SAMPLE SIZE: HOW MANY IS ENOUGH?

Elizabeth Burmeister BN MSc

Nurse Researcher

Nursing Practice Development Unit, Princess Alexandra Hospital and Research Centre for Clinical and Community Practice Innovation, Griffith University

Brisbane, QLD Australia

+ 61 7 3176 7289

Liz_Burmeister@health.qld.gov.au

Leanne M Aitken RN, PhD

Professor of Critical Care Nursing

Research Centre for Clinical and Community Practice Innovation, Griffith University and Princess Alexandra Hospital

Brisbane, QLD Australia

+ 61 7 3176 7256

l.aitken@griffith.edu.au

Keywords: Sample size

Introduction

Sample size is one element of research design that investigators need to consider as they plan their study. Reasons to accurately calculate the required sample size include achieving both a clinically and statistically significant result and ensuring research resources are used efficiently and ethically. Study participants consent to study involvement on the basis that it has the potential to lead to increased knowledge of the concept being studied, however if a study does not include sufficient sample size to answer the question being studied in a valid manner, then enrolling participants may be unethical.

Although sample size is a consideration in qualitative research, the principles that guide the determination of sufficient sample size are different to those that are considered in quantitative research. This paper only examines sample size considerations in quantitative research.

Factors that influence sample sizes

Sufficient sample size is the minimum number of participants required to identify a statistically significant difference if a difference truly exists. Statistical significance does not mean clinical significance. For example, diarrhoea was experienced by patients on 8% fewer days after introduction of a bowel management protocol and was statistically significant[1] but this result may not actually be clinically significant. Before calculating a sample size researchers need to decide what is considered an important or significant clinical difference for their proposed study/question and then calculate the sample size needed to estimate this clinically meaningful difference with statistical precision.

Elements that influence sample size include the effect size, the homogeneity of the sample, the risk of error considered appropriate for the question being studied and the anticipated attrition (loss to follow up) for the study. Considerations related to each of these elements will be discussed.

Effect size is the difference or change expected in your study primary outcome as a result of the intervention being delivered. In order to determine effect size it is essential that the primary outcome being measured is clearly defined. A primary outcome can be collected and measured in a variety of ways, with some examples including physiological data such as blood pressure or heart rate, instrument scores such as quality of life scores or time to event data such as length of stay or survival time.

The sample size calculation should be based on the primary outcome measurement. After a relevant primary outcome measurement has been identified, the expected difference or effect size in that outcome is estimated. Determining an expected difference can be achieved by examining pre-existing data, for example from a previous study or pilot studies or from routinely collected data such as quality audit data. In general, the smaller the anticipated effect size is (i.e. the smaller the difference between groups), the larger the required sample size. For example, if the primary outcome is incidence of delirium and pilot data suggests the intervention is likely to reduce the incidence from 80% to 40%, this will require a smaller sample size than an outcome such as incidence of central line infections where an intervention might be expected to reduce the rate of infection from 5% to 4%.

In considering the effect size of the outcome it is also necessary to ascertain if the study outcome tests will be two-sided or one-sided. Single sided tests are used when the positive (or negative) effects of an intervention are known. For example if an intervention has

previously been tested and proven to reduce the incidence of an outcome when compared to the control group then the difference in that specific direction alone is tested. Two-sided tests are used when the difference in outcomes could be either positive or negative, in other words when an intervention has not been tested previously and the direction (higher or lower) of the difference is not known. Two-sided tests are routinely used in clinical trials as it is essential that either a positive or negative difference or change is detected.

The homogeneity of the sample refers to how similar the participants in the study are to each other and is a reflection of how well the sample reflects the study population. Homogeneity is generally measured using the standard deviation. For example it can be expected that different intensive care units (ICUs) have different patients with different characteristics - higher acuity or longer length of stay(LOS). If a study was using the LOS as an outcome using two different sites then the homogeneity should be examined to ensure the samples from each site do reflect the true population the study is describing.

The risk of error that the researchers consider appropriate must also be contemplated. There are two aspects to consider including the level of significance and the power. The level of significance (referred to as α) defines the strength of identifying an effect when no effect exists, in other words having a false-positive result. A type I error (false-positive) occurs when we wrongly conclude there is a difference, i.e. with an α of 0.05 there is a 5% risk of a false-positive result. The lower the level of α the less likely it is that a type I error will occur. When determining the appropriate level of significance it will be necessary to consider the potential impact of a false-positive result; if the potential impact is serious then a lower level of significance, for example, α= 0.01(1% risk), might be selected.

The power of the study determines the likelihood of not detecting an effect when an effect does actually exist in hypothesis testing studies, in other words having a false-negative result (type II error). A type II error (or false-negative) occurs when we wrongly conclude there is no difference. The higher the power of a study the less likely it is that investigators will fail to detect an effect when an effect does exist. The power of the study is equal to 1-β, where β is the level of acceptability of a false negative result (β =0.20 or 20% is typical, giving a power of 80%), however in a similar fashion to level of significance it may be increased or decreased based on the potential impact of this type of error.

Attrition

Once a required sample size has been calculated it is important to remember that this is the number of participants that are required to complete the study to obtain clinical and statistical differences. It is then essential to consider what the attrition of participants in the study is likely to be, and the sample size should be increased to account for this attrition. Potential attrition varies between study scenarios, for example if all data are collected while a participant is a hospital in-patient and the mortality rate of the patient group is low, then attrition is likely to be low. However, if you are studying a group of patients who have a high mortality rate you will need to take this into account in your sample size. Similarly, if a study involves following patients up post hospital discharge (e.g. at 12 months) a percentage of patients may no longer wish to participate, or they may be difficult to locate. As a result the 12 month sample may be significantly smaller than the baseline sample. As Fernandez et al showed in reporting their study in a cohort of cardiology patients, almost 15% of participants were lost to follow up after 6 weeks[2].

Survey sampling

Sampling methods differ for different types of research. For descriptive surveys, sampling using probability or non probability sampling methods is conducted. The most common type of probability sampling is convenience sampling. Non probability sampling such as random sampling or quota sampling is a more rigorous method of sampling for surveys. As in all research it is essential that survey respondents be a true representation of the study population.

For surveys such as quality of life (QoL) assessment, sampling methods similar to those used in hypothesis testing research are used. However QoL data are often not normally distributed, i.e. the distribution of the data is skewed, therefore sample size calculations need to account for this. This lack of normal distribution is often true for other outcome data, e.g. length of stay. One method of adjusting for a non normal distribution in calculating sample sizes is to transform the outcome variable to a normal distribution for the calculations, for example this may involve using the log or square root of the outcome variable. The same transformation would be used as for the research analyses. Many skewed outcome variables tend toward a normal distribution as the sample size increases, although this may not always be possible to achieve. Another option is to calculate the sample size on a different research outcome that is normally distributed.

Sample size calculation using means

The formula for the sample size required to compare two population means, μ_0 and μ_0, with common variance, σ^2, is:

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{\left(\frac{\mu_0 - \mu_1}{\sigma}\right)^2}$$

For α = 0.05 (level of significance) and β = 0.20 (1- β = power of study) the values of $z_{1-\alpha/2}$ $z_{1-\beta}$ are 1.96 and 0.84, respectively[3]; and $2(z_{1-\alpha/2} + z_{1-\beta})^2$ = 15:68, which can be rounded up to 16, producing the simple formula below:

$$16\ s^2 / d^2 + 1$$

where 'd' is the expected difference between means and 's' is the within-group standard deviation of the individual measurements (indicating the homogeneity of the participants within the groups).[4] For example: if the QOL mean score difference between groups was estimated to be 14 and the within-group standard deviation of the individual QOL scores was 19 then

$$n = 16 \times (19^2/14^2) + 1 = 31 \text{ participants required in each group.}$$

If baseline values of the research are known such as in analysis of covariance (ANCOVA) situations then the difference between the baseline and expected research results is used as the expected difference ‘d’. In addition, for a given effect size, alpha, and power, a larger sample size is required for a two-tailed test than for a one-tailed test.

Different sample size formula are required depending on the research underlying statistical test, for example a t-test for comparing two means, a z-test for comparing two proportions or a log-rank test in time to event analyses. The sample size formula provided in this paper are relevant for tests of comparison between two groups. To calculate sample sizes for studies involving more than two independent groups, the use of Bonferroni’s correction[5] to

amend the alpha level for the sample size calculation would be appropriate. For example, in a study involving examination of the difference in means between four groups, there are six possible comparisons, or t-tests, to be examined. If the overall desired alpha level is 0.05 the sample size formula should be reduced to 0.05 / 6 = 0.0083. In general the formula for more than two groups requires advanced statistical knowledge.

Regression analysis

Similar principles apply when considering an adequate sample size for regression analyses. Multiple regression is used to estimate a relationship between predictors (independent variables) and a continuous dependent variable. Sample size for this type of analysis can use the 20:1 rule[6] which states that the ratio of the sample size to the number of parameters in a regression model should be at least 20 to 1. This rule is appropriate for any regression - dichotomous logistic regression (use the lowest number of events or non events as the effective sample size), survival analysis (number of events), or linear regression (using continuous outcome variable).

The number of parameters to be counted for the sample size calculation should include the number of categories for each variable, in other words if a variable has two potential categories it is counted as two parameters, rather than one. If there are N age categories in your analysis this translates to (N-1) parameters. For example to find the predictors of blood pressure (BP) (dependent variable) with predictors including age group (4 categories) and gender (2 categories) the following would apply:

$$n = ((4 + 2)-1) \times 20 = 100 \text{ participants required in the study.}$$

An alternative method of sample size calculation for multiple regression has been suggested by Green (1991) as:

$$N \geq 50 + 8p$$ where p is the number of predictors[7].

Using the BP study example above and Greens method a sample of $\geq 50 + 8x6 = 98$ participants, therefore a sample of 100 should be sufficient.

These rule of thumb sample size calculations are simple and easy to use but for a truly parsimonious model analysis, and because of the impact of the sample size on the rigour of the research process, it is essential that this aspect of study design is adequately planned, this reinforces the need for a statistician to be part of all research teams.

Conclusion

In summary, a researcher needs to consider the issues related to sample size early in the research planning process. Sample size calculators are available online at no cost for research teams to access and are simple to use. However, if preliminary decisions have not taken place to determine the relevant components of the sample size calculation then a calculator offers no practical advantage. Once the relevant study design and outcome measures have been determined, the other influences on sample size requirements can be specified and include the effect size, homogeneity, risk of error tolerated and the expected attrition.

References

1. Ferrie S, East V. Managing diarrhoea in intensive care. Australian Critical Care (2007) 20, 7-13

2. Fernandez R.S, Davidson P., Griffiths R., Juergens C., Stafford B., Salamonson Y. A pilot randomised controlled trial comparing a health-related lifestyle self-management intervention with standard cardiac rehabilitation following an acute cardiac event: Implications for a larger clinical trial. Australian Critical Care (2009) 22, 17-27

3. Chapter 2: Sample size. http://vanbelle.org/chapters%5Cwebchapter2.pdf. Accessed 30 November 2011.

4. http://www.jerrydallal.com/LHSP/SIZE.HTM. Accessed 10 October 2011.

5. Miller, Rupert G. (1981) Simultaneous statistical inference . 2nd ed. Springer Verlag, pages 6-8

6. Department of Biostatistics, Vanderbilt University; http://biostat.mc.vanderbilt.edu/wiki/Main/ManuscriptChecklist. Accessed 10 October 2011.

7. Green, S.B. How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research,* 1991, *26*, 499-510