Sample Size Planning, Calculation, and Justification

Theresa A Scott, MS

Vanderbilt University
Department of Biostatistics
theresa.scott@vanderbilt.edu
http://biostat.mc.vanderbilt.edu/TheresaScott

Introduction

▷ After you've decided what and whom you're going to study and the design to be used, you must decide *how many 'subjects' to sample*.

- Even the most rigorously executed study may fail to answer its research question if the sample size is *too small*.
- If the sample size is *too large*, the study will be more difficult and costly than necessary while unnecessarily exposing a number of 'subjects' to possible harm.

▷ *Goal*: to estimate an *appropriate number* of 'subjects' for a given study design.

- ie, the number needed to find the results you're looking for.

▷ IMPORTANT: Although a useful guide, sample size calculations give a deceptive impression of statistical objectivity.

- Really only making a ballpark estimate.

Introduction, *cont'd*

▷ Only as accurate as the data and estimates on which they are based, which are often just informed guesses.
▷ Often reveals that the research design is not feasible or that different predictor or outcome variables are needed.
▷ TAKE HOME MESSAGE: Sample size should be estimated *early* in the *design* phase of the study, when major changes are still possible.

▷ In addition to the statistical analysis plan, the sample size section is critical to an IRB proposal and any kind of grant.

- 42% of R01s examined in one review paper were criticized for the sample size justifications or analysis plans.[1]
- Much more involved than a cut-and-paste paragraph.

[1] Inouye & Fiellin, *"An Evidence-Based Guide to Writing Grant Proposals for Clinical Research"*, Annals of Internal Medicine, 142.4 (2005): 274-282.

Underlying principles

▷ **Research hypothesis**:

- Specific version of the research question that summarizes the main elements of the study – the sample, and the predictor and outcome variables – in a form that establishes the basis for the statistical hypothesis tests.[2]
- Should be *simple* (ie, contain *one* predictor and *one* outcome variable); *specific* (ie, leave no ambiguity about the subjects and variables or about how the statistical hypothesis will be applied); and *stated in advance*.
- Example: Use of tricyclic antidepressant medications, assessed with pharmacy records, is more common in patients hospitalized with an admission of myocardial infarction at Longview Hospital in the past year than in controls hospitalized for pneumonia.

[2] NOTE: Hypotheses are not needed for descriptive studies – more to come.

Underlying principles, *cont'd*

- **Null hypothesis**:
  - Formal basis for testing statistical significance;[3] states that there is no association, difference, or effect.
  - eg, Alcohol consumption (in mg/day) is *not* associated with a risk of proteinuria (>300 mg/day) in patients with diabetes.
- **Alternative hypothesis**:
  - Proposition of an association, difference, effect.
  - Can be *one-sided* (ie, specifies a direction).
    - eg, Alcohol consumption is associated with an *increased* risk of proteinuria in patients with diabetes.
  - However, most often *two-sided* – no direction mentioned.
    - Expected by most reviewers; very critical of a one-sided.

[3] Hypothesis testing discussed in more detail in the 'Biostatistics: Types of Data Analysis' lecture.

Underlying principles, *cont'd*

- General process used in **hypothesis testing**:
  - Presume the null hypothesis (eg, no association between the predictor and outcome variables in the population).
  - Based on the data collected in the sample, use statistical tests to determine whether there is sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis (eg, there is an association in the population).
- Reaching a wrong conclusion:
  - **Type I error**: false-positive; rejecting the null hypothesis that is actually true in the population.
  - **Type II error**: false-negative; failing to reject the null hypothesis that is actually not true in the population.
  - Neither can be avoided entirely.

Underlying principles, *cont'd*

▷ **Effect size**:

- Size of the association/difference/effect you expect/wish to be present in the sample.
- Selecting an appropriate size is most difficult aspect of sample size planning.
- REMEMBER: Sample size calculation only as accurate as the data/estimates on which they are based.
  - Find data from prior studies to make an informed guess – needs to be as similar as possible to what you expect to see in your study.
  - Pilot study/studies sometimes needed first.
- *Good rule of thumb*: choose the smallest effect size that would be *clinically* meaningful (and you would hate to miss).
  - Will be okay if true effect size ends up being larger.

Underlying principles, *cont'd*

▷ Establish the maximum chance that you will tolerate of making wrong conclusions:

- α: probability of committing a type I error; aka 'level of statistical significance'.
- β: probability of making a type II error.
- **Power**: $1 - \beta$; probability of *correctly* rejecting the null hypothesis in the sample if the actual effect in the population is equal to (or greater than) the effect size.
- Aim: choose a sufficient number of 'subjects' to keep α and β at an acceptably low level without making the study unnecessarily large (ie, expensive or difficult).
  - α and β decrease as sample size increases.
- Often $\alpha = 0.05, 0.10$; $\beta = 0.20, 0.10$ (power $= 80, 90\%$).

Additional considerations

▷ *Variability of the effect size*:

- Statistical tests depend on being able to show a difference between the groups being compared.
- The greater the variability (spread) in the outcome variable among the subjects, the more likely it is that the values in the groups will overlap, and the more difficult it will be to demonstrate an overall difference between them.
- Use the most *precise* measurements/variables possible.

▷ Often have >1 hypothesis, but should specify a single *primary* hypothesis for sample size planning.

- Helps to focus the study on its main objective and provides a clear basis for the main sample size calculation.
- Useful to *rank* other research questions/specific aims as secondary, etc.

Calculating sample size

▷ Specific method used depends on

- The specific aim(s)/objective(s).
- The study design, including the planned number of measurements per 'subject'.
- The outcome(s) and predictor(s).
- The proposed statistical analysis plan.

▷ Will also need to consider:

- Accrual/Enrollment (response rate for questionnaires).
- Drop-outs (ie, lost to follow-up) and missing data.
- Budgetary constraints.

▷ Requires you to make *assumptions*.

- Assume specific effect size (variability), α, power, etc.

Calculating sample size for **analytic** studies

▷ Often want to show a significant difference/association between 2 groups.

▷ Most traditional 'recipe' in this case:

1. State the null and 1- or 2-sided alternative hypothesis.
2. Select the appropriate statistical test based on the type of predictor and outcome variables in the hypotheses.
3. Choose a reasonable effect size (and variability, if necessary).
4. Specify α and power.
5. Use an appropriate table, formula, or software program to estimate the sample size.

▷ Most common used statistical tests for comparing 2 groups:

- The *t*-test.
- The *Chi-squared* test.

Calculating sample size for **analytic** studies, *cont'd*

▷ The *t*-test:

- Commonly used to determine whether the mean value of a *continuous outcome variable* in one group differs significantly from that in another group.
  - Assumes that the distribution of the variables in each of the 2 groups is approximately normal (bell-shaped).
- Assumptions:
  - Whether the 2 groups are paired or independent.
    - Example 'paired': Comparing the mean BMI of 'subjects' before and after a weight loss program.
    - Example 'independent': Comparing the mean depression score in 'subjects' treated with 2 different antidepressants.
  - The mean value of the variable in each group.
    - Calculation actually uses the 'effect size' (the *difference* in the mean values between the 2 groups).

Calculating sample size for **analytic** studies, *cont'd*

▷ The *t*-test, *cont'd*:

- Assumptions, *cont'd*:
  - The standard deviation (SD) of the variable.
    - If 2 groups are paired: SD of the *difference* in the variable between matched pairs.
    - If 2 groups are independent: SD of the variable itself.
- Rules of thumb:
  - Smaller sample size needed for paired groups – SD of the difference in a variable usually smaller than the SD of a variable.
  - Sample size decreases as the difference in the mean values increases (holding SD constant).
  - Sample size increases as SD increases (holding the difference in the mean values constant).
- Also have *standardized effect size* $= \frac{effectsize}{SD}$.
  - Sample size decreases as standardized effect size increases.

Calculating sample size for **analytic** studies, *cont'd*

▷ Example using the *t*-test:

- *Research question*: Is there a difference in the efficacy of salbutamol and ipratropium bromide for the treatment of asthma?
- *Planned study*: randomized trial of the effect of these drugs on FEV_1 (forced expiratory volume in 1 second) after 2 weeks of treatment.
- *Previous data*: mean FEV_1 in persons with asthma treated with ipratropium was 2.0 liters, with a SD of 1.0 liter.
- *Wish*: to be able to detect a difference of $\geq$10% in mean FEV_1 between the 2 treatment groups.
- *Assumptions*: α (two-sided) $= 0.05$; power $= 0.80$; effect size $= 0.2$ liters (10% *X* 2.0 liters); SD $= 1.0$ liter.
- *Calculation*: A sample size of 394 patients *per group* is needed to detect a difference of $\geq$10% in mean FEV_1 between the 2 (independent) treatment groups with 80% power, using a two-sample t-test and assuming a (two-sided) α of 0.05, a mean FEV_1 of 2.0 liters in the ipratropium group, and a SD of 1.0 liter.

Calculating sample size for **analytic** studies, *cont'd*

▷ The *Chi-square*-test:

- Commonly used to determine whether the proportion of 'subjects' who have a *binary outcome variable* in one group differs significantly from that in another group.
- Assumptions:
  - Whether the 2 groups are matched/paired or independent.
  - The proportion with the outcome in each group.
- Rule of thumb: both the value of the proportions and the difference between them affect sample size – sample size much larger for small proportions.
- Can also calculate assuming a *relative risk* (instead of 2 proportions).
  - IMPORTANT: a relative risk (ie, risk ratio) equals an *odds ratio* in only certain cases.

Calculating sample size for **analytic** studies, *cont'd*

▷ Example using the (uncorrected) *Chi-square*-test:

- *Research question*: Is there a difference in the incidence of skin cancer between elderly smokers and non-smokers?
- *Previous data*: the 5-year incidence of skin cancer is about 0.20 in elderly non-smokers.
- *Wish*: to determine that the 5-year skin cancer incidence is $\geq$0.30 in elderly smokers.
- *Assumptions*: α (two-sided) = 0.05; power = 0.80; P_1 (incidence in smokers) = 0.30; P_2 (incidence in non-smokers) = 0.20.
- *Calculation*: A sample size of 293 elderly smokers and 293 elderly non-smokers is needed to determine that the 5-year skin cancer incidence is $\geq$0.30 in elderly smokers with 80% power, using an (uncorrected) Chi-square test and assuming a (two-sided) α of 0.05 and a 5-year skin cancer incidence of 0.20 in elderly smokers.

Calculating sample size for **analytic** studies, *cont'd*

▷ Alternative recipes – useful when sample size is fixed or the number of subjects who are available or affordable is limited:

- Calculate the power for a given effect and sample size.
- Calculate the effect size for a given power and sample size.
  - Good general rule: always assume $\geq$80% power.

▷ Can also calculate the sample size required to determine whether the *correlation coefficient* between 2 continuous variables is significant (ie, significantly differs from 0) – see 'Designing Clinical Research'.

▷ *'Adjusting for covariates'*: when designing a study, you may decide that $\geq$1 variable will confound the association between the predictor and outcome, and plan to use *regression analysis* to adjust for these confounders.

- Calculated sample size will need to be *increased* ('10:1 rule') – see a statistician.

Calculating sample size for **descriptive** studies

▷ Usually do not involve hypotheses; goal is to calculate *descriptive statistics* (eg, means and proportions).

▷ *Approach*: calculate sample size required to estimate a *confidence interval* (CI) of a specified *confidence level* (eg, 95%) and *total width* (ie, precision).

- For a *continuous* variable: interested in the CI around the mean value of that variable.
- For a *binary* variable: interested in the CI around the estimated *proportion* of 'subjects' with one of the values.

▷ Rules of thumb:

- A larger sample size is needed for a 'tighter' (ie, smaller total width; more precise) CI of any confidence level.
- For a given total width, a larger sample size is needed for a CI with a higher confidence level.

Sample size for **descriptive** studies, *cont'd*

▷ *Assumptions* made for calculating the sample size:

- For a *continuous* variable: the standard deviation of the variable.
- For a *binary* variable: the expected proportion with the variable of interest in the population.
  - If more than half of the population is expected to have the characteristic, then plan the sample size based on the proportion of expected *not to have* the characteristic.
- For *both*:
  - (1) the desired precision (total width) of the CI.
  - (2) the confidence level for the interval (eg, 95 or 99%).

▷ Example: A sample size of 166 'subjects' is needed to estimate the mean IQ among 3rd graders in an urban area with a 99% CI of ± 3 points (ie, a total with of 6 points), assuming a standard deviation of 15 points.

Sample size for **descriptive** studies, *cont'd*

▷ IMPORTANT: descriptive studies of binary variables include studies of the *sensitivity* and *specificity* of a *diagnostic test*.

- Goal of a diagnostic test (or procedure): to correctly classify 'subjects' as having or not having a 'disease' (or symptom).
- Sensitivity: proportion of *true positives*; probability of testing positive when you truly have the 'disease'.
- Specificity: proportion of *true negatives*; probability of testing negative when you truly do not have the 'disease'.
- Example: A sample size of 246 'subjects' is needed to estimate a 95% CI for the sensitivity of a new diagnostic test for pancreatic cancer, assuming $80 \pm 5\%$ of the patients with pancreatic cancer will have positive tests.
  - When studying the *specificity* of a test, must estimate the required number of 'subjects' who *do not* have the 'disease'.

Sample Size Software Programs

▷ PS: Power and Sample Size Calculation

- *Free* from `http://biostat.mc.vanderbilt.edu/PowerSampleSize`.
- Available for Windows and can run on Linux using Wine.
- Handles several common analyses including *t*-test, Chi-square test ('Dichotomous'), and Survival.
- Can generate *graphs* (eg, power vs sample size) and keeps *'log'*.

▷ nQuery Advisor

- *Not* free; handles complex calculations including >2 groups (ie, ANOVA), non-parametric tests, and cross-over designs.
- Has a 'statement' button, can generate graphs, and keeps log.

▷ *Simulations* can also be done for any statistical technique.

- Most valuable for complex analyses, such as mixed effects or GEE models – statistician (most likely) needed.

Additional thoughts/considerations

▷ Consider *strategies for minimizing sample size and maximizing power*, which include using

- Continuous variables,
- Paired measurements,
- Unequal group sizes, and
- A more common (ie, prevalent) binary outcome.

▷ Useful to calculate (and report) a range of sample sizes by assuming different combinations of parameter values – take the largest sample size to 'cover all bases'.

▷ Always *justify* the feasibility of the calculated sample size.

- How long would it take to accrue/enroll the subjects?
- Need to consider the source of subjects, the inclusion/exclusion criteria, the prevalence of the outcome, etc.

What do I include in my sample size write-up?

▷ Key: state all the information *assumed* such that anyone reading your 'Sample Size section' would be able to re-calculate the sample size given. This includes

- The (primary) specific aim.
- The outcome and predictor variable(s).
- Primary comparison of interest (if applicable).
- Parameter estimates (ie, α, power, effect size, variability, etc).
- Data on which you based your assumptions.
- Statistical test used (if applicable).

▷ Including graphs can be very helpful.
▷ Show the reviewers that you have solid reasoning behind your calculations (as well as your statistical analysis plan).

- Acknowledge whether study will be a pilot or feasibility study.

Take home message

▷ Sample size justification is an essential part of every research study, and thus any IRB proposal or grant application.

▷ Calculating the appropriate sample size is rarely as simple as plugging numbers into a formula.

▷ Sample size should be estimated *early* in the *design* phase of the study, when major changes are still possible.

- Involve a statistician from the beginning – you're more likely to be funded/approved if you do so.

▷ References & acknowledgments:

- *Designing Clinical Research* (3rd edition) by Hulley, et al.
- Ayumi Shintani & Jennifer Thompson.