

Applied Research Design and Analysis (ARDA)

Statistical and Clinical Significance

Sample Size

Power

Effect Size

Confidence Intervals (CI)

Dr Kate Williams

1

Recommended Reading

Brace, N., Kemp, R., & Snelgar, R. *SPSS for Psychologists* Palgrave Macmillan.

Coolican, H. *Research methods and statistics in psychology*. Taylor & Francis.

Cumming, G. *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. Routledge.

Pallant, J. *SPSS survival manual*. Open University Press.

Petrie, A., & Sabin, C. *Medical Statistics at a Glance*. Blackwell Science Ltd.



2

Lecture Outline

- Reminder of statistical significance
- **Clinical** significance and how it differs to statistical significance
- Importance and factors to consider when calculating sample size
- Considerations for interpretation of results **other than** significance:
 - Power
 - Effect size
 - Confidence Intervals (CI)
- Calculating appropriate sample sizes
- Interpreting and reporting CI and effect size.



3

Statistical Significance

- Always deal in area of **Probability**: Never certain of an effect/relation
- Two main aspects we are usually (probability) testing:
 - Difference between scores:
 - sets of scores calculated as grand means - E.g., t-test
 - Association between scores
 - E.g., Score for extroversion and altruism - E.g., correlation
- The 'p value' represents the probability we would find our results if the null hypothesis was true.
- Convention for significance testing:
 - p value **less** than 0.05 means **significant**
 - p value **more** than 0.05 means **not significant**

4

Statistical Significance

- 0.05 means there is 5% likelihood the results were due to chance/random (null hypothesis).
- 1.0 means a 100% certainty results were due to chance.
- Psychologists (and other sciences) accept that below 5% of a 'null hypothesis' means results show a significant effect.
 - Convention for reporting significance would be $p < 0.05$.
 - However, you can report significance values of $p < 0.01$, or $p < 0.001$.
- Statistical Significance found through **inferential** statistics (i.e., make inferences about the data).

5

Errors in significance reporting

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

6

Clinical Significance

- **Statistical** significance can in some cases NOT be particularly interesting or even important:
 - Just shows significant effect (difference or relationship).
- In statistical language, 'significant' means there was a less than 5% chance of getting a result under H_0 .
- Does not always mean you can now change the world
- Likewise, not finding a statistically significant result does not always mean there is nothing of interest.



7

Clinical Significance

- **Practical** importance of study/genuine effect on real life.
 - E.g., How effective is a new treatment?
 - E.g., Does one type of intervention work better than another?
- If findings from research are considered clinically significant, they can influence how actual people in real world are treated.
- It can amend way health professionals work and what guidance is given from regulatory bodies.
- Could be used as evidence to allow government to pass new or update legislation.

8

Let's look at the factors of importance for clinical significance....

9

Sample Size

- Some factors involve estimates/being only a certain % confident.
- So, would be great if we could collect data from whole population
 - This would remove 'chance' and estimations that plague results and conclusions.
- But this may be impossible to achieve
 - So we do next best thing = collect data on sample of individuals representative of a population we wish to investigate.
- Then use data to draw conclusions about the population.
- However, information from sample may not fully reflect what is true of population.

10

Sample Size

- By only sampling some of the population we have introduced in to our study what is known as
 - **"the sampling error"**
- To obtain representative sample we need to ensure those selected were done at random.
 - A sampling-frame: list of individuals for the population is drawn-up and people are selected randomly.
 - If this is not possible, a convenience sample is collected.
 - E.g., instead of collecting data from all cancer patients, investigate all cancer patients from a particular hospital.
- Having the largest sample size possible is of most benefit to make meaningful conclusions about the population.

11

Sample Size

- This is fine if our sample could contain any type of person from the population.
- However, in applied research (particularly clinical research) the population to draw the sample from may already be small (e.g., patients with a rare disease).
- Also, ethical, logistical, time, cost issues to consider
 - E.g., you want to examine whether a drug that has been previously used on adults to manage a condition could be administered to children.
- **What issues might arise here and why?**

12

Sample Size

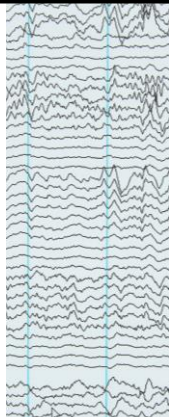
- If sample size is too small, may not detect important underlying effect!
- Balance is key between enough participants to detect effect without running in to problems of trying to get an unrealistically large sample.



13

Calculating Optimal Sample Size

- Need to specify the following quantities at designing stage
- Power:
 - Chance of detecting statistically significant effect
 - Power should be 70% +
- Significance level:
 - Cut-off at which we reject the null (0.05)
- Variability:
 - Variance in observations - e.g., Standard Deviation
- Smallest effect of interest:
 - Magnitude of effect that is clinically important
 - I.e., difference (between means or proportions)
 - Can be expressed as the Standardised difference



14

Calculating Optimal Sample Size

- Power
 - Significance
 - Effect (to a certain degree)
- } Easy to determine
- Providing estimate of variance before any data has been collected is the difficult task.

15

Methods for calculating sample size

- General formulae
- Quick formulae
- Special tables
- Computer packages
- **Altman's nomogram**
 - Requires calculation of standardised difference
 - Join value on left hand axis of nomogram to power we specify on right hand side of axis.
 - **Required sample size is indicated at point which resulting line and sample size axis meet.**
 - Can use nomogram to calculate power of a test for given sample size.
 - Allows retrospective attribution of lack of significance to low sample size.

16

Formulae for Altman's Nomogram

- 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
- Paired t-test = $\frac{2\delta}{\sigma_d}$ Where δ : is the smallest mean difference
 σ_d : the standard deviation of the differences in responses
- Chi-Squared test = To calculate $\bar{p} = \frac{p_1 + p_2}{2}$
 $\sqrt{p(1-p)}$ Where P_1-P_2 : smallest difference in the proportion of 'success' in the 2 groups

17

Power Statement

- Needs to be included either in study protocol or methods section of report
 - Shows sample size is considered during study design stage
- Example of typical statement:
 Eighty-four 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.

18

Quick Formulae

- For Independent t-test and Chi-squared test, can use Lehr's¹ formula for calculating sample size for a power of 80% and a two-tailed significance level of 0.05.

- The formula:

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

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

¹Lehr, R. (1992). Sixteen s squared over d squared: a relation for crude sample size estimates. *Statistics in Medicine*, 11, 1099-1102.

19

What else should we consider?



- Many researchers focus on P value (significant main effect? – Yes, Great! / No, why not?)
- However, we must also consider the Power of the test we are using and the effect size of the outcomes when interpreting results.

20

Power and Effect size

- If you run analysis and get p value of 0.04
 - Technically statistically significant (as < 0.05).
- But if you ran another analysis on different set of data and got p value of 0.0001 (again statistically significant).
- Can you say second result is 'better'/'stronger'?
- **NO!**
- P depends on several factors:
 - Type of test being used
 - Sample size
 - Effect size
 - Power

21

Power

- Power of a test is the probability of detecting an effect (if it is there) and thus **NOT** making a Type II error.
- I.e., detecting a real statistically significant treatment effect of a given size.
- Power is represented as the value $1-\beta$.
- Although we wish for our power to be 100%, there is always small chance we make Type II error.
- However: because we can consider the factors that influence our power we can control it.



22

Power

- Before we begin any study we must ensure good chance of detecting significant result (if there is one).
- Acceptable level of Power is considered at least .7 (70%) but ideally should aim for .8 (80%) or higher.
- Question: Why would conducting **Applied** research with power of less than .7 be an issue?
 - Hint = Think beyond just numbers/values
- E.g., when running a clinical trial, it would be very wasteful of time, money, resources and be ethically irresponsible if it had only a 45% chance of detecting a real treatment effect.



23

Power

- Factors that have direct influence on Power.....
 - Sample size:
 - Power increases with larger sample size.
 - If effect being examined is small then having a larger sample size would increase the power to detect the effect.
 - Variability:
 - The less the variability, the increase in the power.
 - The effect size:
 - Larger effects means greater power.
 - The Significance level:
 - Power is greater if significance level is larger (i.e., deciding before the study that the α will be 0.05 instead of 0.01).

24

Power

- Inspection of Confidence Intervals (CI) for effect of interest gives indication whether power of test was adequate.
- Wide CI results from small sample/data with substantial variability and suggests poor power.
 - 2.85 – 54.64 (wide) vs. 2.85 – 4.38 (narrow)

25

How can we increase Power?

- Increase the sample size
- Increase α
- Increase effect size (previous literature, pilot study, change the levels of an IV).
- Decrease population variance (change aspects of task).
- Aim is to achieve narrow sampling distributions, which subsequently leads to less overlap, and consequently decreases β .
- Stats books show how Power is calculated and reported.
- Need to know effect size and sample size of study to do this.

26

Effect Size

- Sometimes not enough to say an effect is significant.
- May also want to know **Size** of the Effect:
 - E.g., by how much does new drug reduce episodes of depression
- Effect size is the **REAL** effect in the population.
- Has the manipulation of IV created a sizeable difference between our samples on some measure
- Difference between 2 population means (μ_1 and μ_2).
- The further the distance between the 2 population means, the larger the effect size.
- Effect size statistics provide an indication of the magnitude of the difference between groups or scores.

27

Effect Size

- E.g., from CARIAD data
 - Did leaflet help patients deal with waiting for diagnosis compared to patients who didn't receive leaflet?
 - Important for applied work, also want to see if significant difference (if there was one) was sizeable:
 - i.e., leaflet was very much more successful in helping patients as opposed to simply being just a little bit helpful.
- Some cases where effect size are subtle.
- So, what can we do to ensure a small effect is not overlooked?
 - Increase intended sample size.

28

Effect Size

- Popular measures of effect size to compare groups are Cohen's d and eta squared (η^2)/partial eta squared.
- SPSS provides this in inferential statistics box.
 - Remember to tick Display 'Estimates of effect size' in the Options Dialogue box!
- Can also use information from SPSS to calculate effect size statistic either by hand or using online effect size calculators.

29

Interpreting effect size

Guidelines (proposed by Cohen, 1988 – pp. 284-7) for interpreting the value are:
.01 = small; .06 = medium, 0.14 = large

- **IMPORTANT:** It can be confusing but please be aware that although Cohen proposed guidelines for interpreting the strength of the effect size for eta Square, these are not the same as the interpretation for Cohen's d values – so be careful!

30

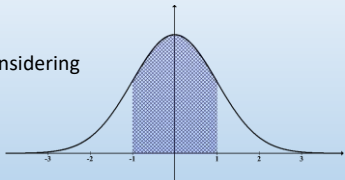
Confidence Intervals (CI)

- Although distribution from sample can tell us if differences are significant, we can use distribution of sample to calculate Confidence Intervals (CI).
- CI are best estimate of range of population value achieved based on sample value.
- It is level of confidence the population mean will fall between two scores/values.
 - E.g., we are 95% confident of the likely range within which a true population mean lies, based on estimate from sample mean and standard error (standard deviation of a sampling distribution).
- Mean of sample can be used as estimate of population mean.
 - Known as estimating population parameters from sample statistics.



31

Confidence Interval of the Mean

- If sample size is large, the sample mean will likely follow normal distribution.
 - We can this when considering the sample mean.
- 
- 95% of the distribution of sample means lies within 1.96 standard deviations (SD) of the population mean.
 - The **standard deviation** is a measure of the amount of variation or [dispersion](#) of a set of values (from the mean)

32

Confidence Interval of the Mean

- With a single sample, the SD is called the standard error of the mean (SEM)
We can calculate the 95% confidence interval of the mean as:
 - (sample mean - (1.96xSEM), sample mean + (1.96xSEM))
- SEM is **calculated** by taking **standard deviation** and dividing it by square root of sample size.
 - **Square root** of a number is a value that, when multiplied by itself, gives the number.
 - Example: $4 \times 4 = 16$, so a **square root** of 16 is 4.
 - Note that $(-4) \times (-4) = 16$ too, so -4 is also a **square root** of 16.
 - The symbol is $\sqrt{\quad}$ which always means the positive **square root**.
- But: Normal distribution can only be used if we know the value of the variance in the population.
 - Also, if sample size is small, the sample mean only follows a normal distribution if the population are normally distributed.

33

Confidence Interval of the Mean

Using the t-distribution:

- Calculate the 95% confidence interval for the mean as:
 - (sample mean - (t.05xSEM), sample mean + (t.05xSEM)).
- t.05 = the percentage point of the t-distribution with (n-1) degrees of freedom and a two-tailed probability.
 - See Appendix of stats books for this value.
- It gives wider CI than using normal distribution to allow for extra uncertainty we have introduced by either estimating population SD or from small sample size.
- Typically use 95% CI but can calculate 99% CI for the mean.
- This is done by multiplying standard error by the t-distribution corresponding to a two-tailed probability of 0.01.
- Gives a wider interval to reflect increased confidence that the range includes the population mean.

34

Confidence Interval for the proportion

- Sampling distribution of a proportion follows a Binominal distribution.
- If the sample size is large then sampling distribution of the proportion is approximately Normal with the mean.
- We can estimate the true probability for each case by the proportion in the sample, $p = r/n$
 - r is the number of individuals in the sample with the characteristics of interest.
- The standard error is estimated by $\frac{\sqrt{p(1-p)}}{n}$

The 95% confidence interval for the proportion is:

$$\left(p - \left[1.96 \times \frac{\sqrt{p(1-p)}}{n} \right], p + \left[1.96 \times \frac{\sqrt{p(1-p)}}{n} \right] \right)$$

35

Interpretation of Confidence Intervals

- How wide is it?
 - Width depends on size of standard error, which depends on sample size, and variability of the data.
 - Small studies on variable data give wide CI, while large studies on less variable data give narrower CI.
 - Wide CI = estimate is not precise
 - Narrow CI = estimate is precise
- What Clinical Implications can be derived?
 - Upper and lower limits provide way of assessing clinical importance
- Are there any values of particular interest?
 - Can check if hypothesised value for population parameter falls within CI. If so, results are consistent with hypothesised value.
 - If not, it's unlikely parameter has the value (remember with 95% CI, chance is 5%).

36