Sample Size Planning

Why is sample size planning necessary?

- Good planning is crucial to the success of any research study and sample size planning is an integral part of study planning for many scientific studies.
- A study with too small a sample size is likely to miss effects that are of scientific importance due to a lack of power (ability to correctly reject a false null hypothesis) while a study with a sample size that is larger than what is necessary wastes resources (bigger sample size = more $$$), may lead to small, clinically insignificant, effects being declared statistically significant and also raise ethical issues for some human studies.

Approaches to sample size planning

- Appropriate sample size planning should be conducted in order to avoid these problems.
- There are different ways to approach the problem of sample size planning.
  1. If estimation of a parameter is of primary interest the sample size can be based on specifying the desired width of a confidence interval. For example opinion surveys are often designed to have confidence intervals of +/- 3%.
  2. The more common approach is to calculate a sample size that achieves a certain pre-specified *power* to reject a null hypothesis (in most cases researchers are interested in detecting some sort of effect or difference between groups and so will want to reject the null hypothesis).

Sample size planning for hypothesis testing

- Sample size planning for hypothesis testing involves:
  1. Specifying the study design & statistical method to be used.
  2. Specifying the null and alternative hypotheses.
  3. Specifying the desired power (usually 80% or 90%).
  4. Specify the significance level, α, i.e. the cutoff for determining statistical significance, usually, but not always = .05.
  5. Specifying the expected *effect size*. The exact nature of the effect size depends on the data and statistical method used for the analysis. For example for a two group comparison of a continuous outcome with the t-test the effect size is the expected difference between the group means divided by the standard deviation of the outcome.

Avoid "shortcuts" in sample size planning

- Finding the appropriate effect size is not always straightforward and many researchers use various "shortcuts" to circumvent this difficulty, the most common of which is to use what are commonly referred to as Cohen's tables or Cohen's effect sizes.
- Basically this method involves declaring effect sizes to be either large, medium, or small and then obtaining the sample size necessary to detect that effect size with a specified power.

Avoid "shortcuts" in sample size planning

- This sort of simplification may seem appealing as it appears to get around the difficulties inherent in estimating expected differences between groups and variation in the outcome, however this is misguided.
- The problem with this method is that there is often no justification for expecting "large", "medium" or "small" effect sizes.
- For example according to Cohen's tables for a two-sample t-test to detect a "medium" effect size (defined as a difference between groups of 0.5 standard deviations) requires 65 subjects in each group for 80% power.

Avoid "shortcuts" in sample size planning

- Thus for example if one is interested in comparing SBP between two groups and the standard deviation is estimated to = 20mm then a medium effect size corresponds to a difference = 10 mm.
- Suppose the true difference = 7 mm then the power < 80% and the study with n = 65 per group is underpowered to detect this difference (which nonetheless may be large enough to be scientifically relevant & interesting).
- The point is that the effect size which is clinically relevant and/or realistically can be expected varies from study to study and cannot simply be labeled as "small", "medium" or "large".

Avoid "shortcuts" in sample size planning

- In fact the use of Cohen's tables is not really "sample size planning" at all.
- If we consider the example of calculating sample size for a two-sample t-test once we specify power (say 90%) and the cutoff for statistical significance (.05) the sample size is completely determined by effect size, thus "large" effect size -> "small" sample size, "medium" effect size -> "medium" sample size and "small" effect size -> "large" sample size.
- Thus asking for a sample size to detect a "medium" effect size is exactly equivalent to simply asking for a "medium" sample size!

Power

- In the past the desired 'power' for sample size calculations tended to be set at 80% but recently some researchers have begun to suggest 90% or even 95%.
- There are good reasons for this (lower type II error rate) but one problem that emerges is that the sample size (and therefore the cost) increases as a non-linear function of power, the closer you get to 100% the faster the sample size increases.
- Another issue is that your power is only achieved if your estimate of effect size is actually the true effect size for the population.

If $EffectSize_{pop} < EffectSize_{est}$ then $power_{true} < power_{desired}$

If $EffectSize_{pop} > EffectSize_{est}$ then $power_{true} > power_{desired}$

Type I error - α

- The type I error rate (the cutoff for declaring statistical significance) is usually set = .05.
- However there are situations, for instance if multiple comparisons are to be done, for which a lower cutoff may be desired (e.g. Bonferroni correction).

Two-sided or one-sided test

- Another consideration is whether the hypothesis test is one-sided or two-sided.
- In the vast majority of cases the test will be two-sided even if it is suspected that the difference between groups should go in one direction.
- In general a one sided test is appropriate when a large difference in one direction would lead to the same action as no difference at all.
- One-sided tests should never be used as a way to make a non-significant result significant (e.g. two-sided p = 0.08).

Effect Size

- Estimating an appropriate effect size is the hardest part of doing a sample size calculation.
- Effect size for comparisons of means requires both an estimated mean difference between the groups and an estimated variance(s) or standard deviation(s) of the outcome of interest.
- Effect size for comparison of proportions requires an estimate of the two proportions, as the variance of the difference in proportions is based on the proportions themselves.

Estimating Effect Size

- A common practice is to base the estimate of effect size on previous studies in the literature.
- For example a previous study may have found a difference in means = 10 with a common standard deviation for the populations = 15 which would require a sample size = 48 per group to obtain a power of 90% to detect a difference given an alpha level = 0.05.
- However just because a previous study with the same groups and outcomes found this difference does not necessarily mean you can expect the same in your study.

Estimating Effect Size from Prior Studies

- What are some reasons why the expected population difference in your study may be different from a previous study from the literature?
- What if there have been several previous studies?

Estimating Effect Size from Pilot Studies

- Another popular option, especially if there are no prior relevant studies, is to use the data from a pilot study to estimate the sample size.
- In general doing a pilot study is a good idea for many reasons but caution needs to be exercised when using the effect size observed in the pilot study as the estimated effect size for the main study.
- The main reason for caution is that the mean difference and SD estimated from the pilot study are also subject to sampling error and this error can be quite large if the sample size for the pilot is small.
- It's a good idea to also calculate a confidence interval for the mean difference observed to see how precise your effect size estimate is.
- If for example you have a mean difference = 20 with a 95% CI = (5, 35) it may be a good idea to use a mean difference < 20 in the calculation to be conservative.

Estimating Effect Size using 'smallest clinically meaningful difference'

- A good alternative to the above approaches is to estimate the effect size you wish to detect based on the concept of wanting sufficient power to detect a 'clinically meaningful' difference.
- The smallest clinically meaningful difference is just the smallest difference between means that would have some sort of clinical significance – for example if a researcher were comparing medications to lower blood pressure, would a difference of 1 mmHg between treatments make a clinical difference, or 5 mmHg or 10 mmHg?
- As sample size rises rapidly with smaller effect size (halving effect size -> quadrupling sample size) it is important not to specify a sample size to detect arbitrarily small and essentially meaningless differences.

Sample Size for Regression Models

- **Sample sizes for regression models are generally based on two approaches:**

1. **Rules of thumb involving the number of independent (predictor) variables to be included in the models.**
2. **Sample sizes based on effect sizes (betas or correlations for linear regression, adjusted odds ratios for logistic regression, hazard ratios for Cox regression).**

Sample Size for Linear Regression: Rules of Thumb

- **There have been many rules of thumb proposed for multiple linear regression based on the number of predictors including:**
  1. **N >= 10*m**
  2. **N > 50 + m**
  3. **N >= 20*m**
  4. **N >=50 + 8*m**
- **S Green (Multivariate Behavioral Research, 1 July 1991) found that (4) seemed to work best but argued that effect size should also be taken into account.**

Sample Size for Logistic Regression: Rules of Thumb

- **The power for logistic regression analysis is based not only on the total sample size but also on the balance between the outcomes – for example a study with 200 subjects, 100 with the event and 100 without the event has more power than a study with 1000 subjects, 10 with the event and 990 without the event.**
- **Therefore when planning the sample size the expected proportion with and without the event of interest also needs to be estimated.**
- **A commonly used rule of thumb is**:

  Number of observations in smaller outcome group >= 10*number of predictors.
- **Thus for example if you had 10 predictors and the expected proportion of events was 60% you would need how many observations?**

Sample Size for Logistic Regression: Rules of Thumb

- **A commonly used rule of thumb is**:

  Number of observations in smaller outcome group >= 10*number of predictors.
- **Thus for example if you had 10 predictors and the expected proportion of events was 60% you would need how many observations?**

Sample Size for Logistic Regression: Rules of Thumb

- A recent paper [Vittinghoff & McCulloch, Am J Epidemiology, 2007] stated that this rule of thumb may be too conservative for studies for which confounder control, rather than prediction, was the main motivation for using logistic regression and that smaller sample sizes may be possible depending on the effect size for the main predictor of interest.
- Another study [FE Harrell et al, Statistics in Medicine, 1996] stated that up to 20 'events per variable' may be necessary for prediction models.

Sample Size for Cox PH Regression: Rules of Thumb

- Sample size rules of thumb for Cox proportional hazards regression models are similar to those for logistic regression except that rather than being based on the number in the smallest outcome group they are based on the number of events (i.e. non-censored observations).
- Thus a study in which very few are censored has more power than those in which the percentage censored is high – thus the expected proportion of observed events needs to be taken into account.

Sample Size for Regression Models based on effect sizes

- Basing the regression sample sizes on effect size is particularly useful when the models are used for confounding and the focus of the analysis is on one predictor variable.
- Such calculations can be done by using the NCSS-PASS software.
- For linear regression the effect size is measured by the correlation, for logistic regression by the odds ratio but the software also takes into account the amount of control for other variables, i.e. the R^2 for the other (confounding) variables in the model.
- The more variability that is controlled by the other variables the smaller the required sample size.

Sample Size for Clustered Data

- Sample sizes for clustered data including data collected using clinical trials and cluster randomization will generally need to be larger than those for non-clustered data.
- This is due to the fact that observations in a single cluster (e.g. patients in a clinic, students in a school) will not be independent of one another and the correlation between them will reduce the power thereby making a larger sample size necessary).
- The ratio of the sample size required for cluster randomization to that required for simple randomization is called the 'design effect' and depends on how correlated observations are within each cluster [see reference for details].

Other Considerations regarding Sample Size

- **Post-hoc power calculation** – this is a power calculation done after the study is completed based on the observed effect size in the study.
- For example if you completed your study and observed a mean difference = 5 and an SD = 10 with a sample size of 200, you would calculate the power based on these numbers.
- Most statisticians regard post-hoc power calculations as useless –
  - They give no extra information – if the p-value is small the post-hoc power will be high, if the p-value is large the post-hoc power will be small.
  - Once the study is over the power calculation really has not use, if you obtained a significant result you had enough power (or you made a type I error), if the result is not significant either you did not have enough power or the null hypothesis is true.
  - Unfortunately reviewers seem to commonly request post-hoc power so you may have no choice.

Other Considerations regarding Sample Size

- What if the sample size is fixed by other considerations?
  - Amount of money available for the study.
  - For rare conditions may be limited by total number of patients available in the clinic, hospital or even country.

• In this case power can be calculated for the specified effect size.
• If the power is then very low this can be used to argue for a larger budget, or perhaps that the particular study may not be useful.
• Alternatively the power and sample size can be fixed and the effect size that can be detected with a given power is then calculated.
• If the effect size is reasonable then the study may proceed.

Sample Size Software

- **NCSS-PASS** – very comprehensive and reasonably user friendly, but expensive $US 750 for a single user license. 7-day free trial version available.
- **SISA** – free online program can do simple calculations for two group comparisons of means or proportions. http://www.quantitativeskills.com/sisa/index.htm
- **G-power** – free downloadable program. Less user-friendly and comprehensive than PASS but more comprehensive than most online programs. http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/
- **Java Applets for Sample Size:** http://www.cs.uiowa.edu/~rlenth/Power/
- **CUHK CCT:** Survival Analysis sample size http://www.cct.cuhk.edu.hk/stat/index.htm

References

1. RV Lenth, Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician*, Vol. 55, No. 3 (Aug., 2001), pp. 187-193.
2. FE Harrell et al, Multivariate prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996: 15:361-87.
3. S Green, How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research 1991;* 26:3 499-510.
4. SM Kerry, JM Bland, Sample size in cluster randomization. *BMJ* 1998; 316:549.
5. E Vittinghoff, CE McCulloch, Relaxing the rule of 10 events per variable in logistic and Cox regression. *American Journal of Epidemiology 2006;* 165(6): 710-8.