

Sample Size Estimation

Wayne Enanoria, PhD, MPH
Center for Infectious Disease Preparedness
University of California at Berkeley

15 August 2007

1 Introduction

It is important to understand how to analyze complex survey samples using statistical packages like **R**. Due to limited time and resources, we are often not able to take measurements on everyone in a population under study. As a result, we may take a sample of the population using the various techniques we have for sampling. After we collect the data, the fundamental questions to ask are "What results would we obtain if we had data on the entire population?". Since we did not take measurements on everyone, we will need to use our study sample as a "representation" of our total population.

The purpose of this article is to illustrate how to calculate a desired sample size estimate for a simple proportion. The statistical program **R** will be used to illustrate the calculations.

2 The Estimation of Sample Size Assuming Simple Random Sampling

In setting up a survey, one needs to think about taking a random sample of some finite size and collecting data on each sampling unit. A fundamental question in thinking about a sample is "how many sampling units do I include in my sample?". In order to perform a sampling size calculation, an investigator needs to identify two things:

1. the margin of error we will allow; for example, we would like to estimate the proportion of people who die after a disaster with a margin of error of plus or minus 10% (degree of accuracy).
2. the level of statistical certainty or confidence interval (α = probability of Type I error);

Thus, we designate that there is some margin of error d in the estimated proportion p of units in relation to the true proportion P that one could expect. There is some (hopefully small) risk α that we are willing to bet that the actual error is larger than d . In doing our sample size calculation, we are assuming:

$$Pr(|p - P| \geq d) = \alpha \quad (1)$$

If simple random sampling is assumed, we can estimate p from the data and calculate the variance of p , $V(p)$, :

$$V(p) = E(p - P)^2 = \frac{S^2}{n} \left(\frac{N - n}{N} \right) = \frac{PQ}{n} \left(\frac{N - n}{N - 1} \right) \quad (2)$$

and the standard error of p , σ_p :

$$\sigma_p = \sqrt{V(p)} = \sqrt{\frac{N - n}{N - 1}} \sqrt{\frac{PQ}{n}} \quad (3)$$

If we assume that the distribution of this proportion is normally distributed, we can then express the degree of precision, d , with the sample size, n , as follows (Cochran 1977):

$$d = Z \sqrt{\frac{N - n}{N - 1}} \sqrt{\frac{PQ}{n}} \quad (4)$$

where Z is the abscissa of the normal curve that cuts an area of α at the two tails (Cochran 1977). By solving for n , we get:

$$n = \frac{\frac{Z^2 PQ}{d^2}}{1 + \frac{1}{N} \left(\frac{Z^2 PQ}{d^2} - 1 \right)} \quad (5)$$

For practical use, we can substitute an advance estimate of p for P . If N is very large, the equation reduces to:

$$n_0 = \frac{Z^2 pq}{d^2}. \quad (6)$$

Thus, we can use this formula in our calculation of a desired sample size (assuming simple random sampling).

3 Sample Size Calculations using R

In R, we will first create a function that takes several inputs and gives us the answer we want: a sample size, n . In order for our function to work, we need to give it certain values often referred to as *parameters*. For a sample size calculation, these parameters are:

- abscissa of the normal curve (called Z);
- estimated proportion that one is trying to estimate in the population (called p);
- degree of accuracy desired, assumed to be half the confidence interval (called d);
- the probability of Type I error (called *alpha*).

```

> srssamp <- function(alpha, pp, dd) {
+   zz <- qnorm(1 - alpha/2)
+   qq <- 1 - pp
+   nn <- (zz^2 * pp * qq)/dd^2
+   nn
+ }

```

Now we can run our new function, giving the function the parameters it needs to calculate sample size.

```

> srssamp(alpha = 0.05, pp = 0.5, dd = 0.1)

```

```

[1] 96.03647

```

Note that for a proportion equal to 50 percent + or - 10 percent, the estimated sample size is the widely quoted $n=96$ individuals, assuming simple random sample (Brogan 1994, p. 303). The nice thing about our function is that we can now vary the input parameters to see how the inputs change our estimate of the desired sample size. For a margin of error equal to 5 percent or 15 percent (versus 10 percent), we can easily calculate this with our function by changing our input value for *dd*.

```

> srssamp(alpha = 0.05, pp = 0.5, dd = 0.05)

```

```

[1] 384.1459

```

```

> srssamp(alpha = 0.05, pp = 0.5, dd = 0.15)

```

```

[1] 42.68288

```

We can also look at the estimated sample size and how it varies with varying proportions that we are trying to estimate in the population.

```

> estprop <- seq(0.01:1, by = 0.01)
> srssamp(alpha = 0.05, pp = estprop, dd = 0.1)

```

```

[1] 3.803044 7.529259 11.178645 14.751202 18.246929 21.665828 25.007897 28.273137
[9] 31.461548 34.573129 37.607882 40.565805 43.446899 46.251164 48.978600 51.629207
[17] 54.202984 56.699932 59.120051 61.463341 63.729802 65.919433 68.032236 70.068209
[25] 72.027353 73.909668 75.715153 77.443810 79.095637 80.670635 82.168804 83.590144
[33] 84.934655 86.202336 87.393188 88.507211 89.544405 90.504770 91.388305 92.195012
[41] 92.924889 93.577937 94.154156 94.653545 95.076106 95.421837 95.690739 95.882812
[49] 95.998056 96.036471 95.998056 95.882812 95.690739 95.421837 95.076106 94.653545
[57] 94.154156 93.577937 92.924889 92.195012 91.388305 90.504770 89.544405 88.507211
[65] 87.393188 86.202336 84.934655 83.590144 82.168804 80.670635 79.095637 77.443810
[73] 75.715153 73.909668 72.027353 70.068209 68.032236 65.919433 63.729802 61.463341
[81] 59.120051 56.699932 54.202984 51.629207 48.978600 46.251164 43.446899 40.565805
[89] 37.607882 34.573129 31.461548 28.273137 25.007897 21.665828 18.246929 14.751202
[97] 11.178645 7.529259 3.803044 0.000000

```

We can modify our program if we want to include an estimated sample size for a cluster sample of a given design effect.

```

> de.samp <- function(alpha, pp, dd, deseff) {
+   zz <- qnorm(1 - alpha/2)
+   qq <- 1 - pp
+   nn <- deseff * ((zz^2 * pp * qq)/dd^2)
+   nn
+ }
> de.samp(alpha = 0.05, pp = 0.5, dd = 0.1, deseff = 2)

[1] 192.0729

```

We could still use this new function if we wanted to calculate a sample size assuming simple random sampling by assuming that the design effect equals 1.0.

```

> de.samp(alpha = 0.05, pp = 0.5, dd = 0.1, deseff = 1)

[1] 96.03647

```

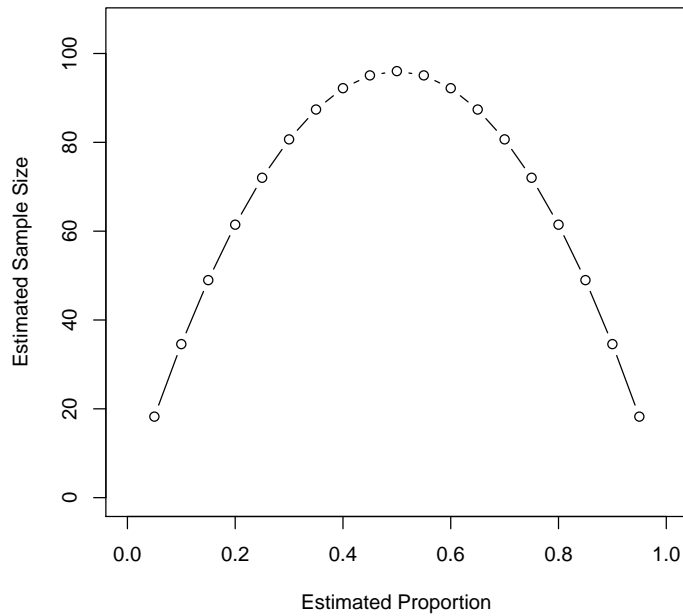
Let's modify our function to include a plot of the estimated sample size by the estimate proportion we are trying to measure in the population.

```

> de.samp.plot <- function(alpha, pp, dd, deseff, sampplot = c("TRUE", "FALSE")) {
+   sampplot <- match.arg(sampplot, several.ok = FALSE)
+   zz <- qnorm(1 - alpha/2)
+   qq <- 1 - pp
+   nn <- deseff * ((zz^2 * pp * qq)/dd^2)
+   if (sampplot == "TRUE") {
+     nn
+     ylim.max <- max(nn) + 10
+     plot(pp, nn, type = "b", xlab = "Estimated Proportion", ylab = "Estimated Sample Size",
+          xlim = range(0, 1), ylim = range(0, ylim.max))
+   }
+   else {
+     nn
+   }
+ }

```

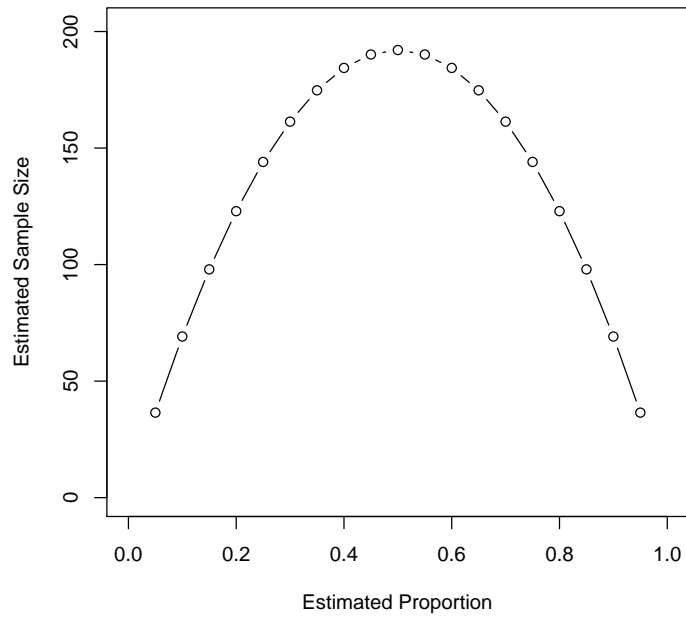
And now, let's run our new function.



If we wanted to run the function without producing the graph, we could modify our command.

```
> de.samp.plot(alpha = 0.05, pp = estprop, dd = 0.1, deseff = 1, sampplot = "FALSE")
[1] 18.24693 34.57313 48.97860 61.46334 72.02735 80.67064 87.39319 92.19501 95.07611 96.07611
[11] 95.07611 92.19501 87.39319 80.67064 72.02735 61.46334 48.97860 34.57313 18.24693
```

If we want to calculate the desired sample size for a rapid needs assessment assuming that our design effect is equal to 2.0, we can re-run our function and produce a graph.



4 References

1. Brogan D, Flagg EW, Deming M, Waldman R. Increasing the accuracy of the Expanded Programme on Immunization's Cluster Survey Design. *Ann Epidemiol* 1994;4:302-311.
2. Cochran WG. *Sampling Techniques*. John Wiley & Sons:New York, 1977. pp.74-76.