Institute for Health Research

THE UNIVERSITY OF NOTRE DAME AUSTRALIA

Institute for Health Research

Biostatistics Lunch
Lecture Series

Fremantle Broome Sydney

nd.edu.au

Institute for Health Research
THE UNIVERSITY OF NOTRE DAME AUSTRALIA
For the health and wellbeing of people and communities in need.

BASIC INFORMATION FOR POWER CALCULATIONS

IHR BIOSTATISTICS LUNCH LECTURE SERIES PRESENTED BY
DR PAOLA CHIVERS
RESEARCH AND BIOSTATISTICS: INSTITUTE FOR HEALTH RESEARCH
THE UNIVERSITY OF NOTRE DAME AUSTRALIA
paola.chivers@nd.edu.au

Research
Fundamental aspect – sample size

Research studies try to find answers to questions

→ Research Question

Since we can't study the whole population, we need to sample.

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Common sample size mistakes

- X Not conducting a power calculation *a priori*
- X Unrealistic assumptions
- X Not accounting for potential drop outs/missing data
- X Selecting sample size based on:
  - Another study
  - Convenience
  - Insufficient funding

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Power Calculations

Power calculation is an estimate
of sample size based on assumptions.

Your best guess!

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

When do I need a Power Calculation?

Any quantitative study
- Except when no data available
- Pilot study

Power is the probability of detecting an effect if it truly exists.

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Why do I need a power calculation?

- Good research practice!
- To determine the sample size (e.g. participants) needed to answer the primary research question (i.e. hypothesis)
- Power calculations help to avoid Type I (false positive) and Type II (false negative) errors.
- Adequate power increases the reproducibility of results
- Requirement for project planning, ethics, grants and publications.

Undersized – can't conclude/find real results!

Oversized – find unimportant results!

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Conducting Power calculations

We recommend you always seek advice from a biostatistician.

A biostatistician is best placed to support you in making sure the right information is used to conduct your power calculation.

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Calculating sample size

1. Main purpose of the study.
   - Primary outcome measure
   - Measurement of outcome measure
2. Need a reasonable knowledge of likely outcome.
3. Estimate of resources available (time, money, recruitment)

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Power balance

- Enough cases to have a good chance to find the unknown truth
- Not to many that we
  - waste resources
  - unnecessary inconvenience
  - unnecessary risk exposure

Ethical practice!

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Main factors affecting a power calculation

- Magnitude of clinically significant differences
- Precision and variance of measurements
- Certainty to avoid Type I and II errors
- Type of statistical test
- One-tailed or two-tailed

Clinically important difference to detect is key!

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

I don't know the variability!

What do you do if you don't know the standard deviation of your outcome measure?

Educated Guess!

Estimate the lowest and highest value standard deviation that you think is plausible and run a power analysis accordingly.

This will provide a sample range upon which you can determine if the study is feasible (i.e. sample recruitment is achievable).

IMPORTANT Reasonable knowledge of the likely outcome!

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Critical

- What is the smallest effect that is clinically important!
- Clinical significance is the practical importance of your treatment/intervention/difference
  - genuine, palpable, change in quality of life, decreased side effects
  - Improved patient management, reduced cost burdens / inconveniences

Type I and II errors and power

Type I error
- False positive
- Convention set alpha <.05
- 5% of occasions of reporting a false positive

Type II error
- False negative
- Convention set Power (1-β) of 0.8-0.9
- If a differences truly exists, then 20% - 10% of occasions will be a false negative

Pitfalls of Power

Pitfalls of underpowering and overpowering a study

	Not a statistically significant difference	Statistically significant difference
Important difference	Underpowered = too small of a sample	
Unimportant difference		Overpowered = too large of a sample

Adapted from Seaman, Seaman and Allen (2015), Table 3. http://asq.org/quality-progress/2015/07/statistics-roundtable/the-significance-of-power.html

Basic information for a power calculation.

1. A main hypothesis to test with null and alternative.
   - Informs your main outcome variable: how is it measured?
   - Informs the type of statistical test to be undertaken
   - Informs the decision between one tailed or two tailed test
2. Significance level (usually .05).
3. Smallest effect size for clinical importance
   - % change between groups, difference between means?
   - Based on evidence from the literature or pilot study
4. Variance of the outcome variable
   - Standard deviation?
   - Based on evidence from the literature or pilot study
5. Power (1-β) of the test (usually 0.8-0.9)

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Basic information for a power calculation.

1. A main hypothesis to test with null and alternative.
   - Informs your main outcome variable: how is it measured?
   - Informs the type of statistical test to be undertaken
   - Informs the decision between one tailed or two tailed test
2. Significance level (usually .05).
3. Smallest effect size for clinical importance (effect size)
   - % change between groups, difference between means?
   - Based on evidence from the literature or pilot study
4. Variance of the outcome variable
   - Standard deviation?
   - Based on evidence from the literature or pilot study
5. Power (1-β) of the test (usually 0.8-0.9)

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Clinical Trial example – scale data

Jones et al 2003

Scenario

- Clinical trial involving hypertensive patients
- New antihypertensive drug Jabba Juice compared to bendrofluazide

? Clinically important difference:
- mean blood pressure of Jabba Juice and
- mean blood pressure of bendrofluazide

? Standard deviation of blood pressure for each

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Clinical Trial example – scale data

Jones et al 2003

What is the null hypothesis?	That Jabba Juice will be no more effective than bendrofluazide in treating new presentations of hypertension.
What level do we want to avoid a type I error at?	We set this to 0.05
What level do we want to avoid a type II error at (1-β) ?	We set this to 0.8
What is the "clinically important difference" we want to detect?	From this study we want to detect a minimum 10mmHg difference between treatments.
What type of data and analysis are likely?	Continuous normally distributed data. To be analysed using a t test.
What is the standard deviation of blood pressure in this group of patients?	From other studies we know that the standard deviation is 20mm Hg.

Adapted Table 2 from Jones et al., 2003

Based on the above assumptions, a two tailed test, a total sample size of 128 patients (64 per group) would be required to test the hypothesis.

Clinical Trial example – categorical data Jones et al 2003

Scenario

- Clinical trial involving myocardial infarction
- New post-thrombolysis drug Jedi Flow

? Clinically important difference:

- Mortality proportion for standard treatment
- Mortality proportion for Jedi Flow

Clinical Trial example

Jones et al 2003

What is the null hypothesis?	That adding Jedi Flow will be no more effective than thrombolysis alone in improving the mortality rate in acute MI.
What level do we want to avoid a type I error at?	We set this to 0.05
What level do we want to avoid a type II error at (1-β) ?	We set this to 0.8
What is the "clinically important difference" we want to detect?	3%
What is the mortality rate using thrombolysis alone?	12%

Adapted Table 4 from Jones et al., 2003

Based on the above assumptions, a two tailed test, a total sample size of 3,000 patients (1,500 per group) would be required to test the hypothesis.

Example – dependent t-test

McCrum-Gardner, 2010

Scenario

- Chlofibrate changes mean cholesterol level
- Pre-post experimental design

? Clinically important difference:

- Pre and post mean cholesterol (or mean difference)
- Standard deviation

Example – dependent t-test

McCrum-Gardner, 2010

What is the null hypothesis?	Clofibrate does not change the cholesterol level.
What level do we want to avoid a type I error at?	We set this to 0.05
What level do we want to avoid a type II error at (1-β) ?	We set this to 0.8
What is the "clinically important difference" we want to detect?	Mean difference of 40mg/dl with standard deviation of 50
What type of data and analysis are likely?	Continuous normally distributed data to be analysed using a dependent (paired) t-test
What is the standard deviation of cholesterol in this group?	From other studies we know that the standard deviation is 50 mg/dl

Based on the above assumptions, a two tailed test, a sample size of 14 patients would be required to test the hypothesis.

More than one outcome measure

Sample estimation should be conducted on your **PRIMARY** outcome measure.

However if you have multiple outcomes of interest you could estimate the sample size for each outcome and power to the largest sample size required.

Example – multiple outcomes

- Based on Wright (2020), using a mean (SD) for game breathlessness 66 (16) and soccer training 51 (16), an assumed correlation between groups of 0.5, alpha of .05 and power (1-β) of 0.8; using a t-test, a sample of 12 participants would be required to compare breathlessness during a game versus during training [Hypothesis 1].
- In addition, based on Wright (2020), using a mean (SD) for game lower limb 61 (16) and soccer training 49 (16), an assumed correlation between groups of 0.5, alpha of .05 and power (1-β) of 0.8; using a t-test, a sample of 16 participants would be required [Hypothesis 2].
- For Hypothesis 3, using correlations based on Wright (2020) for within player associations between differential and global RPE (0.38-0.67), alpha of .05, power (1-β) of 0.8, using correlation, a sample of 11 to 5 participants would be required respectively [Hulley et al., 2013].
- Therefore for this study, a sample of 16 participants would be adequate to answer the proposed hypotheses.

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Example – no known data

A priori power calculation was conducted using PASS 2019 v19.0.4 to estimate the required sample size using power 0.9 and alpha 0.05. The assumptions used for this calculation were based on clinical evidence as no existing information is available to inform the power calculation. These assumptions for one standard deviation were 65% agreement between measures (both pathological 30%, both non-pathological 30%) and 35% disagreement (5%:35%). For one standard deviation, a sample of 55 pairs was determined. For two standard deviations, the assumption of 55% agreement between measures (both pathological 25%, both non-pathological 30%) and 45% disagreement (5%:40%), with a sample of 34 pairs determined. Therefore, for this study it was determined a minimum of 55 clinical patient records with both measures would be analysed.

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

General guidelines for basic power calculations

Statistical Test	What you need (previous studies/pilot) or estimate (educated guess)
Independent T-test	• means of two groups (or difference between the means) • Standard deviations of the two groups
Correlation	• Expected correlation in the population • Accepted correlation for your study
Chi square test	• Positive proportion of both groups
Simple logistic regression	• Probability of positive outcome at the mean predictor variable • Probability of positive outcome at one standard deviation above the mean predictor variable
ICC	• Observations/subject (n) • Expected and acceptable reliability

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Additional study design considerations that affect power

- Repeated measures designs are more powerful than measures at one time point
- Consider number of categories in groups
  - reducing the number of categories increases the sample size per category
  - which group to use as the comparison group
- Drop outs, non-adherence and response rates
- Missing data implications (listwise deletion)

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Sample size calculated is not feasible

For studies where the determined sample cannot be obtained
e.g. Rare diseases.

Inverse the problem

- Calculate the power that can be obtained from the feasible sample
- Evaluate the power against study objectives

Other options:

- Pooling resources, change primary outcome, reduce variability, **don't proceed**.

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Limitation
Power analysis and assumptions

- Based on best case scenario estimate based on assumptions or educated guess.
- The estimate is only as good as your assumptions.
- If assumptions are incorrect then then the power estimated is likely incorrect.
- Does not account for drop-outs, sub group analysis.
- If variability is larger then estimated, serious issue with power.

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Post-hoc Power calculations

Summary: Power Calculations

References and further reading

Grace-Martin, K. 5 reasons to run sample size calculations before collecting data. https://www.theanalysisfactor.com/5-reasons-sample-size-calculations/

Grace-Martin, K. 5 steps for calculating sample size. https://www.theanalysisfactor.com/5-steps-for-calculating-sample-size/

Hickey, G. L., Grant, S. W., Dunning, J., & Siepe, M. (2018). Statistical primer: sample size and power calculations-why, when and how?. European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery, 54(1), 4–9. https://doi.org/10.1093/ejcts/ezy169

Institute for Digital Research & Education Statistical Consulting. Introduction to power analysis. UCLA. https://stats.idre.ucla.edu/other/mult-pkg/seminars/intro-power/

Jones SR, Carley S, Harrison M. An introduction to power and sample size estimationEmergency Medicine Journal 2003; 20:453-458. http://dx.doi.org/10.1136/emj.20.5.453

McCrum-Gardner, E. 2010. Sample size and power calculations made simple. International Journal of Therapy and Rehabilitation; 17(1)10-14. https://www.uv.es/uvetica/files/McCrum_Gardner2010.pdf

Nayak B. K. (2010). Understanding the relevance of sample size calculation. Indian journal of ophthalmology, 58(6), 469–470. https://doi.org/10.4103/0301-4738.71673

Seaman CR, Seaman JE and Allen IE. (2015). The significance of power. Quality Progress. http://asq.org/quality-progress/2015/07/statistics-roundtable/the-significance-of-power.html

Walter, S.D., Eliasziw, M., Donner, A. (1998). Sample size and optimal designs for reliability studies. Statistics in Medicine, 17, 101-110.