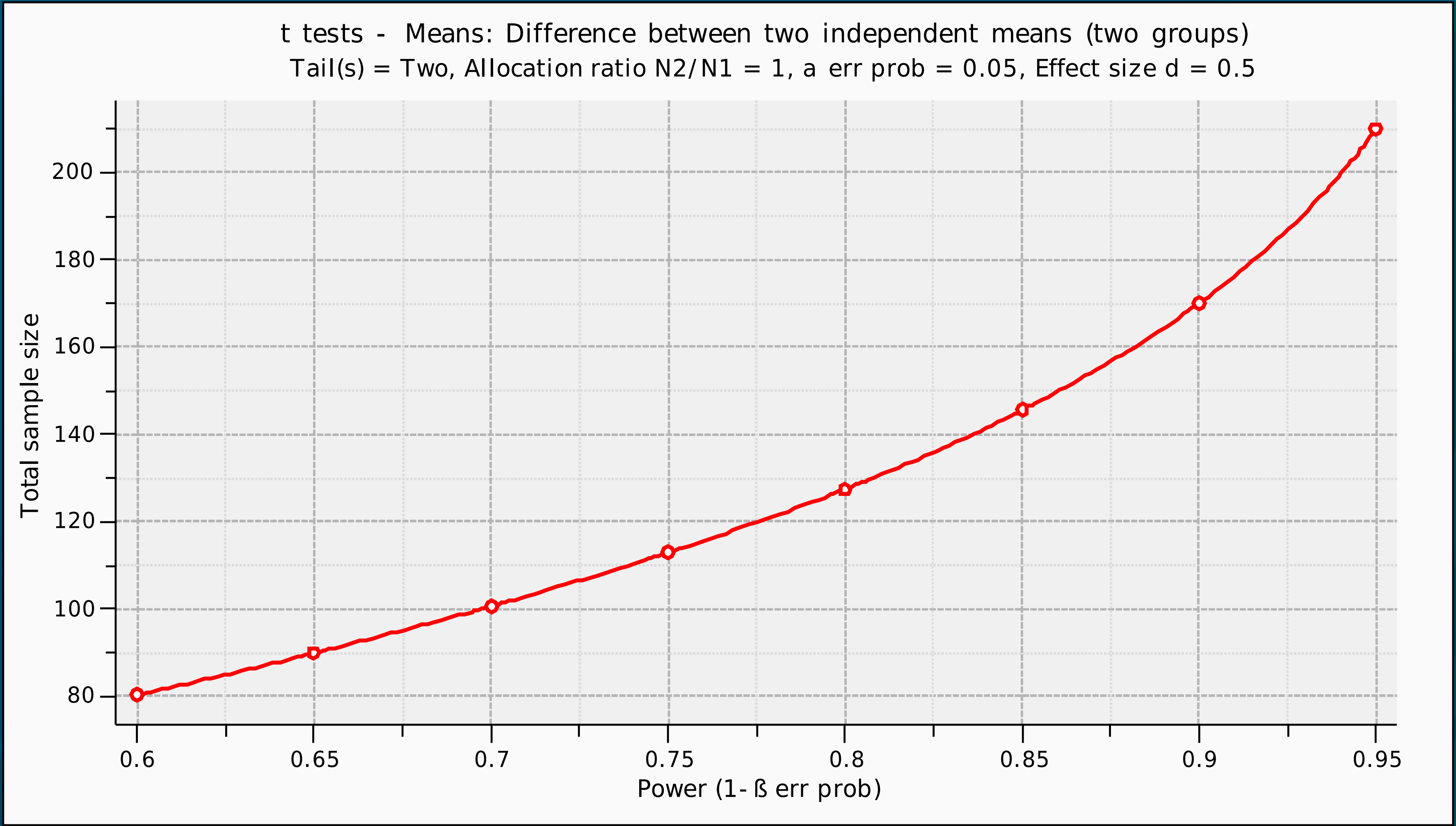
$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

Finding needed sample size

- Experimental design: what is the minimum sample size needed to attain the desired level of significance, power, and effect size (assuming there is a real relationship)?
- Choose level of significance: $\alpha = 0.05$
- Choose power: usually $1-\beta = 0.80$
- Choose desired effect size: (from literature or Cohen's rules of thumb)
- → use GPower or similar to calculate the required sample size (do this for your project!)

Power vs. sample size

- Fix α=0.05, effect size d=0.50:



SPSS tips!

- Plan out the characteristics of your variables before you start data-entry
- You can reorder variables: cluster related ones
- Create a var holding unique ID# for each case
- Variable names may not have spaces: try “_”
- Add descriptive labels to your variables
- Code missing data using values like '999'
  - Then tell SPSS about it in Variable View
- Clean up your output file (*.spv) before turn-in
  - Add text boxes, headers; delete junk