



"turning data into dollars"

Tom's Ten Data Tips – February 2009

Survey Sampling

Sampling applies whenever we choose to draw inferences from a limited number of observations rather than the entire "population." There can be many reasons to use sampling like not having all data available, or prohibitive costs for acquiring all data. Sometimes all the data are there (as in a census), but are too unwieldy to handle.

The word "population" has a special meaning in statistics. Our conclusions based on the sample refer to a wider group we want to make inferences about. But strictly speaking we mean each and every *measurement* pertaining to the population. It is important to realize we draw inferences about real "things" out there on the basis of a proxy: the set of recording made about it, which is perforce an imperfect representation.

1. Be 'Greedy' When Sampling

When you need to calculate the "optimal" size of a sample, it is *not* the time to be lazy and take a 'stab' at this number. In any practical research design, there are always constraints, monetary constraints on the size of the sample(s) being one of them. And when you carelessly add too many subjects to your sample, you are implicitly wasting resources that ought to be used to further *other* research objectives. The fact of the matter is that you need to spend at least five minutes doing elementary calculations. Sometimes this means lifting your statistics 101 book off the shelf. So be it.

Not all strata necessarily need to be of equal size for the same significance level, nor is there any research objective served by choosing "round" numbers, etc. This is a profession, not an avocation.

2. Sample Size Has (Typically) Nothing To Do With Population Size

For all practical intents and purposes, one can say that a sample of 1000 can represent a population of 100,000, 1,000,000 or 10,000,000 with the *same* degree of precision. Likewise, when you sample from a dish you are cooking, you sample the same amount regardless of whether you are cooking for two or twelve people.

This phenomenon runs counter to many people's intuition. Precision of a sample depends on size, not the proportion of the population you have sampled. There is an asymptote when you have sampled a large chunk, say, 5-10% of the population or more, but it is rarely used (and appears in very few statistics books). The correction factor for standard deviation in large samples (as a fraction of the population): $\sqrt{N-n/N-1}$ where N is the size of the population and n is the size of the sample. You can see it decreases the standard error because it is always smaller than 1, but it doesn't amount to much until n grows large relative to N.

All this of course assumes that your sample is unbiased, more on that in tip# 3.

3. Bias Can Come From Selection And Responders

Two major sources of bias that can invalidate surveys are selection and respondent bias. Selection bias occurs when the sample is not (fully) representative of the population. No matter how careful the study, selection bias prevents drawing valid inferences. When you sample from the telephone book, and call during daytime hours, you're less likely to draw employed people in your sample. When you sample the first N records from a database, there could be a correlation with the ordering, etc.

Respondent bias occurs when the method employed leads to invalid results despite the fact that the responders are a perfect representation of the population. This can happen because the question method invokes socially desirable responses, or acquiescence, etc.

4. Selection Error Itself Has Two Sources: Bias And Random Error

You can differentiate between two *unrelated* sources of sampling error: bias and random error. Sampling bias, as we have seen can come from sampling "the wrong subjects". When you interview people at a shopping mall during office hours, employees will be under represented.

What most laypeople refer to when they mention "sampling error" or "uncertainty" stems from the fact that a random sample of 1000, if redrawn from the same population (say, the electoral role), will have

differing numbers of voters in favor of a party. The *expected* number of Democrats is equal to the population proportion, but the *actual* number will be different every time you redraw. The latter is called random or sampling error.

5. Stratified Sampling Can Be Cheaper (And Safer)

Stratified sampling means that you assemble your sample while explicitly taking subgroups (strata) into account. You do this when you know or expect subgroups to differ significantly. Suppose a researcher wants to study consumer spending behavior. If you know from previous research that gender is correlated with spending, it would be wise to consider stratifying males and females. Needless to say, the assumption that women are spendthrifts is purely hypothetical in this example!

Let's say the population is made up of 40% men, and 60% women. Strata are *always* mutually exclusive. When you use proportional sampling, you would draw 3 women for every 2 men in your sample. A purely "random" sample might contain slightly different proportions. In this way you safeguard against sample fluctuations. When used in this way, stratification leads to a "safer" sample. If you would use disproportionate (or optimum) allocation you will choose a number per stratum that minimizes your sample size yet maximizes accuracy of conclusions. This typically means oversampling the rare category, so in this example you would choose the same number of men as women. This will (usually) lead to a more economical sample, with equally accurate conclusions for men and women.

6. Be Weary Of Point Estimates

When a test result is presented as a point result, essential information is missing. An example of a "point result": "70% of respondents support legalization of euthanasia." What is missing is not only the size of the sample but also how sure you are of this conclusion. To overcome *both* problems, the same study results could also be presented as: "Between 66.3% and 73.6% of respondents support legalization of euthanasia, with 95% certainty."

Bear in mind that the ubiquitous 95% significance level generally used in social sciences is arbitrary but still needs to be made explicit. It informs us how likely it is that the 'true' population value lies within this range. Also note that the confidence interval, in this case 66.3-73.6 provides a measure that illustrates the *accuracy* of measurement.

Accuracy is influenced by two things: the method chosen and the size of the sample. To the reader neither method nor sample size are relevant – merely accuracy. The sample size by itself may be informative and interesting, but it doesn't tell you how accurate the expected results will be (that also depends on the *method* used).

7. When You Increase Sample Size By Factor N, Accuracy Goes Up By Only \sqrt{N}

The cost for setting up a poll breaks down into a one time cost to set up the study, and marginal costs for each additional responder that gets drawn into the sample. For practical purposes, the setup costs of a study is a fixed investment. The same more or less holds for analysis costs which are the same irrespective of sample size. Data cleaning on the other hand can be estimated as a linear function of the number of records, with some possibilities to gain economy of scale for *very* large projects. Depending on these costs one can choose to increase sample size. Unfortunately there is a law of diminishing returns: to be *twice* as accurate, the sample size needs to be *four* times as large. This is where the square root of the incremental growth comes in.

8. 'Standard' Confidence Intervals Can Be Improved Upon

In most research reports, the "standard" confidence interval for proportions is used. An example would be: 36% of voters \pm 2% are in favor of candidate A. This is made up of the center (36%) which is the *expected outcome*, \pm z-score* σ_p where σ_p is the standard deviation given by $\sqrt{p(1-p)/N}$; p=expected proportion (in this case 36%), and N is the sample size. The z-score follows from the well known normal distribution. Note that for 'smallish' samples (say, less than 60 subjects), you use student's t distribution, because the approximation to a z-distribution is insufficiently accurate.

It is slightly more exact to calculate the confidence interval based on the *exact* binomial distribution rather than the normal approximation of the binomial distribution. The sampling distribution of a proportion is not a continuous distribution because proportions don't *exactly* follow the normal distribution. A simple way to approximate this approximation (J) is to add 0.5/N to the confidence interval on either side. We then get: $p \pm$ z-score* $\sigma_p + 0.5/N$ where σ_p is the standard deviation given by $\sqrt{p(1-p)/N}$; p=expected proportion, and N is the sample size. As you can tell, the difference is immaterial in large samples.

9. Percentage Confidence Intervals Are Really *Asymmetric*

An outcome like $36\% \pm 2\%$ is an example of a symmetrical confidence interval. However, for proportions the "true" confidence interval is *never* exactly symmetric (except around 50%). This is easy to see when you 'slide' the percentage towards 0: for 1% this would lead to a confidence interval of -1% to 3%. That can't possibly be right, now can it? The reason why a confidence interval for the proportion is asymmetric is because the standard deviation is calculated as $\sqrt{p(1-p)/N}$ (with p =expected proportion, and N the sample size), and you can see that the density distribution thus varies with the proportion. Therefore, the nearer the proportion lies to 0 or 1, the less accurate a *symmetric* confidence interval is. The math gets a little bit more involved. For your convenience, a link to a site where you can plug in your numbers and get the asymmetric confidence intervals is <http://faculty.vassar.edu/lowry/prop1.html>

10. When Can You Say That "Enough Is Enough?"

The problem of finding a "large enough" sample is that you need to define what represents "enough." The answer to that question drives deeper into the motivation of why you are doing research in the first place, and is (or at least should be) part of the initial briefing. One empirical fact that can be hard to determine beforehand is the amount of variance in the population you are studying. After all, if you knew all there is to know, you wouldn't need to do research, right?

Increasing the sample size (see also tip# 7), and avoiding as many sources of bias (tips# 3 & 4), serve to *converge* on a representative sample. For bias this is tricky and never absolutely sure, but for 'genuine' random fluctuations there are quantitative methods to assess how close you are (at a particular confidence level, see tip# 6) to estimated population parameters. The trade-off between investing more for a more accurate answer then needs to be made transparent to business owners commissioning the research. The final call is a management call, never a research decision, but it needs to be properly informed by someone with sufficient statistical background.