Determining Sample Sizes using R

Dr. James. Ready

January 19, 2019

The purpose of this document is to differentiate between sampling for proportion of a population and sampling for inferential research and provide pre-written R functions to assist you in determining the appropriate sample size.

Sampling for proportion is often used in public policy and marketing research to obtain information about a characteristic of the population. Examples of this type of sampling include voting preferences, views on political issues, or thoughts on new products. Conversely, inferential testing involves statistical tests about relationships or differences in a sample. Understanding the relationship between two variables (e.g., job satisfaction and intent to quit) or differences in a variable based on a group or category (e.g., job satisfaction differences between male and female participants) are examples where inferential testing is employed.

Sampling for Proportion

When the proportion of a population is the unit of interest, then sampling requires four variables: margin of error, confidence level, population size, and the sample proportion.

Margin of Error (or Confidence Interval)

The Margin of Error (MOE) is the number of percentage points the results will differ from the real population value. MOE is a researcher-defined item. The larger the MOE, the smaller the sample size. Conversely, the smaller the MOE, the larger the sample. Start with a MOE of 5%, and adjust accordingly based on the cost of the calculated sample required.

Confidence Level

The Confidence Level (CL) is the probability that the value of a parameter falls within a specified range of values. CL is expressed as a %, such as 95%. A CL of 95% means that a researcher wants to be 95% confident that the resultant value (e.g., mean of the population) is true. The CL is inverse to the MOE; the larger the CL, the larger the sample size. The smaller the CL, the smaller the sample size. Start with 95% and adjust accordingly based on the cost of the calculated sample required.

Population Size

The population size influences the sample size; however, once the population exceeds 20,000, the sample size will become more static. A researcher performing proportional sampling will need to ascertain the size of the population to ensure representation.

Sample Proportion

If a researcher has prior studies to fall back upon, the results of the research can inform the sample proportion. For example, if prior research has shown that 60% of the population has a preference for an item, then this can be used in the sample size calculation. However, if the proportion is unknown, then it's wise to use 50% as the proportion.

Determining Sampling Sizes

First, load the **samplingbook** package (Manitz, 2016). This library contains specific functions that facilitate selecting the appropriate sample size based on a series of inputs. The following code will check your system for the package. If the package is not found, it will install it from the Cloud (along with any required libraries) and make prewritten functions accessible.

```
if(!require(samplingbook)){
    install.packages("samplingbook")
}
```

```
## Loading required package: samplingbook

## Loading required package: pps

## Loading required package: sampling

## Loading required package: survey

## Loading required package: grid

## Loading required package: Matrix

## Loading required package: survival

##
## Attaching package: 'survival'

## The following objects are masked from 'package:sampling':
##
##     cluster, strata

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart
```

```
library(samplingbook)
```

Once the **samplingbook** package is loaded, you can enter variable information that will allow you to determine the appropriate sample size.

Sample with a Finite Population

First, consider a scenario where a finite population is known. A researcher wants to survey company employees about how they feel about a certain topic. The company has 114 employees, and the researcher wants a 95% confidence level with a MOE of 5%. No prior research was used to determine the proportion of the employees having any feelings towards the topic. Using the parameters MOE (e) = .05, proportion (P) = .5, sample size (N) = 114, and confidence level ($level$) = .95, a sample size of 88 will be needed.

```
sample.size.prop(e = .05, P = 0.5, N = 114, level = 0.95)
```

```
##
## sample.size.prop object: Sample size for proportion estimate
## With finite population correction: N=114, precision e=0.05 and expected proportion P=0.5
##
## Sample size needed: 88
```

Using the same parameters, but increasing the size of an organization from 114 to 500 will change the required sample size from 88 to 218 -

```
sample.size.prop(e = .05, P = 0.5, N = 500, level = 0.95)
```

```
##
## sample.size.prop object: Sample size for proportion estimate
## With finite population correction: N=500, precision e=0.05 and expected proportion P=0.5
##
## Sample size needed: 218
```

If a sample size of 218 is too difficult to obtain, then a researcher can change the MOE; say from .05 to .10. This simple change reduces the sample size from 218 to 81 -

```
sample.size.prop(e =.10, P = 0.5, N = 500, level = 0.95)
```

```
##
## sample.size.prop object: Sample size for proportion estimate
## With finite population correction: N=500, precision e=0.1 and expected proportion P=0.5
##
## Sample size needed: 81
```

Sampling with an Infinite or Unknown Population

If the population is large (e.g., a State population) or is not known, eliminating the "N=xxx" variable from the function will result in a sample size of 385 -

```
sample.size.prop(e = .05, P = 0.5, level = 0.95)
```

```
##
## sample.size.prop object: Sample size for proportion estimate
## Without finite population correction: N=Inf, precision e=0.05 and expected proportion P=0.5
##
## Sample size needed: 385
```

Changing the precision from +/-5% to +/-10% reduces the required sample size from 385 to 97 -

```
sample.size.prop(e = .10, P = 0.5, level = 0.95)
```

```
##
## sample.size.prop object: Sample size for proportion estimate
## Without finite population correction: N=Inf, precision e=0.1 and expected proportion P=0.5
##
## Sample size needed: 97
```

Finally, changing the proportion from .50 to .70, reduces the required sample size from 97 to 81.

```
sample.size.prop(e = .10, P = 0.7, level = 0.95)
```

```
##
## sample.size.prop object: Sample size for proportion estimate
## Without finite population correction: N=Inf, precision e=0.1 and expected proportion P=0.7
##
## Sample size needed: 81
```

Sampling for Inferential Testing

Determining an appropriate sample size for inferential testing involves performing a power analysis. In a power analysis, a researcher generally makes an a priori decision about an estimated effect size and determines the appropriate Type I and Type II error risk before determining the sample size. Before a power analysis can be performed, an understanding of terms is required.

Effect Size

An effect size is a quantitative measure of a phenomenon. An example of an effect size is the correlation statistic representing the relationship between two variables. The American Psychological Association (2019) recommends reporting effect sizes in the results section of a manuscript (p. 89). Psychologist Jacob Cohen (1923-1998) developed many effect size measurements that are widely used today. Common effect size measurements and their related size in simple terms (Cohen, 1988, 1992), are -

- Difference between two populations: A Cohen's h effect size of .20, .50, and .80 would be considered small, medium, and large, respectively.
- Independence between cell counts: A Cohen's w effect size of .10, .30, and .50 would represent a small, medium, and large difference, respectively.
- Difference between two means: A Cohen's d effect size of .20, .50, and .80 would be considered small, medium, and large, respectively.
- Differences between three or more means: A Cohen's f effect size of .10, .25, and .40 would represent a small, medium, and large difference, respectively.
- Relationship between two variables: A Pearson r test statistic of .10, .30, and .50 would be considered small/weak, medium/moderate, and large/strong, respectively.
- Strength of multiple regression model: An f^2 effect size of .02, .15, and .35 would be represented as small, medium, and large, respectively.

There are other effect size measures; many developed to address differences in sample sizes between two or more groups. A researcher should explore effect size measurements based on the hypotheses developed and statistical tests selected. A source for effect size measurements and conversions between effect sizes is Ellis (2010).

Significance Level

The significance level, or α, is the measure of the strength of evidence set by the researcher to reject the null hypothesis and conclude statistical significance. The widely-used $\alpha = .05$ stems from a blending of the work of Ronald Fisher in 1925 and the team of Jerzy Neyman and Egon Pearson in 1933. Ronald Fisher (1990) suggested that a 1-in-20 chance of error was sufficient to reject a null hypothesis. Neyman and Pearson (1933) called this cut-off the significance level. Since then, researchers have used this as the default α value.

Statistical Power

While the α value is set by a researcher to reject the null hypothesis and avoid a Type I error, the statistical power of the test, or β, is the likelihood that a study will detect an effect when there is an effect there to be detected. Cohen (1988, 1992) posited that in social science, the risk of not rejecting the null hypothesis when one should, a Type II error, should be four times less critical. Using Cohen's rule, the β is often set to .80 (1-(4xα)) in social science.

Degrees of Freedom

The definition of *degrees of freedom* (*df*) in statistics is the number of independent variables that can be assigned to a statistical distribution. The purpose of understanding the degrees of freedom relates to the critical value necessary to reject a null hypothesis. Detailed calculations of the degrees of freedom for all tests can be complicated and is outside the scope of this document.

To determine sample sizes for inferential testing, only the chi-square test and ANOVA require a researcher to calculate the degrees of freedom a priori. In a simple test, *df* is the sample size minus 1 (*N*-1), which means that if one had 15 observations, the *df* would be 14. In a test involving two samples, the *df* is (*N1* + *N2*) – 2. To illustrate, if one was performing a *t*-test (see later) and each group had 30 observations, the *df* would be 60-2 or 58.

Two- vs. One-sided Test

A two-sided test relates to Null Significance Hypothesis Testing (NHST). Under NHST, a researcher is examining whether a test statistic is equal to or different from 0. If the test statistic is positive or negative, the null hypothesis (H_0) would be rejected, and the alternative hypothesis (H_A) would be accepted.

A one-sided test relates to directional hypothesis testing. Under directional hypothesis testing, a researcher makes an a priori decision to posit either positive or negative differences from 0 (e.g., $H_0 < 0$, $H_0 > 0$). If a directional hypothesis is used, a researcher is only considering one side of a normal distribution. Because of this consideration, a smaller sample is necessary.

Sample Size Selection Methods

Calculating sample sizes is based on the statistical test used. Six groups of tests are covered in this document.

First, load the *pwr* library (Champely, 2018). This library contains specific functions that facilitate selecting the appropriate sample size based on a series of inputs. The following code will check your system to see if

the *pwr* library is installed. If not, it will install it from the Cloud, along with any required packages, and make the functions accessible to you.

```
if(!require(pwr)){
    install.packages("pwr")
}
```

```
## Loading required package: pwr
```

```
library(pwr)
```

Difference between two proportions (Proportion test)

A test of proportions (**stats::prop.test**) is used examine the difference between two populations. To determine a sample size for a proportion test, four parameters must be estimated: Cohen's h, α, the β, and whether NHST or directional hypotheses testing will be used. To illustrate, assuming an effect size $h = .50$, an $\alpha = .05$, a $\beta = .80$, and an NHST-type hypothesis, a sample size of 31 is required.

```
pwr.p.test(h = .50,
           sig.level = .05,
           power = .80,
           alternative = "two.sided")
```

```
##
##      proportion power calculation for binomial distribution (arcsine transformation)
##
##               h = 0.5
##               n = 31.39544
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

If directional hypothesis testing was used, for example assuming a sample proportion would be greater than another sample proportion, then a sample of 25 would be needed.

```
pwr.p.test(h = .50,
           sig.level = .05,
           power = .80,
           alternative = "greater")
```

```
##
##      proportion power calculation for binomial distribution (arcsine transformation)
##
##               h = 0.5
##               n = 24.73023
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
```

Independence between cell counts (Chi-square test)

A chi-square test is used to answer a research question involving the association of categorical variables. Four parameters must be estimated to determine a sample size in a chi-square test: the effect size (Cohen's W), the degrees of freedom, the significance level, and statistical power. Degrees of Freedom (df) are determined by the number of data rows (r) and columns (c). To determine the df in a chi-square test, the formula is *(r-1)(c-1)*. For example, a study involving gender (M/F) in one dimension, and height (below 72 inches and 72 inches and above) in another dimension, would have 1 degree of freedom: (2-1)(2-1). In contrast, a chi-square test involving a 4-by-3 matrix would have a df of ([4-1][3-1]) or 6.

To illustrate, assuming a Cohen's $w = .30$, $df = 2$, an $\alpha = .05$, a $\beta = .80$, a sample size of 87 is required.

```
pwr.chisq.test(w = .3,
               df = 1,
               sig.level = 0.05,
               power = .80)
```

```
##
##      Chi squared power calculation
##
##               w = 0.3
##               N = 87.20954
##              df = 1
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: N is the number of observations
```

Using the same parameters but changing the df to 6, changes the required sample size to 151.

```
pwr.chisq.test(w = .3,
               df = 6,
               sig.level = 0.05,
               power = .80)
```

```
##
##      Chi squared power calculation
##
##               w = 0.3
##               N = 151.381
##              df = 6
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: N is the number of observations
```

The differences in the df illustrate that the more freedom a variable has to vary, the larger the sample size.

Difference between Means

There are three types of tests where are researcher can examine the difference in the mean: a one-sample, a two-sample, and a paired-sample.

One-sample *t*-test A one-sample t-test is used when a researcher is comparing the mean value of an item against a hypothesized value. This test can be used for both NHST and directional hypotheses. For example, in a situation where a researcher is hypothesizing that there is a moderate effect size difference (Cohen's d = .50), in any direction, a α = .05, and a β = .80, a sample size of 33 would be required.

```
pwr.t.test(d = .50,
           sig.level = .05,
           power = .80,
           type = "one.sample",
           alternative = "two.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 33.36713
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Changing the research by hypothesizing that the sample mean would be greater than the hypothesized mean requires only a one-sided test, and results in a slightly lower sample size.

```
pwr.t.test(d = .50,
           sig.level = .05,
           power = .80,
           type = "one.sample",
           alternative = "greater")
```

```
##
##      One-sample t test power calculation
##
##               n = 26.13753
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
```

Independent Samples *t*-test An Independent Samples t-test, or two-sample test, is used to compare the means of two different samples. For example, examining the difference in the Graduate Record Exam (GRE) scores by gender could be performed by this type of test. To illustrate, assuming a two-sided test (NHST), Cohen's d = .30, an α = .05, a β = .80, and a type of test involving two independent samples, a sample size of 64 in each group is required.

```
pwr::pwr.t.test(d = .50,
                sig.level = .05,
                power = .80,
                type = "two.sample",
                alternative = "two.sided")
```

```
##
```

```
##      Two-sample t test power calculation
##
##               n = 63.76561
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Changing the research by hypothesizing that the sample mean for one group would be greater than the sample mean of the other requires only a one-sided test, and results in a slightly lower sample size in each group.

```
pwr::pwr.t.test(d = .50,
                sig.level = .05,
                power = .80,
                type = "two.sample",
                alternative = "greater")
```

```
##
##      Two-sample t test power calculation
##
##               n = 50.1508
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
##
## NOTE: n is number in *each* group
```

Paired samples *t*-test A paired-sample t-test, also called a dependent sample t-test, is used when a researcher measures the difference in scores between two observations. A typical application of the paired-sample t-test is in the implementation of a pre-test, post-test or repeated-measures research design where an observation occurs before and after a treatment or intervention. For example, testing students on a skill, performing instruction, and testing the students again.

To illustrate, assuming a two-sided test (NHST), Cohen's $d = .30$, an $\alpha = .05$, a $\beta = .80$, and a type of test involving paired samples, a sample size of 33 in each group is required.

```
pwr.t.test(d = .50,
           sig.level = .05,
           power = .80,
           type = "paired",
           alternative = "two.sided")
```

```
##
##      Paired t test power calculation
##
##               n = 33.36713
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
```

```
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

Changing the research by hypothesizing that the difference in means would be would be greater (e.g., post-test scores would be higher than pre-test scores) requires only a one-sided test, and results in a slightly lower sample size.

```
pwr.t.test(d = .50,
           sig.level = .05,
           power = .80,
           type = "paired",
           alternative = "greater")
```

```
##
##      Paired t test power calculation
##
##               n = 26.13753
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
##
## NOTE: n is number of *pairs*
```

Analysis of Variances (ANOVA) between three or more groups

The Analysis of Variances (ANOVA) is similar to the Independent Samples t-test. Where the t-test only measures the difference in the means based on two categories, the ANOVA is used for three or more groups. For example, if a researcher wanted to measure the mean value of employee job satisfaction from five different locations of a company, an ANOVA would be an appropriate test. To illustrate, assuming a measure is recorded from five different locations (df = k-1 = 4), a moderate effect size (f = .25), an α = .05, and a β = .80, a sample size of 45 in each group are required (225, overall).

```
# identifying a sample size based on the four required parameters
pwr::pwr.anova.test(k = 4,
                    f = .25,
                    sig.level = .05,
                    power = .80)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 4
##               n = 44.59927
##               f = 0.25
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

Relationship between two variables (correlation)

One of the most common statistical testing approaches is to examine the relationship between two variables using the Pearson Product-Moment Correlation Coefficient (r). To illustrate, assuming an NHST-type hypothesis, an $r = .30$, an $\alpha = .05$, and a $\beta = .80$, a sample size of 84 is required.

```
# identifying a sample size based on the four required parameters
pwr::pwr.r.test(r = .30,
                sig.level = .05,
                power = .80,
                alternative = "two.sided")
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 84.07364
##               r = 0.3
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Changing the research question from an NHST-type hypothesis to a directional-type hypothesis (H_0: $r => 0$), then the sample size would be reduced to 67.

```
# identifying a sample size based on the four required parameters
pwr.r.test(r = .30,
           sig.level = .05,
           power = .80,
           alternative = "greater")
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 66.55463
##               r = 0.3
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
```

Relationships between two variables can also be measured by the non-parmetric, distribution-free Spearman Rank-Order Correlation Coefficient (rho or ρ) or the Kendall Rank-Order Correlation Coefficient (tau or τ). Determining sample sizes using Spearman and Kendall coefficients can use the same methodology as that for Pearson.

Predicting Change in an interval/continuous Dependent Variable (DV) based on two or more Independent Variables (IV; multiple regression)

A common form of examining prediction is through multiple linear regression. Under this form of regression, the dependent variable must be an interval/continuous variable, and the independent variables can take the form of interval, continuous, or category. To determine an appropriate sample size, a researcher needs to identify the number of independent variables, the estimated effect size (f^2), the α, and the β. To illustrate, assuming six independent variables, a moderate effect size ($f^2 = .15$), an $\alpha = .05$, and a $\beta = .80$, a sample size of 90 is required.

```r
pwr::pwr.f2.test(u = 6,
                 f2 = .15,
                 sig.level = .05,
                 power = .80)
```

```
##
##      Multiple regression power calculation
##
##               u = 6
##               v = 90.30998
##              f2 = 0.15
##       sig.level = 0.05
##           power = 0.8
```

Sometimes, ordinal independent variables are misclassified as interval. Ordinal variables should be 'dummy-coded,' which is outside the scope of this document. Assuming that one of the IVs in our scenario is ordinal and requires recoding into four unique IVs, the number of variables will increase from 6 to 9 (6-1+4). The associated sample size will also increase from 90 to 103.

```r
pwr::pwr.f2.test(u = 9,
                 f2 = .15,
                 sig.level = .05,
                 power = .80)
```

```
##
##      Multiple regression power calculation
##
##               u = 9
##               v = 103.0567
##              f2 = 0.15
##       sig.level = 0.05
##           power = 0.8
```

References:

American Psychological Association (2019). *Publication manual of the American Psychological Association* (7th ed.). American Psychological Association.

Champely, S. (2018, March 3). *pwr: Basic functions for power analysis.* Retrieved from https://cran.r-project.org/web/packages/pwr/pwr.pdf

Cohen, J. (1988). *Statistical power analysis for the behavior sciences.* Routledge.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. https://dx.doi.org/10.1037/0033-2909.112.1.155

Ellis, P. D. (2010). *An essential guide to effect sizes.* Cambridge University Press.

Fisher, R. A. (1990). *Statistical methods, experimental design, and scientific inference* (14th ed.). Oxford University Press.

Manitz, J. (2017, May 23). *samplingbook: Survey sampling procedure.* Retrieved from https://cran.r-project.org/web/packages/samplingbook/samplingbook.pdf.

Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society, 29*(4), 492-510. https://doi.org/10.1017/S030500410001152X