An Introduction to Sampling

Sampling is the process of selecting a subset of units from the population. We use sampling formulas to determine how many to select because it is based on the characteristics of this sample that we make inferences about the population. In preparing to take a sample, there many questions we might have, including:

- Can we make procedural errors that influence the results, but are not related to the sample itself?
- Can we make errors in our sampling?
- How can we determine the size and the accuracy of the sample result?

SYSTEMATIC ERRORS (NON-SAMPLING ERRORS)

Systematic errors result from decisions that bias the sample selection or response to your survey. Four common mistakes are made:

1. Population Specification Error: This error is one of not understanding who you should be surveying. As a simple example, imagine you are preparing a survey about the consumption of breakfast cereals. Who do you survey? It might be the entire family, the mother, or the children. The family consumes cereal, the mother purchases, and the children influence her choice.

2. Sample Frame Error: A frame error occurs when the wrong sub-population is specified from which the sample is drawn. A classic frame error occurred in predicting the 1936 presidential election between Roosevelt and Landon. The sample frame used was from car registrations and telephone directories. In 1936, car and telephone owners were largely Republicans. While the results may have reflected the sample, the predictions were not accurate for the US as a whole and the results wrongly predicted a Republican victory.

3. Selection Error: Selection error results when the respondents self select their participation... those who are interested respond. Selection error can be controlled by going extra lengths to get participation. Typical steps include initiating pre-survey contact requesting cooperation, actual surveying, post survey follow-up if a response is not received, a second survey request, and finally interviews using an alternate modes such as telephone or person to person.

4. Non-Response: Non response errors occur when non-respondents are different than those who respond. This may occur because either the potential respondent was not contacted (they did check their e-mail) or they refused to respond (they were all grumpy old men or beautiful young women afraid of strangers). Again, the extent of this non-response error can be checked through follow-up surveys using alternate modes.
**NON-SYSTEMATIC ERRORS (SAMPLING ERRORS)**

Sampling errors occur because of variation in the number or representativeness of the sample that responds. Sampling errors can be controlled by (1) Careful sample designs, (2) Large samples, and (3) Multiple contacts to assure representative response. Two types of samples may be drawn, a probability sample where every person in the sample has an equal and known probability of being selected, and a non-probability sample where the probability of a person being selected is unknown.

Non-probability samples include *quota samples, referral samples, and convenience samples*. A *quota sample* assures that various subgroups of the population are represented on relevant sample characteristics. The quota sample does not use population proportions. For example a quota sample may be used to make sure you have at least 35 people who have an income more than $250,000. Respondents from group would rarely be interviewed because of their small incidence in the population as a whole.

A *referral sample* is used to locate a population of rare individuals by referral. Locating 100 adult croquet players may be difficult were it not by referral.

*Convenience samples* are based on convenience and may include members of affiliation groups, interest groups, or random intercepts on your website. The objective is to collect as much data as possible, regardless of where they come from. Probability samples include *simple random samples, systematic samples, and stratified samples*.

A *simple random sample* occurs where every element has a known and equal probability of being selected. A true random sample is rarely used because we rarely have a sample frame that lists every person we could sample.

A *systematic sample* occurs all potential respondents have a known and equal chance of being selected. Typically a systematic sample would select every nth person from the list of potential respondents. For example, a systematic sample of 400 customers (out of the 3000 total customers) would be conducted by computing a respondent selection frequency \((N/n) = 3000/400 = 7.5\). Then this number is rounded to 8 and a random number is selected from 1-8 (suppose the result is 3). We would select a systematic sample by first selecting the third customer on the list and then every 8th thereafter. This will form a simple random sample of respondents if the customer list is not systematically ordered in some way.

A *stratified sample* is sometimes desirable if the population is to be broken up into different groups based on one or more characteristics of the population. In this case, the strata are identified. Strata may defined as any groups: Credit card users Vs noncredit card users, by gender, age, industry, purchasers Vs non-purchasers, current customers Vs past customers, etc. Once the strata are identified, a simple random sample is drawn within each strata. Once the survey is completed, the strata are then weighted back to the population proportions.
Calculating Sample Size

How many people do you need to interview to get results representative of the target population with the level of confidence that you are willing to accept? Given an existing sample, what is the level of precision you now have? The answer to these questions can be found using a sample size table and depends on the degree of sampling precision and accuracy you require.

Computing sample size is based on the confidence interval and the confidence level.

ESTIMATING SAMPLE SIZE

Given the sampling methods discussed above, the final sample determination is that how many should be sampled? The most common method of sample size determination is that based on proportions. For example, suppose we are preparing for the winter Olympics and are interested in estimating “the proportion of out of state skiers that took at least one overnight trip.” We might use this number of people that would consider traveling to the Olympics.

In this case, the sample size is estimated using proportions. \( s_p = \sqrt{\frac{p(1-p)}{n}} \) where \( p \) is the proportion of “out of state skiers that took at least one overnight trip”. The most conservative number for this proportion is .50 and if the desired accuracy was .05 and the formula would appear as:

\[
\frac{(Number\ of\ Standard\ Errors)^2 \times (proportion) \times (1-proportion))\ /\ (Accuracy)}{(1+((Number\ of\ Standard\ Errors)^2 \times (proportion) \times (1-proportion))\ /\ (Accuracy)\ -\ 1)\ /\ (the\ proportion\ size)}
\]

While this formula is easily entered into a spreadsheet, this formula results in the following sample size determination table.
**Sample Size Terminology**

The following is an Edited version of “Inside the paper’s election polls”, an article by Elsa McDowell that appeared in The Charleston Post and Courier, November 8, 2002:

The beauty of... election polls is that they are straightforward. They use statistical formulae to estimate how many people will vote one way and how many will vote another. No spin. No qualifying clauses to muddy the picture.

### Table 1: Sample Size Lookup Table for 95% Confidence

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Variability Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50/50%</td>
</tr>
<tr>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>50</td>
<td>14.2</td>
</tr>
<tr>
<td>75</td>
<td>11.5</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>150</td>
<td>8.2</td>
</tr>
<tr>
<td>200</td>
<td>7.1</td>
</tr>
<tr>
<td>250</td>
<td>6.3</td>
</tr>
<tr>
<td>300</td>
<td>5.8</td>
</tr>
<tr>
<td>400</td>
<td>5</td>
</tr>
<tr>
<td>500</td>
<td>4.5</td>
</tr>
<tr>
<td>600</td>
<td>4.1</td>
</tr>
<tr>
<td>800</td>
<td>3.5</td>
</tr>
<tr>
<td>1000</td>
<td>3.1</td>
</tr>
<tr>
<td>1500</td>
<td>2.5</td>
</tr>
<tr>
<td>2000</td>
<td>2.2</td>
</tr>
<tr>
<td>2500</td>
<td>2</td>
</tr>
<tr>
<td>5000</td>
<td>1.4</td>
</tr>
<tr>
<td>10000</td>
<td>1</td>
</tr>
<tr>
<td>15000</td>
<td>0.82</td>
</tr>
<tr>
<td>25000</td>
<td>0.63</td>
</tr>
<tr>
<td>50000</td>
<td>0.4</td>
</tr>
</tbody>
</table>

This table is compute using the following formula:

\[
\frac{\text{((Number of Standard Errors)^2 * (proportion)*(1-proportion)) / (Accuracy)}}{1+(\text{((Number of Standard Errors)^2 * (proportion)*(1-proportion)) / (Accuracy)-1)) / (the population size)}
\]

This formula is easily entered into a spreadsheet, to compute a sample size determination table.
The difficulty of... election polls is that they are not always straightforward. How else could you explain that a poll done by one candidate shows him in the lead and that a poll done by his opponent shows her in the lead? Statisticians say there are ways to twist questions or interpret answers to give one candidate an advantage over another.

One reader took issue with a recent poll results run in The Post and Courier. He questioned whether the methodology was described in enough detail, whether the sample size was adequate.

He was right about one point. The story did not make clear who was polled. It said “voters” and failed to elaborate. It should have indicated that the people polled were registered and likely to vote in the November elections.

His next point is debatable. He said the sample size of 625 likely voters was insufficient for a state with nearly 4 million residents and suggested at least 800 should have been polled.

Brad Coker, the researcher responsible for the study responded that “the standard sample size used by polling groups nationally is 625. It produces, as the story stated, a margin of error of plus-or-minus 4 percent. Increasing the sample size to 800 would have produced a margin of error of plus-or-minus 3.5 - more accurate, but not so much more accurate to justify the additional cost.”

“Many people do not understand how sample sizes work. They believe that, the larger the pool, the larger the sample size needs to be.” “It’s not like that. You can take a drop of blood from a 400-pound person and it will contain the same data you would get if you took it from a 100-pound person,” he said.

The reader’s next concern was that the margin of error of plus-or-minus 4 applies only to the group viewed in its entirety. “If ‘minorities’ constituted 27 percent of the total sample, then only 169 were sampled. The margin of error then skyrocketed into double digits.” Coker said the reader is right and wrong. The margin of error does jump for subgroups, but does not reach double digits. In this case, it moves from plus-or-minus 4 to plus-or-minus 6 to 8.

Two days before The Post and Courier ran the... poll, there was a short story... about a... poll commissioned by MSNBC. The... poll indicated incumbent Gov. Jim Hodges with a 45-43 percent lead. (Our) poll indicated challenger Mark Sanford was ahead 45 to 41 percent. When the margin of error is considered, both polls show the race is still a toss-up.

**TERMINOLOGY**

The confidence interval is the plus-or-minus figure that represents the accuracy of the reported statistic. Consider the following example:

A Canadian national sample showed “Who Canadians spend their money on for Mother’s Day.” Eighty-two percent of Canadians expect to buy gifts for their mom, compared to 20 percent for their wife and 15 percent for their mother-in-law. In terms of spending, Canadians expect to spend $93 on their wife this Mother’s Day versus $58 on their mother. The national findings are accurate, plus or minus 2.75 percent, 19 times out of 20.
For example, if you use a confidence interval of 2.75 and 82% percent of your sample indicates they will “buy a gift for mom” you can be “confident (95% or 99%)” that if you had asked the question to ALL CANADIANS, somewhere between 79.25% (82%-2.75%) and 84.75% (82%+2.75%) would have picked that answer.

The confidence level tells you how confident you are of this result. It is expressed as a percentage of times that different samples (if repeated samples were drawn) would produce this result. The 95% confidence level means that 19 times out of twenty that results would fall in this - + interval confidence interval. The 95% confidence level is the most commonly used.

When you put the confidence level and the confidence interval together, you can say that you are 95% (19 out of 20) sure that the true percentage of the population that will “buy a gift for mom” is between 79.25% and 84.75%.

Wider confidence intervals increase the certainty that the true answer is within the range specified. These wider confidence intervals come from smaller sample sizes. When the cost of an error is extremely high (a multi-million dollar decision is at stake) the confidence interval should be kept small. Increasing the sample size can do this.

What Influences the Size of the Confidence Intervals

Sampling theory teaches us that the accuracy of our estimates are dependent on such factors as the dispersion and skewness of the population’s responses (divergent opinions or characteristics verses similar opinions or characteristics), the sample size, and the size of the population. Controlling these variables contributes to the incidence (and elimination) of sampling error. Note that other “non-sampling” error, such as bad question design and selection of a “bad” sample list is not controlled by sample size.

EFFECTS OF A LARGER SAMPLE SIZE ON CONFIDENCE INTERVAL

Larger sample sizes generally produce a more accurate picture of the true characteristics of the population. Larger samples tighten the size of the confidence interval, making your estimate much more accurate. However, this relationship is not linear. Increasing sample size from 500 to 1000 reduces the confidence interval from +/-.4.38 to +/- 3.1.

EFFECTS OF DISPERSION (PROPORTION) ON ACCURACY

The accuracy of an estimate also depends on the dispersion and skewness of the population on the question being asked. A sample of Republicans would give a less dispersed evaluation of a Republican president than would a sample of Democrats. Likewise, a sample of Catholic priests would have less variability on the issue of abortion than would a survey of the general population. Accuracy of the sample estimate increases as the dispersion in the population decreases.
To cover all eventualities, you should use a worse case scenario (50%) as the percentage when estimating confidence intervals and required sample size. Once your data has been collected, you can use the observed proportion in a specific question to obtain a more accurate estimate of the confidence interval.

**THE EFFECT OF POPULATION SIZE ON SAMPLE SIZE**

The size of the population also influences the size of the confidence, but not as much as you might expect. If your survey of 1000 respondents were from a population of 100,000, the confidence interval is +- 3.08%, while the confidence interval widens to only +- 3.1% when sampled from a population of 1,000,000. The size of the population has little effect on the confidence interval relative to the effect of the actual sample size.

*Non-Sampling Errors.* Non-sampling errors cannot be compensated for by increased sample size. Often, larger samples accentuate non-sampling errors rather than reduce them. Non-sampling errors come from samples that are not truly random, bad scales, misleading questions, incomplete surveys and so on.

**We’re Here to Help!**

Qualtrics.com provides the most advanced online survey building, data collection (via panels or corporate / personal contacts), real-time view of survey results, and advanced “dashboard reporting tools”.

If you are interested in learning more about how the Qualtrics professional services team can help you with a conjoint analysis research project, contact us at research@qualtrics.com.