



Statistical Power, Effect Size and Sample Sizes

Stacey Coleman

● Statistical Power, Effect and Sample Sizes

- Statistical significance versus clinical significance
- The importance of sample size and the factors that need to be considered to calculate an appropriate sample size
- Power
- Effect sizes

● Statistical Significance

- We are never **certain** of an effect/relation, so we always deal in the area of probability (p-values!)
- We (usually) test for one of two main things:
 - *Differences between scores*
 - Usually sets of scores calculated as grand means e.g., t-test
 - *Associations between scores*
 - Is there a relationship between (x) and (y) e.g., correlation

● Reminders!

- 0.05 means there is a 5% likelihood that the results were due to chance/random (null hypothesis)
- 1.0 would mean a 100% certainty that the results were down to chance
- In statistical language, 'significant' simply means that there was a less than 5% chance of getting a result under what would be correct for the null hypothesis to be true...

● Reminders!

- But! Finding a statistically significant result doesn't necessarily mean that you can now go out and change the world
 - Likewise, not finding a statistically significant result doesn't necessarily mean that there is nothing of interest...Statistical significance is determined through conducting inferential tests (the stuff you've been doing), and is noted as the sig. or sig. value in SPSS

● Reminders!

- **Type 1 error:** (find a significant p-value in error)
 - Reject the null hypothesis when it is correct
 - Accept the alternative hypothesis in error
 - Displayed like this: Type I error
- **Type 2 error:** (find a non-significant p-value in error)
 - Reject the alternative hypothesis in error
 - Accept the null hypothesis by mistake when there really is an underlying effect
 - Displayed like this: Type II error

● Clinical Significance

○ *The practical importance of a study, with a genuine effect on real life!*

- E.g., how effective is a new treatment?
- E.g., does one type of intervention work better than another?

● Clinical Significance

- If findings from research are considered **clinically significant** then they are likely to influence how actual patients in the real world are treated
- It could amend the way in which health professionals work and what guidance is given from regulatory bodies, and even provide supporting evidence to allow government to pass new or update legislation

● Sample Size

- We typically use estimates or % chance/confidence
- It would be great if we could collect data from an entire population – we could remove this chance and estimation element
 - This is impossible to achieve in most cases, and not needed
- So, we select a sample of individuals who we believe are representative of the population we are looking to investigate

● Sample Size

- Common sense suggests that having the largest sample size possible would be of most benefit in order to make meaningful conclusions about the population
 - If your sample size is too small, then you may not detect an important underlying effect.
 - But – in clinical studies in particular, the population that we can draw our sample from may already be small (e.g., patients with a rare disease).

● Sample Size

- So, we need to try and get the balance right between getting enough P's to detect an effect without running in to the problem of trying to get an unrealistically large sample.
- We can calculate an optimal sample size – this must be done during **the design stage**

Calculating Sample Size

We need to specify the following quantities during the **design stage**:

POWER

- The chance of detecting a statistically significant effect (if there is one). Power should be 70%+.

SIGNIFICANCE LEVEL

- Usually remains at 0.05 (point where we reject the null)

SMALLEST EFFECT OF INTEREST

- Magnitude of the effect that is clinically important – i.e., the difference (this can be expressed as the **Standardised Difference**)

● Calculating Sample Size

- Power and significance are relatively easy to specify
- Obtaining a value for the standardised difference requires two further things:
 1. An estimate of the smallest difference in means needed
 2. An assumed equal standard deviation of the observations (either in groups or across responses)

● Methods for Calculating Sample Size

General formulae

Quick formulate

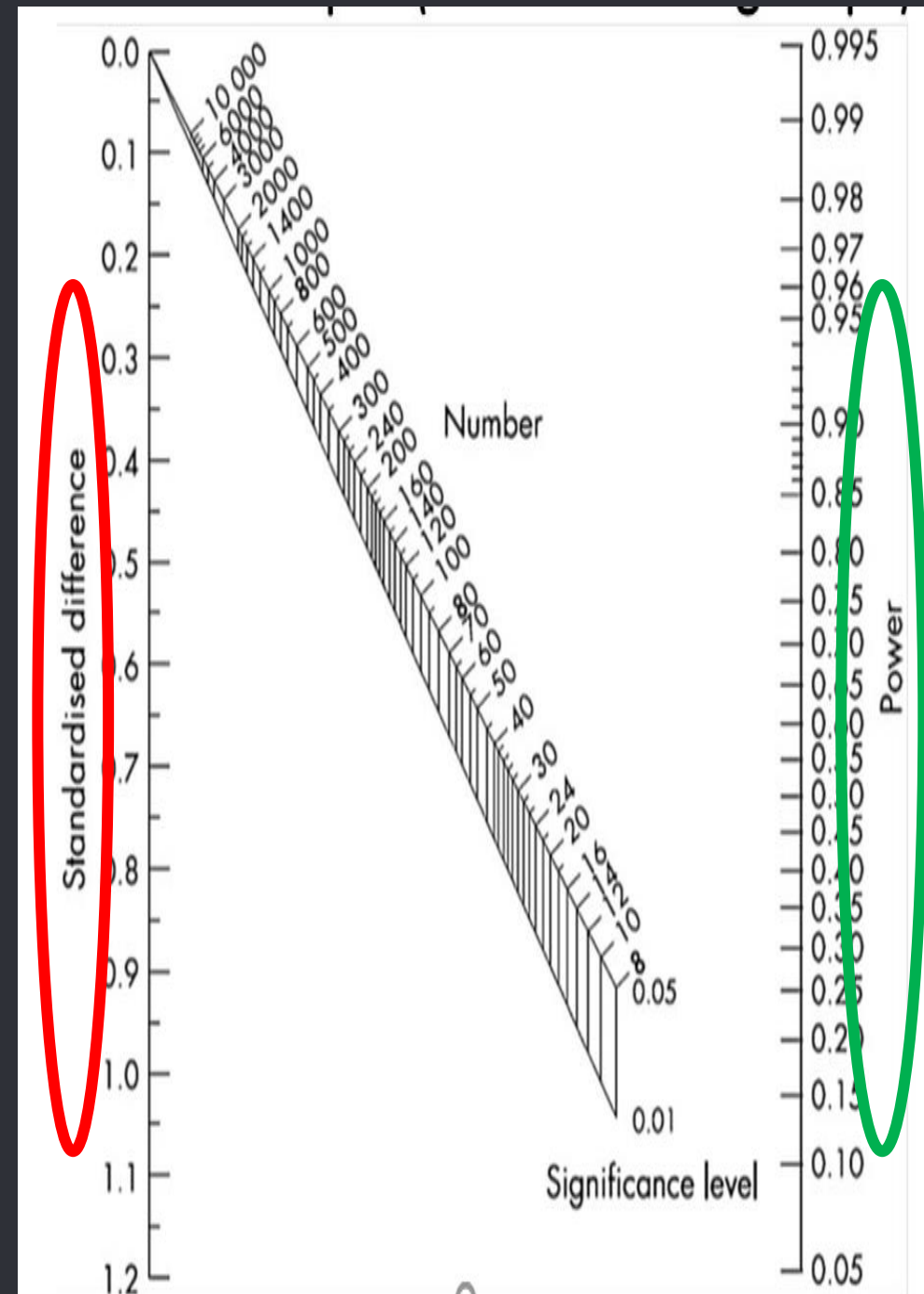
Special tables

Computer packages

1. Altman's Nomogram

Requires the calculation of the **standardised difference** and to join values on the left-hand axis of the nomogram to the power we have specified on the right-hand side of the axis

NB: also need to specify alpha level - .05 or .01



1. Altman's Nomogram

The required **sample size** is indicated at the point at which the resulting line and sample size axis meet

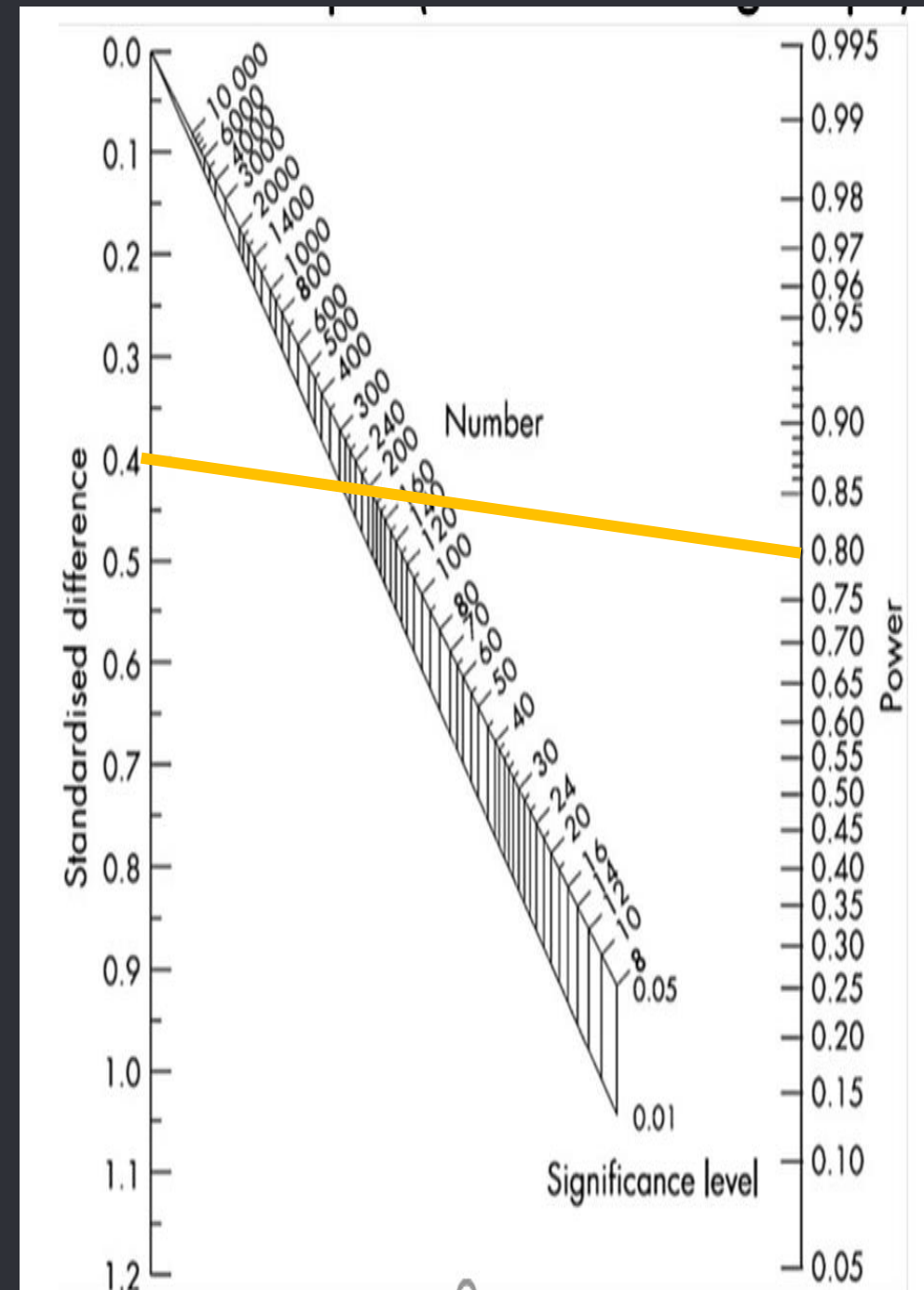
Example

Power = 80%

Standardised difference = 0.4

Alpha level = 0.05

200 participants



● Calculating the Standardised Difference

- The formulae for calculating the standardised difference will depend upon the test you are looking to conduct:

Independent t-test:

$$\frac{\delta}{\sigma}$$

Where δ : is the smallest difference in means
 σ : assumed equal standard deviation of the
observations in each of the 2 groups

● Calculating the Standardised Difference

Paired t-test:

$$\frac{2\delta}{\sigma_d}$$

Where δ : is the smallest mean difference

σ_d

σ_d : the standard deviation of the differences in responses

- Calculating the Standardised Difference

Chi² test:

$$\frac{P_1 - P_2}{\sqrt{\bar{p}(1-\bar{p})}}$$

$$\sqrt{\bar{p}(1-\bar{p})}$$

$P_1 - P_2$ = the smallest difference in the proportion of 'success' in the groups

To calculate $\bar{p} = \frac{P_1 + P_2}{2}$

● 2. Quick Formulae: Lehr's Formula

For independent t-test and the Chi² test, you can use Lehr's formula for calculating the **sample size** for a power of 80% and a two-tailed significance level of 0.05

$$\frac{16}{(\text{Standardised difference})^2}$$

Note: the numerator of 21 is used (instead of 16) to relate to a power of 90%.

● Power Statement

- With applied research, you will more than likely need to include a **power statement** as part of your study protocol or in the method section of a report to show that you considered the sample size during the designing stages of your study
- A typical statement could be:
 - “84 patients in each group were required to have a 90% chance of detecting a difference in means of 2.5 days (SD = 5 days) at the 5% level of significance using an Independent samples t-test.”

● Effect size

- The REAL effect in the population; the magnitude of the difference between groups
 - Sometimes it's not enough to simply know that an effect is significant
- We may also want to know the size of the effect
 - E.g., we may want to know by how much does a new drug reduce episodes of depression
- Whether the manipulation of an IV has created a sizeable difference between our samples on some measure
 - The further the difference between the two means, the larger the effect size

● Effect size

- The popular measures of effect size to compare groups are Cohen's d and eta squared (η^2)/partial eta squared.
- In Psychology we typically use **partial eta squared**, which is what SPSS provides us with in the very last column of our inferential statistics box when conducting more complex analyses (such as ANOVA).
 - If we have remembered to tick Display 'Estimates of effect size' in the Options Dialogue box!

● Effect size: Eta Squared

- Guidelines proposed by Cohen (1988) for interpreting the eta squared value are:

0.01 – 0.06 = small

0.06 – 0.14 = medium

> 0.14 = large

- *NB: eta squared more appropriate to report as the measure of effect size when your model has more than one IV.*

● Effect size: Cohen's d

- Cohen's d reporting of effect size:

0.2 = small

0.5 = medium

0.8 = large

- *NB: Cohen's d indicates the size of the difference between two means in standard deviation units. Useful to report when you are only comparing two means.*

● Effect size: Pearson's r

- Correlation co-efficient values are also relevant measures of effect size

0.2 = small

0.5 = medium

0.8 = large

- *NB: Pearson's r for use with correlational / relationship studies – not suitable for comparison of differences in groups means*

● Task for Today

1. Practice using Altman's nomogram to estimate sample sizes
2. Write up power statements for each of the scenarios

Group work:

- Assessment of attraction should be chosen
- Work done on AI images
- Qualtrics accounts
- Materials started (information sheet, consent form, debrief)