

MSc Applied Social and Health Psychology

Applied Research, Ethics, and Analysis (AREA)

Statistical and Clinical Significance

Sample Size

Power

Effect Size

Confidence Intervals (CI)

Dr Kate Williams

Recommended Reading

- Brace, N., Kemp, R., & Snelgar, R. (2016). *SPSS for Psychologists* (6th ed.). Hants: Palgrave Macmillan.
- Coolican, H. (2014). *Research methods and statistics in psychology*. London: Taylor & Francis.
- Cumming, G. (2012). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. London: Routledge.
- Pallant, J. (2016). *SPSS survival manual*. (5th Ed.). Buckingham: Open University Press.
- Petrie, A., & Sabin, C. (2000). *Medical Statistics at a Glance*. Blackwell Science Ltd.

Lecture Outline

- Reminder of statistical significance.
- Discussion of *Clinical* significance and how it differs to statistical significance.
- Importance of sample size and factors to consider when calculating it.
- Run through of what we need to consider for interpretation of results ***other than*** significance:
 - Power
 - Effect size
 - Confidence Intervals (CI)
- Run through calculating appropriate sample sizes, and look at how we interpret and report CI and effect size.

Statistical Significance

- Probability: We are never certain of an effect/relation.
- Always deal in the area of probability.
- Two main things we are usually (probability) testing:
 - Difference between scores:
 - usually sets of scores calculated as grand means - E.g., t-test
 - Association between scores
 - E.g., Score for Ps on 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**

Statistical Significance

- 0.05 means there is a 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 even $p < 0.001$.
- Statistical Significance is found through conducting *inferential* statistics (i.e., make inferences about the data).

Errors in significance reporting

- Type 1 error: (find a significant p-value)
 - 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)
 - 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

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

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?
- If findings from research are considered clinically significant, they can influence how actual patients in the real world are treated.
- It can amend the way health professionals work and what guidance is given from regulatory bodies.
- It could even go as far as providing supporting evidence to allow government to pass new or update legislation.

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

Sample Size

- Some factors in today's lecture involve estimates or being only a certain % confident.
- When conducting research it would be great if we could collect data from the whole population
 - This would remove the 'chance' and estimation elements that plague our results and conclusions.
- But this is impossible to achieve
 - So we do next best thing = collect data on sample of individuals representative of a population we wish to investigate.
- We then use data and results to draw conclusions about the population.
- However, information from the sample may not fully reflect what is true of the population.

Sample Size

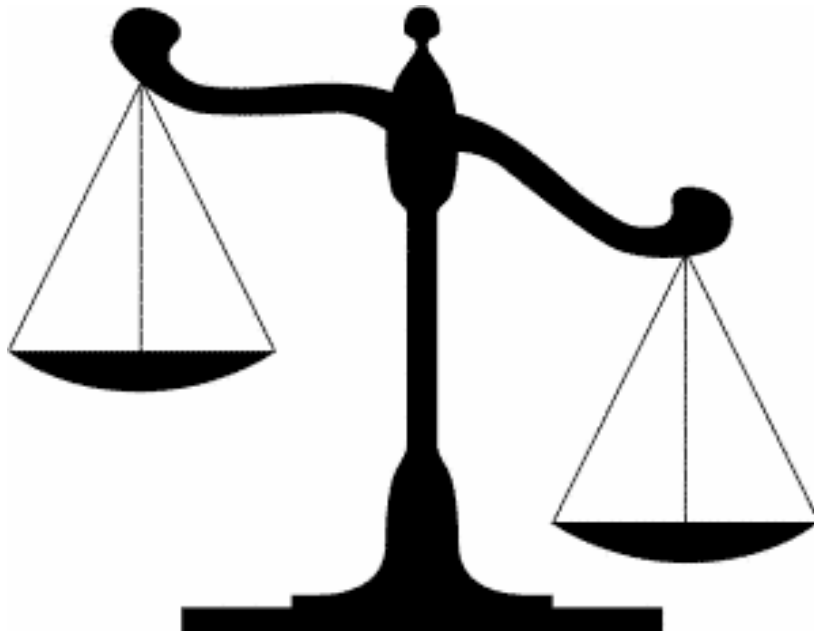
- By only sampling some of the population we have introduced in to our study what is known as
 - “*the sampling error*”
- To obtain a representative sample we need to ensure that those selected were done so at random.
 - A sampling-frame i.e., a list of individuals for the population is drawn-up and people are selected randomly.
 - If this is not possible then 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.

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).
- You may encounter ethical, logistical, time, cost issues.
 - 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?*

Sample Size

- But: If your sample size is too small then you may not detect an important underlying effect!
- Balance is key between getting enough Ps to detect an effect without running in to problems of trying to get an unrealistically large sample.



Calculating Optimal Sample Size

- Need to specify the following quantities at **designing** stage
 - Power:
 - Chance of detecting a statistically significant effect (if one).
 - Power should be 70% +.
 - Significance level:
 - Cut-off at which we reject the null (0.05).
 - Variability:
 - Variance in the 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.

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.

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 the axis.
 - ***The required sample size is indicated at the point at which the resulting line and sample size axis meet.***
 - Can use nomogram to calculate power of a test for a given sample size.
 - Allows retrospective attribution of lack of significance to low sample size.

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 $\bar{p} = \frac{p_1 + p_2}{2}$

$$\frac{\bar{P}_1 - \bar{P}_2}{\sqrt{P(1-p)}}$$

Where $P_1 - P_2$: smallest difference in the proportion of 'success' in the 2 groups

Power Statement

- Needs to be included either as part of study protocol or in methods section of a report.
- Why = to show you considered the sample size during the designing stages of the 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.

Quick Formulae

- For an Independent t-test and the Chi-squared test you can use Lehr's¹ formula for calculating the 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.

Worksheet: Calculating Sample size

What else should we consider?

- Many researchers focus on P value (is there a significant main effect? – Yes, Great! / No, why not?)
- However, we must also consider Power of our test and effect size of any difference when interpreting results.
- Furthermore, there are some things we should consider before we even start collecting any data.
 - E.g., how many Ps we need to ensure that we do detect an observable effect when there is one there.
- We have run through sample size
- Now let's have a look at the other areas of consideration in more detail...

Power and Effect size

- Say you run an analysis and you get a p value of 0.04.
 - This is technically statistically significant (as it's < 0.05).
- Say you ran another analysis on a different set of data and you got a p value of 0.0001 (again statistically significant).
- Can you say that the second result is 'better'/'stronger'?
- **NO!**
- P depends on several factors:
 - The type of test being used
 - The sample size
 - The effect size
 - The Power

Power

- The power of a test is the probability of detecting an effect (if it is there) and thus **NOT** making a Type II error.
- In other words: 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 a small chance that we make a Type II error.
- However: because we can consider the factors that influence our power we can control it.

Power

- Before we begin any study we must ensure a good chance of detecting a significant result (if there is one).
- An acceptable level of Power is considered to be at least .7 (70%) but ideally should aim for .8 (80%) or higher.
- Question: Why would conducting *Applied* research with a 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.

Power

- Factors that have a direct influence on Power.....
 - The sample size:
 - Power increases with a larger sample size.
 - If effect being examined is small then having a larger sample size would increase the power in order 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).

Power

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

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 the 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.

Effect Size

- Sometimes not enough to say that an effect is significant.
- May also want to know the **Size** of the Effect:
 - E.g., by how much does new drug reduce episodes of depression
- The effect size is the **REAL** effect in the population.
- Has the manipulation of an IV created a sizeable difference between our samples on some measure
- The 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.

Effect Size

- E.g., from CARIAD data
 - First want to see whether giving patients the leaflet helped them deal with waiting for their diagnosis compared to the patients who didn't receive the leaflet.
 - Importantly for applied work, also want to see whether that significant difference (if there was one) was sizeable:
 - i.e., the leaflet was very much more successful in helping patients as opposed to simply being just a little bit helpful.
- There are some cases where the effect size might be quite subtle.
- So, what can we do to ensure that a small effect is not overlooked?
 - We can increase the intended sample size.

Effect Size

- The popular measures of effect size to compare groups are Cohen's d and eta squared (η^2)/partial eta squared.
- Psychology typically uses partial eta squared
- SPSS provides this in the last column of inferential statistics box when conducting more complex analyses (such as ANOVA).
 - Remember to tick Display 'Estimates of effect size' in the Options Dialogue box!
- SPSS does not provide effect size statistics for t-tests
- Use information from SPSS to calculate the effect size statistic either by hand or by using online effect size calculators:
 - <http://www.uccs.edu/~lbecker/>

Calculating effect size using Eta-Squared

- Ranges from 0 to 1, represents the proportion of variance in the DV that's explained by the IV (Group).
- Use the values from your SPSS output in the formulas:

- **Independent samples t-test =**

$$\text{Eta Squared} = \frac{t^2}{t^2 + (N1 + N2 - 2)}$$

- **Paired samples t-test =**

$$\text{Eta Squared} = \frac{t^2}{t^2 + (N - 1)}$$

- The Guidelines (proposed by Cohen, 1988 – pp. 284-7) for interpreting the value are:
- .01 = small; .06 = medium, 0.14 = large
- Use your stats books to see how effect size is reported.

Task:

Based on the (hypothetical) CARIAD research dataset provided on Moodle

Calculate the effect size of the comparison between the Control group and the Intervention group on PANAS at time 2.

Confidence Intervals (CI)

- Although distribution from sample can tell us whether differences are significant or not, we can use distribution of sample to calculate Confidence Intervals (CI).
- CI are the best estimate of the range of the population value achieved based on the sample value.
- It is the level of confidence that the population mean will fall between 2 scores/values.
 - E.g., we are 95% confident of the likely range within which a true population mean lies, based on an estimate from a sample mean and standard error (standard deviation of a sampling distribution).
- The mean of our sample can be used as the estimate of the population mean.
- This is known as estimating population parameters from sample statistics.

Confidence Interval of the Mean

- If sample size is large, sample mean follows normal distribution
- We can use normal distribution when considering sample mean
- So, 95% of the distribution of sample means lies within 1.96 standard deviations (SD) of the population mean.
- When we have a single sample, the SD is called the standard error of the mean (SEM) & we calculate the 95% confidence interval of the mean as:
 - (sample mean – (1.96xSEM), sample mean + (1.96xSEM))
- But: the Normal distribution can only be used if we know the value of the variance in the population.
 - Also, if the sample size is small the sample mean only follows a normal distribution if the population are normally distributed.

Confidence Interval of the Mean

Using the t-distribution:

- Calculate the 95% confidence interval for the mean as:
 - (sample mean – ($t_{.05} \times \text{SEM}$), sample mean + ($t_{.05} \times \text{SEM}$)).
- $t_{.05}$ = the percentage point of the t-distribution with (n-1) degrees of freedom and a two-tailed probability.
 - See the Appendix of your 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.

Confidence Interval for the proportion

- The sampling distribution of a proportion follows a Binominal distribution.
- If the sample size is large then the 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:

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

Interpretation of Confidence Intervals

- How wide is it?
 - Width depends on size of the standard error, which depends on the sample size, and the 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 & lower limits provide way of assessing clinical importance
- Are there any values of particular interest?
 - Allows you to check whether a hypothesised value for the population parameter falls within the CI. If so then results are consistent with hypothesised value. If not then it's unlikely the parameter has the value (remember with 95% CI, chance is 5%).

Worksheet: Calculating Confidence Intervals (CI)