Power Analysis and Statistical Analysis Plan - Alessandra Marvelo

Flynn Entwistle[a], **Gisella Sutanto**[a], **and Yoon Hee Cha**[a]

[a]University of Sydney

This version was compiled on April 9, 2024

Alessandra is formulating her ethics application for an upcoming study evaluating the effects of a patient decision aid on their intention to use opioids to ease back pain. The study will be a double blinded randomized controlled trial, constructed analagously to *The impact of a patient decision aid on intention to undergo surgery for Subacromial Pain Syndrome: An online randomised controlled trial* (Zadro et al., 2022). Alessandra has requested us to calculate the sample size necessary to achieve 90% power in detecting a 12% difference of group mean in the primary outcome. Additionally, a prospective statistical analysis plan based on the aims of the study is provided.

Introduction

Alessandra is formulating her ethics application for an upcoming randomised controlled trial (RCT) evaluating the effect of a patient decision aid on their intention to use opioids to ease back pain. Eligible participants in the study are over 18 years old and report lower back pain. The patient decision aid given to the intervention group consists of a brochure outlining the potential consequences of opioid use and recommended alternatives, whilst the control group are given a series of short videos related to back pain, but without mention of opioids.

Primary outcomes. The primary outcome of the study - which will be used for the power analysis - is *Treatment Intention*. It indicates the willingness of participants to use opioids for symptomatic relief of back pain, measured on a Likert scale.

In *Zadro, et al. (2022)*, treatment intention was discretised into a dichotomous outcome for participants recording $\geq 50\%$ or $< 50\%$ on the scale - they used this as their primary outcome, and was the variable they considered when computing their required sample size estimates.

Aims of consultation. We start by reproducing the power analysis in *Zadro, et al. (2022)*, showing their working and providing the same calculation at Alessandra's requested 90% power. We then proceed to conduct the power analysis for *Treatment intention* using the corresponding group standard deviations reported in *Zadro, et al. (2022)*. Supplementary information verifying appropriateness of the two-sample t-test on Likert scale data, and how this may differ depending on the number of points on the scale is also provided. Finally, we provide a plan for statistical analysis modeled off *Zadro, et al. (2022)*.

Power Analysis

We investigate what sample size is sufficient to detect a 12% difference in between control and intervention groups with 90% power, two-tailed $\alpha = 0.05$.

Reproducing power analysis of *Zadro, et al. (2022)*. Zadro, et al. (2022) conducted their power analysis based on *Treatment Intention* discretised into a dichotomous variable with participants recording $\geq 50\%$ or $< 50\%$ on the scale. We explain how they arrived at their sample size calculation to achieve **80%** power, and provide the same calculation at the requested **90%** power in case the client would like to replicate this.

The power analysis for a dichotomous variable consists of a test of proportions, in which effect size d is calculated using the formula:

$$d = \frac{|\Phi^{-1}(p_1) - \Phi^{-1}(p_2)|}{\sqrt{2}}$$

where $\Phi^{-1}(p)$ is the inverse cumulative distribution function of the standard normal distribution, and p_1 and p_2 are the proportions in the two groups.

Using this, *Zadro, et al. (2022)* calculated $n = 198$ as the group sample size necessary to achieve **80%** power. Replicating this at the **90%** power threshold, we recorded $n = 276$ per group would be suitable. As *Zadro, et al. (2022)* recorded a ~3% dropout rate among participants, we recommend adding a 10% buffer - bringing the total requisite sample size per group to $n = 304$.

Power analysis for Treatment Intention. From our initial consultation, Alessandra indicated *Treatment Intention* will be treated as her primary outcome. For a two-sample t-test on numerical data, the effect size (Cohen's d) is calculated using the formula:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Where $\bar{x}_i$ represents the sample mean of group i, and s is the pooled standard deviation of the two groups.

Using the sample standard deviations from *Zadro, et al. (2022)* and theoretical sample means set 12% apart from each other, we calculated that a sample size of $n = 241$ per group would be required for 90% power. Adding 10% to buffer against dropout, we recommend $n = 265$ per group.

Validating statistical assumptions. Being collected on an integer scale, the primary outcome is considered discrete numerical data. To satisfy the normality assumption of the t-test, continuous numerical data is typically needed. However, if the sample size is large enough and/or there are a sufficient number of points on the Likert scale used in the survey, we can empirically prove the normality assumption is met regardless.

Through a simulation study, we found that for $n = 241$ observations per group (consistent with our earlier power analysis) collected on a 15-point Likert scale, the use of discrete numerical data did not lower the power of the t-test below the theoretical computation. The methodology of this experiment is outlined in **Appendix A**.

We also simulate how an empirical estimate of power changes in accordance with the number of points on a Likert scale, across a

range of group sample sizes. This is done to highlight the importance of the scale range in the final survey design. **Appendix B** holds the visualisation and interpretation of this experiment.

Prospective statistical analysis plan

All analyses will be blinded to group status and will be by intention-to-treat. Descriptive statistics will be used to summarise participant demographics, presenting symptoms, current treatment, and outcome data. All analysis conducted will be interpreted at a significance level of $\alpha = 0.05$. The primary outcome will be transformed to a 0-100 scale, with between-group differences being analysed using a student two-sided two-sample T-test. Analysis adjusted for baseline scores and other demographic and clinical characteristics will also be performed using analysis of covariance models with the outcome change from baseline as the dependent variable. A secondary analysis including data from all follow-ups will be conducted using repeated measures with a 15-point Likert scale. The remaining outcome measures will also be regarded as secondary and analysed separately with a logistic regression for dichotomous measures and linear regression analysis for continuous outcome.

do we know that she wants to do a logistic regression?

Statistical analysis guide. ***Should we include this alongside the statement for ethics approval?***

- Generate summary statistics and view the distribution of observations for participant demographics, symptoms, and outcome responses.
- T-tests for comparisons on numerical outcomes, Chi-sq for categorical outcomes: Two-sample t-test (based on the likert scale of 0-15) for control and intervention comparisons. (In reference to the paper, a secondary outcome was made by binning treatment intention into >= and < 50%).
- Perform logistic regression on effect of intervention on intention to try / not try Opioids. Sensitivity analysis on time it took participants to complete the survey.

Bibiography

Zadro, J.R. et al. (2022) 'The impact of a patient decision aid on intention to undergo surgery for Subacromial Pain Syndrome: An online randomised controlled trial', Patient Education and Counseling, 105(9), pp. 2951–2961. doi:10.1016/j.pec.2022.05.005.

Appendix A: Validating t-test assumptions on discrete numerical data

Without continuous numerical data, the statistical power of the t-test may be lower than recorded by the theoretical calculation. Hence, it was necessary for us to validate that the use of discrete numerical data did not lower the power of the t-test beyond the theoretical computation.

To do so, we simulated $n = 241$ observations (as per our power computation in the previous section) from continous normal distributions with a 12% difference in means, and standard deviation consistent with that of Zadro, et al. (2022). 10000 t-tests were conducted from our simulated data. From these t-tests the power was empirically calculated to be ~90% - the same as the theoretical computation.

We then repeated this procedure, but with the observations rounded to the nearest integer to become discrete numerical data. The empirical power computation (~90%) was unchanged by the use of discrete numerical data for the given sample and effect sizes, so we can say with confidence that the use of discrete numerical data does has no bearing on the power reached by our at our computed group sample sizes on a 15-point scale.

Appendix B: Power simulation across Likert scale ranges

The previous simulation assumed that treatment intention was measured on a 15-point Likert scale, as was done in *Zadro, et al. (2022)*. It was shown that, for our calculated sample size, 15 points provided an effect size large enough to achieve 90% power. However, as the number of points on the scale decreases, the effect size shrinks, and the number of samples needed to achieve the same power begins to increase. We provide a simulation to show how the power could be expected to change across a range of sample sizes and scale ranges.

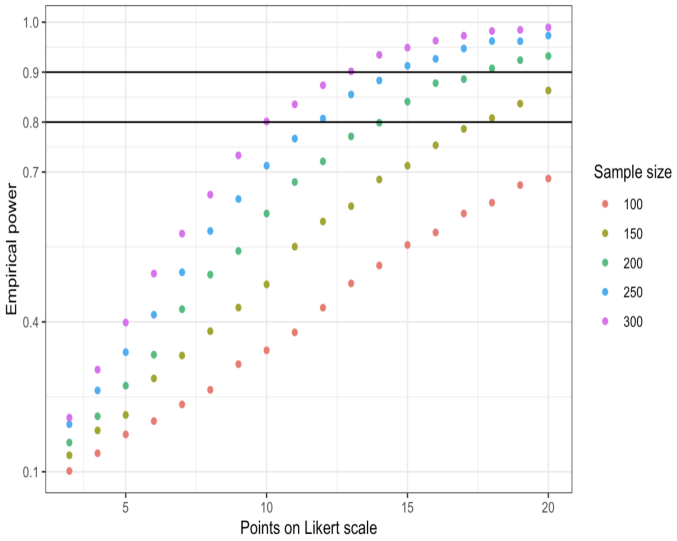

Fig. 1. Statistical power met by sample size and scale range

In figure 1, note that power increases alongside scale and group sample size. Consistent with our earlier sample size calculation, power > 90% for $n = 250$ samples on a 15-point Likert scale. Care should be taken to increase sample size appropriately if you choose to use a scale with less than 15-points.

Note: All scale ranges $\neq 15$ approximate the effect size based on rescaling the standard deviation to the new range of the Likert scale. Graph only shows a 'ballpark' estimate of power at scale sizes $\neq 15$.

Questions

- Inconsistent tense - should we be addressing to 'you', or third personal 'Alessandra' / 'the client'?