

## ESTIMATING SAMPLE SIZES FOR SURVEYS

Your primary source for this topic should be the class readings; this memo should be seen as a supplement that attempts to clarify some issues that traditional presentations leave unclear. First, I give the general recipe that's presented in textbooks for simple random sampling designs. Then I try to clarify how actual practice usually deviates from it:

1. SET A CONFIDENCE LEVEL: This part is easy. The longstanding convention is "95% confidence." This is one less thing to decide, and therefore one less source of existential angst in your life. It implies that the  $t$ -value in the formulas can be replaced with the specific value  $t = 1.96$  (the critical  $t$  value for a 2-tailed  $\alpha = .05$  --  $.025$  in each tail).
2. SET A TOLERABLE ERROR (TE) LEVEL: Tolerable error ( $te$ ) is the degree of sampling precision you are aiming for and think you can afford; specifically, it is half the width of the full confidence interval--the half on either side of the result you estimate -- i.e., "your estimate  $\pm te$ ". To set  $te$ , first you need to decide whether your major outcome is likely to be a percentage (proportion) vs. a mean on some scale.

For estimates expressed as percentages/proportions (e.g., "72% of students have smoked at least one cigarette"), a  $te$  of 2% means that you'll be 95% confident that your estimate is within  $\pm 2\%$  of the "true value"-- in this case, between 70 and 74% (i.e.,  $72\% \pm 2\%$ ).<sup>1</sup>

For estimates expressed as means,  $te$  isn't a percentage -- e.g., it isn't  $\pm 2\%$  of the mean value and you wouldn't use ".02" in your calculations (below). Instead, it is some value in the same metric as the outcome measure. It might be  $te = \$200$  ("mean = \$1700 in monthly income  $\pm \$200$ "), or  $te = 6$  lbs. ("mean = 15 lbs of weight loss  $\pm 6$  lbs"). It would only be  $te = .02$  if .02 happened to be a meaningful increment in the response scale, e.g.,  $\pm .02$  ounces, where .02 refers to ounces, not "2%").

*Ceteris paribus*, more precision is better than less precision. But in practice, reductions in  $te$  can be quite expensive, and you may decide that the goals of the survey don't really require extreme precision anyway. My suggestion is to use a spreadsheet, and do the sample size calculations for a range of  $te$  values (e.g., for proportions, make each column a different  $te$  ranging from  $\pm 1\%$  at the precise end to  $\pm 6\%$  at the fuzzy end).

---

<sup>1</sup> I put "true value" in quotation marks because  $te$  only refers to sampling error (how well the sample matches the population), not measurement error (the reliability and validity of your measures).

3. ESTIMATE THE EXPECTED STANDARD DEVIATION: Here's the part that everyone finds baffling. "If I knew so much that I could specify the standard deviation (and other parameters of the outcome distribution), why would I bother conducting the survey?" In general, we are completely clueless about the expected standard deviation, yet we need it to solve for  $n$ , the sample size. Here are some suggestions for making guesses that aren't completely arbitrary:

(a) Look at previous research: What  $sd$ 's have other studies found? This is less scientific than it sounds; it's really a judgment task. You have to decide: What makes another study relevant to this study? To the extent that the other survey differs from your proposed survey in time, location, population, question content, and question format, it may or may not be a good predictor of  $sd$  in your survey. (Question format is particularly important; if they asked a dichotomous yes/no question, and you are planning to ask using a 5-point scale, their  $sd$  is completely uninformative.)

(b) If you are estimating proportions, you can assume the true proportion you are estimating (e.g., the % of citizens who'll vote for the *Let's Abolish Tenure and Send Professors to Labor Camps* initiative) is 50%. Why 50% Because  $50\% = .50$  is the most conservative assumption, for the following reason. For a proportion, the standard deviation will be at its maximum when  $p = .50$ . (Warning: If this assumption is unduly conservative for your topic, you may spend too much money on sampling.)

(c) For outcomes involving means, decide on the plausible range (maximum value - minimum value) that will cover 95% of all cases; i.e., all but the extreme outliers. Statistically, if the distribution is approximately normal (bell-shaped), 95% of the values will fall within  $\pm 2$   $sd$ 's of the mean. Therefore, you can take your range (i.e., max - min), divide it by 4, and come up with an estimate of the standard deviation. (This approach is easy for a 7-point scale; you just take the end points 1 & 7, then  $7-1 = 6$ , and  $6/4 = 1.50$ . The approach fails for extremely skewed distributions; e.g., dollar scales that are bounded at 0 on one end but unbounded on the high end, like punitive damage awards.)

Because of the ambiguity of forecasting a  $sd$ , I recommend that you examine a range of  $sd$ 's (i.e., a sensitivity analysis), and select the most conservative  $n$  that you can afford. This adds a second dimension to your planning spreadsheet; you have a vector of  $te$  values (columns) and a vector of  $sd$  values (rows, or vice versa), and in each cell of the matrix you calculate the  $n$  you'd need to achieve that  $te$  level given that amount of variation.

4. Estimate n. Actually, I recommend you estimate a *range* of n's, under different combinations of tolerable errors and expected standard deviations.<sup>2</sup>

(a) Before applying the finite population correction (below), the formula for n' is:

- For means:  $n' = sd^2 / (te/t)^2 = sd^2 / (te/1.96)^2$  In our case,  $sd^2 / (.04/1.96)^2 = sd^2 / .0004$ . So plug in your range of guesses for the standard deviation (sd) and you'll get a range of n's.
- For proportions<sup>3</sup>:  $n' = pq / (te/t)^2 = pq / (te/1.96)^2$  In our case,  $pq / (.04/1.96)^2 = pq / .0004$ . In this case you can plug in .25 as the most conservative pq estimate (for a proportion = 50%,  $pq = .5 * .5 = .25$ ). So  $n' = .25 / .0004 = 625$ .

(b) Now apply the finite population correction ("fpc"):  $n = n' / (1 + n'/N)$ . (This adjusts for the fact that your population isn't infinitely large, and you are sampling without replacement.) If you want, you can merge this step with the previous calculations, but since n' appears twice, the formula becomes unmanageable: e.g.,  $n = [pq / (te/t)^2] / (1 + [pq / (te/t)^2] / N)$ . More importantly, you probably want to keep this step separate because you *might not know the population size (N)*. If that's the case, you'll have to *estimate n for a range of plausible guesses about N*. In other words, you'll end up with a range of n estimates across alternative combinations of te, sd, and N. But you'll find the fpc doesn't matter much if your population is quite large. As seen in the following table (where I assumed an n' of 500), the fpc only makes a notable difference for the first row of the table:

N	n	n - n'	% of n'
1,000	333	-167	67%
5,000	455	-45	91%
10,000	476	-24	95%
50,000	495	-5	99%
100,000	498	-2	>99%
500,000	500	0	100%
1,000,000	500	0	100%

5. Adjust your sample size (n) to account for low response rates: Ideally, your lit review will tell you how low the response rate is likely to be. For example, if you expected a 70% response rate and estimate n = 500, then you'll need to sample  $500 / .70 = 714$  people to make sure you get the 500

<sup>2</sup> My formulas don't mention the "standard error," but note that it is embedded within them. We know that  $te = se * t$ , or  $te = se * 1.96$  for a 95% confidence interval. So roughly, te is twice the standard error. If we set  $te = .04$  (for +/- 4%), then  $te = .04 = se * 1.96$ . Algebraically, we can solve for se:  $se = te / t = .04 / 1.96 = .02$ .

<sup>3</sup> Note that in class, I said  $n' = pq / se^2$ . Since  $se = te / t$ , this is the same formula. Also, p. 82 of Kalton's book might look different, but it isn't. First, his example uses percentages (0-100) rather than proportions (0-1), so his te is a nice round 2 rather than .02 for ±2%. But his P and Q are 50% rather than .50, and  $(50% * 50%) / 2^2 = .5 * .5 / .02^2 = 625$  so it works out the same. His formula is  $n' = 1.96^2 PQ / 2^2$  which could be restated as  $n' = t^2 pq / te^2$ . Algebraically, we find that our formula of  $n' = pq / (te/t)^2$  is identical; since  $pq / (te/t)^2 = pq / (te^2/t^2) = t^2 pq / te^2$ .

respondents you'll need. If you don't know what response rate to expect, you have yet another dimension to vary in your planning spreadsheets; now you can see why I recommend using Excel or comparable spreadsheet software!

Multiple variables: Usually there are many different variables you are trying to estimate in your survey. Pick the one you need the most precision on, and base your calculations on that.

Multiple strata: The simplest, most conservative (but inefficient) method is to compute  $n$  separately for each stratum. Typically, this means that you'll have the same  $n'$  (and roughly the same  $n$  after the  $fpc$ ) for each stratum, even though some strata are much smaller in the population than others. (You'd correct for this oversampling in the analyses by applying sample weights to the responses. See the readings for more detail.) There are more complex formulas for handling complex stratified/cluster design hybrids.

Clusters: Ideally, one would sample enough clusters to statistically represent the full population of clusters at a given level of precision. Realistically, this would require far too many clusters to be affordable, and it is doubtful you would really need all that information anyway. But cluster designs still require extra  $N$ . Technically, a cluster design's total sample size needs to be adjusted for the "design effect." See Kalton for the details, but the basic intuition is that a given cluster of 100 people might well be far more homogeneous than a random sample of 100 people, and as a result, your cluster sample will provide less information. Computing the design effect requires one to make assumptions about the "intracluster correlation" and you may feel you have little basis for doing so. Instead, a crude rule of thumb is to simply compute your sample size target as if you were doing a SRS, but then **double the resulting  $N$**  to compensate for the design effect. (The true design effect might be even larger.) Ouch! So what you gain logistically from a cluster design you lose in precision – there's no free lunch. If you think it is unrealistic to include more than a small number of clusters in your study (out of the total number of clusters possible), then it may be better to pick those clusters purposively rather than randomly. Purposive sampling can involve the choice of "typical" clusters, or instead "extreme clusters," or a mix of different types of clusters – your research questions should dictate your choice. What you don't want is a survey of California schools that comes out completely differently depending on whether Los Angeles Unified School District ended up being one of the clusters or not. (If you choose clusters purposively, you would need to weight the data to reflect this before analysis.)

---

You may take comfort in the fact that this wasn't obvious to me; I had to work it out to be sure there wasn't a conflict!