Sample Size Planning
for Behavioral Science Research

Douglas G. Bonett

Center for Statistical Analysis in the Social Sciences

University of California, Santa Cruz

Revised 3/15/2014

Overview

- Importance of sample size planning in behavioral research
- Sample size requirements for desired precision
- Sample size requirements for desired power
- Precision for a specified sample size
- Power for a specified sample size
- Sampling in two stages
- Examples
- Learning materials on CSASS website (lecture notes, study guide, R functions, SAS programs)

Importance of sample size planning

Sample size planning is especially important in studies where statistical methods will be used to analyze the data and there are tangible costs of recruiting, measuring, or treating participants.

If the sample size is too small, statistical tests may not detect important effects (low power), and effect size confidence intervals will be uselessly wide.

Using a sample size that is unnecessarily large is wasteful of a valuable and finite human resource. A study that uses too many participants could reduce the number of participants that are available to other researchers.

Importance of sample size planning (*continued*)

Funding agencies usually require a justification of the proposed sample size. An increasing number of journals now require authors to provide a sample size justification as explained in the following journal policy statements.

“State how the intended sample size was determined.” (*APA Publication Manual)*

“The Method section should make clear what criteria were used to determine the sample size.” (*New Statistical Guidelines for Journals of the Psychonomic Society*).

“Authors should indicate how the sample size was determined.” *(Consolidated Standards of Reporting Trails).*

Why do journals want sample size justification?

Several studies have shown that most published behavioral science articles should not have found “significant” results because the sample sizes were too small to reliably detect the reported effect sizes. This suggests that the reported effect sizes were inflated due to random sampling error because inflated effect sizes are more likely to give *p*-values below the .05 level. Sample size planning should reduce the number of underpowered studies, which in turn should reduce the positive bias in reported effect sizes.

Behavioral science publications seldom provide an adequate description of the meaning and importance of reported effect sizes and confidence intervals. Authors who provide a sample size justification will naturally need to explain why the expected effect size should have practical or theoretical importance.

Illustration of effect size inflation

Suppose the true *d* value (standardized mean difference) is equal to 0. The following table gives the expected absolute *d* value in studies where *p* < .05.

Sample size per group	Average |*d*|
10	1.15
20	0.78
30	0.62
40	0.53
80	0.37

These results illustrate how reported effect size can be highly inflated in under-powered studies.

Sample size for desired precision

A $100(1-\alpha)\%$ confidence interval for μ is

$$L = \hat{\mu} - t_{\alpha/2;df}\sqrt{\hat{\sigma}^2/n} \qquad U = \hat{\mu} + t_{\alpha/2;df}\sqrt{\hat{\sigma}^2/n}$$

where $df = n - 1$, $t_{\alpha/2;df}$ is a two-sided critical t-value, and y_i is a quantitative measurement of some attribute for participant i.

The width (w) of this confidence interval is $U - L = 2t_{\alpha/2;df}\sqrt{\hat{\sigma}^2/n}$ and solving for n gives

$$n = 4\hat{\sigma}^2(t_{\alpha/2;df}/w)^2.$$

Sample size for desired precision *(continued)*

Prior to conducting the study, we will not have the estimate of σ^2 and $\hat{\sigma}^2$ must be replaced with a *planning value* of σ^2, denoted as $\tilde{\sigma}^2$.

Planning values of σ^2 can be obtained from pilot studies or prior research.

Since $df = n - 1$ and n is unknown, $t_{\alpha/2;df}$ is unknown but can be approximated by $z_{\alpha/2}$, a two-sided critical z-value. With these two substitutions, we obtain

$$n = 4\tilde{\sigma}^2(z_{\alpha/2}/w)^2.$$

Sample size for desired precision *(continued)*

An examination of this sample size formula

$$n = 4\tilde{\sigma}^2 (z_{\alpha/2}/w)^2$$

shows that larger sample sizes are needed for:

- higher levels of confidence
- narrower confidence intervals
- larger variances

Example

We want to estimate the mean job satisfaction score for a population of 4,782 public school teachers, and we will use a job satisfaction questionnaire (measured on a 1 to 10 scale) that has been used in previous studies. A review of the literature suggests that the variance of the job satisfaction scale is about 6.0. We want the 95% confidence interval for μ (the mean job satisfaction score for all 4,782 teachers) to have a width of about 1.5. The required sample size is approximately

$$n = 4(6.0)(1.96/1.5)^2 \approx 41.$$

Sample size for desired power

The one-sample t-test is a test of H_0: $\mu = h$. The *power* of this test is the probability of rejecting H_0.

The required sample size to test H_0: $\mu = h$ with a specified α value and with desired power is approximately

$$n = \tilde{\sigma}^2(z_{\alpha/2} + z_{\beta})^2/(\tilde{\mu} - h)^2$$

where β = 1 – power and $\tilde{\mu}$ is a planning value of μ.

Sample size for desired power *(continued)*

An examination of this sample size formula

$$n = \tilde{\sigma}^2(z_{\alpha/2} + z_\beta)^2/(\tilde{\mu} - h)^2$$

shows that larger sample sizes are needed for:

- higher levels of power (smaller β)
- smaller α values
- larger variances
- smaller absolute values of $\tilde{\mu} - h$ (smaller effect size)

Example

The EPA estimates that lead in drinking water is responsible for more than 500,000 new cases of learning disabilities in children each year. Lead contaminated drinking water is most prevalent in homes built before 1970. The legal lead concentration limit for drinking water is 15 ppb. We want to obtain water samples from pre-1970 apartment buildings in San Francisco and test H_0: $\mu = 15$ with α = .05 such that the power of the test is about .9. Results from a similar study in Los Angeles were used to set the planning value of the mean lead concentration to 20 ppb with a variance of 250. The number of San Francisco apartments to sample is approximately

$$n = 250(1.96 + 1.28)^2/(15 - 20)^2 \approx 105.$$

Power and precision for a specified sample size

In some studies, the researcher will have access to, or will only have the resources to, take a sample of some specified size.

In these studies, it informative approximate the anticipated width of a planned confidence interval or the power of a planned test.

If the confidence interval with is too large, or if the power is too small, the researcher might decide to abandon the proposed study or attempt to obtain a larger sample size.

Power for a specified sample size

The power of one-sample t-test H_0: $\mu = h$ in a sample of size n is approximately equal to the area under a standard unit normal distribution that is to the left of

$$z = |\tilde{\mu} - h|/\sqrt{\tilde{\sigma}^2/n} \; - \; t_{\alpha/2;df}$$

where $df = n - 1$. This area is easily computed using the **pnorm** function in R, and the **qt** function in R is useful for finding $t_{\alpha/2;df}$.

Precision for a specified sample size

The anticipated width of a confidence interval for μ in a sample of size n is

$$w = 2t_{\alpha/2;df}\sqrt{\tilde{\sigma}^2/n}$$

where $df = n - 1$.

Example

We plan to measure ventromedial prefrontal cortex brain activity (an area associated with response to reward) using fMRI in a sample of n = 25 pathological gamblers. Based on research from previous studies of non-gamblers, we set h = 45 (the mean brain activity score for non-gamblers observed in previous studies) and $\tilde{\sigma}^2$ = 100. The researcher expects $\tilde{\mu}$ = 50 for gamblers and will use α = .05.

$$z = |50 - 45|/\sqrt{100/25} - 2.06 = 0.44$$

The estimated power is pnorm(0.44) = .67.

Two-stage sampling for desired precision

Most confidence interval sample size formulas require planning values. In applications where planning values are difficult to specify but sampling can be performed in two stages, a confidence interval can be computed in a small sample of size n_0 and its width w_o determined. The number of additional participants (n^+) to sample in the second stage is

$$n^+ = n_0\left[\left(\frac{w_0}{w}\right)^2 - 1\right]$$

where w is the desired width.

Example

The sample size needed to estimate a slope coefficient in a multiple linear regression with desired precision requires a planning value for the multiple correlation between predictor variable *j* and all other predictor variables. This value is difficulty to specify. A first-stage sample of n = 50 is obtained, and the width of a 95% confidence interval for the slope coefficient of primary interest was 47.3. We would like the width of the confidence interval for this slope coefficient to be about 30. The number of additional participants to sample is

$$50[(47.3/30)^2 - 1] \approx 75$$

for a final sample size of n = 125.

Example: Two-group analysis of means

We want to compare the performance of 1-person and 3-person teams on a particular type of writing task that must be completed in 30 minutes. The quality of the written report will be scored on a 1 to 10 scale. We set $\tilde{\sigma}^2$ = 5.0 and expect a 2-point difference in the population mean ratings. For α = .05 and power of 1 – β = .95, the required number of teams per group is approximately

$$n_j = 2\tilde{\sigma}^2\left(z_{\alpha/2} + z_\beta\right)^2/(\tilde{\mu}_1 - \tilde{\mu}_2)^2$$

$$= 2(5.0)(1.96 + 1.65)^2/4 \approx 33.$$

Example: Two-group analysis of proportions

About 10,000 people in the United States may be wrongfully convicted of serious crimes each year with many of these cases resulting from mistaken eyewitness identification. A two-group experiment is planned to compare simultaneous and sequential photo lineup procedures. We will estimate the proportion of participants who correctly identify the suspect in each type of photo lineup after viewing a 4-second surveillance video of the suspect. After reviewing the literature on eyewitness accuracy, we set $\tilde{\pi}_1$ = .6 and $\tilde{\pi}_2$ = .75. We want a 95% confidence interval for $\pi_1 - \pi_2$ to have a width of 0.2. The sample size requirement per group is approximately

$$n_j = 4[\tilde{\pi}_1(1 - \tilde{\pi}_1) + \tilde{\pi}_2(1 - \tilde{\pi}_2)](z_{\alpha/2}/w)^2$$

$$= 4[.6(.4) + .75(.25)](1.96/0.2)^2 \approx 165.$$

Example: Two-group standardized mean difference

We want to compare two methods of treating phobia in children and we will use electrodermal responses to fear-producing objects as the dependent variable. The electrodermal scores do not have a clear psychological meaning and so it is difficult to specify a desired width of the confidence interval. However, we expect the standardized mean difference to be about 0.6 and would like a 95% confidence interval for the population standardized mean difference (δ) to have a width of about 0.5. The required sample size per group is approximately

$$n_j = (\tilde{\delta}^2 + 8)(z_{\alpha/2}/w)^2$$

$$= (0.6^2 + 8)\ (1.96/0.5)^2 \approx 129.$$

Example: Paired-samples t-test

We are planning a study that compares two smart phones in a population of college students. A sample of college students will be given both smart phones to use for one month and will rate each phone on a 1-10 scale at the end of the evaluation period. A review of the literature suggests that the correlation between these types of ratings could be as low as .4. We set $\tilde{\delta}$ = 0.5, α = .05, β = .1, and $\tilde{\rho}_{12}$ = .4. The number of college students we need to sample is approximately

$$n = 2\tilde{\sigma}^2(1 - \tilde{\rho}_{12})(z_{\alpha/2} + z_\beta)^2/(\tilde{\mu}_1 - \tilde{\mu}_2)^2$$

$$= 2(1 - \tilde{\rho}_{12})(z_{\alpha/2} + z_\beta)^2/\tilde{\delta}^2$$

$$= 2(1 - .4)(1.96 + 1.28)^2/0.25 \approx 51.$$

Example: Partial correlation

We want to assess the correlation between amount of violent video playing and aggressive behavior in a study population of high school male students. Hours of TV viewing and father's aggressiveness will be used as two (s = 2) control variables. After reviewing the literature, we decide to use .5 as the planning value of the partial correlation. We want the 95% confidence interval to have a width of about 0.3. The approximate sample size requirement is

$$n = 4(1 - \tilde{\rho}_{yx}^2)^2(z_{\alpha/2}/w)^2 + 3 + s$$

$$= 4(1 - .5^2)^2(1.96/0.3)^2 + 3 + 2 \approx 97.$$

Example: Squared multiple correlation

We want to estimate the squared multiple correlation between a measure of public speaking skill and four (q = 4) predictor variables in a study population of college freshman. We believe the squared multiple correlation will be about .3 and would like the 95% confidence interval to have a width of about .2. The approximate sample size requirement is

$$n = 16\tilde{\rho}_{y.\mathbf{x}}^2(1 - \tilde{\rho}_{y.\mathbf{x}}^2)^2(z_{\alpha/2}/w)^2 + q + 2$$

$$= 16(.3)(.7)^2(1.96/0.2)^2 + 4 + 2 \approx 232.$$

Example: 2 x 2 factorial design

We want to estimate a main effect $(\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2$ in a between-subjects design with 95% confidence, a desired confidence interval width of 2.0, and a planning value of 8.0 for the average within-group error variance. The contrast coefficients are .5, .5, -.5, and -.5. The sample size requirement per group is approximately

$$n_j = 4\tilde{\sigma}^2\left(\textstyle\sum_{j=1}^{k} c_j^2\right)(z_{\alpha/2}/w)^2$$

$$= 4(8.0)(.25 + .25 + .25 + .25)(1.96/2.0)^2 \approx 31.$$

Example: 2x2 factorial design with covariates

In the previous example, the sample size requirement was 31 per group (124 total). If it is not feasible to obtain that many participants, the sample size requirement can be reduced by including one or more (s) covariates in the design.

The sample size per group required to estimate the main effect in the previous example with a covariate that correlates .5 with the dependent variable is approximately

$$n_j = 4\tilde{\sigma}^2(1 - \tilde{\rho}^2)(\sum_{j=1}^{k} c_j^2)(z_{\alpha/2}/w)^2 + s$$

$$= 4(8.0)(1 - .25)(1)(1.96/2.0)^2 + 1 \approx 22.$$

Example: Interrater agreement

A sample of parole candidate files will be subjectively reviewed by two expert raters, and each rater will assign an “Approve” or “Disapprove” recommendation for each candidate. A 95% confidence interval for Guilford’s *G*-index of agreement will be computed from the sample of candidate files. Using a planning value of .8 for the *G*-index and a desired confidence interval width of .2, the required number of files that should be reviewed by both raters is approximately

$$n = 4(1 - \tilde{G}^2)(z_{\alpha/2}/w)^2$$

$$= 4(1 – .64)(1.96/0.2)^2 \approx 139.$$

Example: Cronbach's alpha reliability

A researcher wants a 95% confidence interval of Cronbach's alpha for a newly developed 10-item (m = 10) measure of "Integrity" using a random sample of working adults. In a previous study using college students, the sample value of Cronbach's alpha was .87 and will be used as a planning value. The researcher hopes the 95% confidence interval in the planned study will have lower and upper limits of about $\tilde{L}$ = .82 and $\tilde{U}$ = .92. The approximate sample size requirement is

$$n = [8m/(m-1)](z_{\alpha/2}/ln[(1-\tilde{L})/(1-\tilde{U})])^2 + 2$$

$$= [80/9][1.96/ln(2.25)]^2 + 2 \approx 54.$$

Example: “Big data” regression analysis

Suppose a company has a database of 750 million online customer transactions and wants to predict the purchase amount using about 100 customer characteristics as predictor variables. Instead fitting a regression model to all 750 million cases (which could take hours of computer time), the model can be fit very quickly to a random sample of cases such that the 99.99% confidence interval for the prediction error has an upper to lower endpoint ratio of 1.02 (an extremely narrow CI). The regression model can be fit to a random sample of

$$n = 2[3.89/ln(1.02)]^2 + 100 \approx 77{,}300 \text{ cases}$$

which would be about 10,000 times faster than analyzing the complete dataset and will produce virtually the same results.

Learning materials on CSASS website

- PowerPoint slides
- Lecture Notes
- Study Guide
- R functions
- SAS programs

Thank you.

Questions or comments?