MarketVision RESEARCH

March 2017

SAMPLING ERROR & SURVEY RELIABILITY

Frequently, researchers wish to make some statement about how reliable a particular piece of research is in guiding business decisions. To do this, we often rely on the error that is attributable to sampling. Since we are not interviewing all potential respondents, there is some chance that our results (from a sample) would not match those obtained had we interviewed all potential respondents (the universe).

The calculation of this sampling error is very easy for questions dealing with proportions of respondents.[1] The sampling tolerance (error associated to sampling) is calculated using the common formula of:

$$\pm t \times \sqrt{\frac{pq}{n}}$$

In this formula, t corresponds to the t-statistic, which is determined by the confidence level at which you will test the significance of the difference. Typically, significance testing is conducted at the 95% confidence level. The corresponding t-statistic is 1.96. The value p represents the proportion of respondents who give a particular response and q represents the proportion of respondents who don't give that response (q can also be expressed as 1 - p). Finally, n represents the sample size.

Let's consider a simple example of 300 respondents. Assume 60% of the respondents say, "yes" to a particular question. Let's determine the sampling error for tests of significance at 95% confidence. Here n = 300, p = 0.6, q = 0.4 (q = 1 – p or q = 1 - 0.6 = 0.4), and t = 1.96. Therefore, it follows that the sampling tolerance (at 95% confidence) is:

$$\pm 1.96 \times \sqrt{\frac{0.6 \times 0.4}{300}}$$

$$\pm 1.96 \times \sqrt{\frac{0.24}{300}}$$

$$\pm 1.96 \times \sqrt{0.0008}$$

$$\pm 1.96 \times 0.02828$$

$$\pm 0.055$$

$$\pm 5.5\%$$

The confidence interval (or sampling tolerance or sampling error) on this proportion is 5.5%. The interpretation of this number is as follows:

If we were to complete this study 100 times using the same techniques and methods, in 95 of the 100 cases the true universe proportion would be within 5.5 percentage points of the proportion that we determine from the sample.

[1]The calculation of sampling error around a mean of numeric responses can be done as well. To calculate the sampling tolerance for a mean, the term sqrt(pq/n) is replaced with the standard error of the mean and is multiplied by the t-value as shown above.

We say "95 out of 100 times" because of the confidence level that we selected;[2] here we used 95%. We could have used 90%, or 99%, or even 99.9%. It does not mean that we are 95% confident that the true universe proportion is within 5.5 percentage points of the sample proportion.

The t-statistics for various confidence levels are shown below.

Confidence Level	T-Statistic
80%	1.282
90%	1.645
95%	1.960
99%	2.576

The 95% confidence level is commonly used, and many researchers round the t-statistic to 2.0 for simplicity.

Recall the formula presented above for the sampling error:

$$\pm t \times \sqrt{\frac{pq}{n}}$$

Note that the sampling error is based on three measures:

- The confidence level (the t-statistic)
- The variability of respondents' answers (the p term)
- The sample size (n)

The variability of respondents' answers refers to how much respondents agree on their responses, that is, how high or low the *p* term is. For example, consider just the *pq* term from the previous formula for various values of *p.*

p	q	pq
.1	.9	.09
.2	.8	.16
.3	.7	.21
.4	.6	.24
.5	.5	.25
.6	.4	.24
.7	.3	.21
.8	.2	.16
.9	.1	.09

We can see that the largest *pq* term occurs when the proportion (p) is exactly half (0.5) and changes only a little between 0.3 and 0.7.

We can use this information to form a worst-case scenario for any sample size and level of confidence; that is, when half of the respondents say "yes" to a question and half say "no." The following table shows sampling errors for a number of such worst-case scenarios, when p = 0.5: *p.*

Sampling Error of a Proportion		
	Confidence level	
n	*90%*	*95%*
60	10.6%	12.7%
80	9.2%	11.0%
100	8.2%	9.8%
150	6.7%	8.0%
200	5.8%	6.9%
250	5.2%	6.2%
300	4.7%	5.7%
400	4.1%	4.9%
500	3.7%	4.4%
600	3.4%	4.0%
700	3.1%	3.7%
800	2.9%	3.5%
900	2.7%	3.3%
1000	2.6%	3.1%
2000	1.8%	2.2%
2500	1.6%	2.0%
5000	1.2%	1.4%

[2]When testing at the 95% confidence level, there is a 5% chance that we will find a significant difference that doesn't truly exist. This chance, or the alpha level (α) is the complement of the confidence level, and is also referred to as the probability of making a Type I error. Not finding a significant difference that truly does exist is referred to as a Type II error, and its probability s expressed as the beta level (β), which is the complement of the power of the significance test.

Note that the confidence interval (or sampling error) decreases at a much slower rate than the sample size increases. Recall that the formula includes the square root of the sample size, so a doubling of the sample size will decrease the sampling tolerance by a factor of the square root of 2. Because of this square root relationship, a sample size must be quadrupled to reduce the sampling error by half.

Put another way, doubling the sample size decreases the sampling tolerance to 70% of the previous tolerance (= 1/1.41 = 0.707).

Assumptions Behind These Calculations

While these calculations are relatively straightforward, there are a number of important assumptions to remember when applying this formula to our work. These are outlined below.

Reasonably Large Sample Sizes

This formula should only be applied to samples with a minimum base of 60, and bases over 100 are better. Do not use this formula with small samples sizes.

Proportion (p) not near 0 or 1

This formula works well when the p term is roughly between 0.15 and 0.85. When p is very near zero or very near one, different formulae are more appropriate.

Perfect Random Sampling

The usage of this formula also depends on the sampling design. First, the formula assumes a simple random sample. If the sample used had quotas or weighting, this formula will understate the confidence interval[3] — at times by a factor of 2 or more. Second, this formula assumes a 100% response rate. While this second part of the assumption is generally ignored, it is important to keep in mind in cases when the refusals are particularly high or when the refusals are less likely to be random.

Large Universe

This formula assumes that the universe is sufficiently large, and that the sampling fraction (the sample size divided by the universe size) is small, typically under 8%. As the sampling fraction gets larger, the risk attributed to sampling is decreased, expressed as a smaller confidence interval.[4]

One Sample, Two Sample

So far we have been talking about the sampling errors associated with the difference between our sample results and the true (but unknown) proportion in the universe. This difference is very useful in fields of political polling or sampling to determine an incidence, as the government frequently reports. In marketing research, we sometimes care about the incidence of brand users or the proportion of people likely to purchase a specific new product. More often we use significance testing to identify if two groups are significantly different from each other.

The test used to determine if two sample segments are statistically different is referred to as a two-sample test.[5] The formula shown above is a one-sample test. The two-sample test is similar to a one-sample test but takes into consideration that both sample segments have sampling error.

The formula for this test is shown below.

$$\pm t \times \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

[3]This inflation of the variance is often referred to as the design effect, or sometimes just deff.

[4]Such adjustments fall under the heading of "finite population correction" factors.

[5]It is important to note that a two-sample test is different than a two-tail test. We won't discuss a two-tail versus one-tail test here, but just know that they are not the same thing.

Here, p, q, and n each have the same meanings they did above, but there is now a value for each of the two sample segments under consideration, denoted by subscripts.

Let's consider the following example. In a sample of lottery players we find that 60% of males are frequent players and 52% of females are frequent players. Can we say that men are significantly more likely to be frequent players than females? Our sample included 300 men and 250 women. Let's test at the 95% confidence interval. The difference in proportions required for a significant difference is calculated as follows:

$$\pm 1.96 \times \sqrt{\frac{0.6 \times 0.4}{300} + \frac{0.52 \times 0.48}{250}}$$

$$\pm 1.96 \times \sqrt{\frac{0.24}{300} + \frac{0.2496}{250}}$$

$$\pm 1.96 \times \sqrt{0.0008 + 0.0009984}$$

$$\pm 1.96 \times \sqrt{0.0017984}$$

$$\pm 1.96 \times 0.0424075$$

$$\pm 0.083$$

$$\pm 8.3\%$$

To determine if this is a significant difference, we must determine if the difference between our samples is larger than the difference required. The observed difference is 8% (60% - 52%), which is not larger than the required difference, so the two groups are not significantly different.

When the two proportions (p) are similar and the two sample sizes (n) are similar, the difference required for two proportions to be significantly different is approximately 1.4 times larger than the one sample test. The following table shows the difference required for a statistically significant difference between two sample segments when $p = 0.5$ and the sample sizes in the two cells are equal. The assumptions for the one-sample test also apply to the two-sample test. In addition, one other assumption must be stated.

Independent Samples

This test of differences is only appropriate when there are two groups and no respondent is in both groups. For example, if we have one group of all lottery players and one group of frequent lottery players, it is wrong to test if these two groups are significantly different because the frequent players are a subset of all players. In effect, you would be testing if someone were different from himself or herself.

Sample Size Determination

So far we have discussed using these formulae only to analyze data. We can also use these formulae to help determine sample sizes. Assume that we want to determine a sample size in which a difference of 6 percentage points between two groups would be significant at the 95% confidence

Difference Required for Statistical Significance Two Sample Test

Confidence Level			Confidence Level		
n	***90%***	***95%***	***n***	***90%***	***95%***
100	11.6%	13.9%	800	4.1%	0%
200	8.2%	9.8%	900	3.9%	0%
300	6.7%	8.0%	1000	3.7%	0%
400	5.8%	6.9%	1500	3.0%	0%
500	5.2%	6.2%	2000	2.6%	0%
600	4.7%	5.7%	2500	2.3%	0%
700	4.4%	5.2%	5000	1.6%	0%

level. We can use the following formula (derived from the previous one) to determine the sample size per sample segment.

$$n = \sqrt{\frac{2*t^2*pq}{a^2}}$$

These terms each have the same meaning as above. The new term, *d,* represents the difference (expressed as a decimal) that should be considered significant. Again, the *pq* term is frequently set to $p = 0.5$, $q = 0.5$, $pq = 0.25$, as a worst-case scenario. Therefore, we can calculate the required sample size as shown below.

$$n = \sqrt{\frac{2*1.96*p*q}{0.06*0.06}}; \quad n = \frac{1.9208}{0.0036}; \quad n = 534$$

Therefore, with a sample size of 534 in each sample segment, a difference between segments of 6 percentage points would always be significant.[6] Note that this represents a two-sample test. To determine a sample size for a one sample test, simply exclude the "2" from the numerator.

While this is a powerful tool, it is probably myopic to determine sample sizes for a research study on this criterion alone. Other considerations include analyses to be performed, number of segments likely to be investigated, and type of information to be collected.

Sampling versus Non-Sampling Errors

Now, the remaining question is what do researchers mean when they say reliable? And what does it mean when someone says they would like their research to have a statistically reliable sample?

To answer this we actually leave the world of sampling statistics. Instead, we focus on issues of questionnaire writing. Questions are said to produce reliable results if on repeated administration they produce the same results; reliability has virtually nothing to do with sampling theory.

An example of a very reliable question is gender. Questions that produce less reliable results include income (people lie), questions dealing with long term memory (people forget), and questions that are uninteresting or are asked in too much detail (people don't care). Question order also has a major impact on the reliability of a question. That is why in tracking studies, even changing one question early in the survey can erode a researcher's ability to make meaningful comparisons.

There is another point to consider when talking about reliability; reliability indicates that a respondent will give the same answer on repeated administration. It doesn't indicate that he or she gives the correct answer—that measurement is referred to as validity.[7] A valid measure must be reliable but a reliable measure isn't necessarily valid.

So then, why do sampling discussions include reliability so often?

When respondents err answering a question, they can do so either at random or systematically. If they do so systematically, there is a consistent directional bias. For example, some respondents tend to overstate their income. Such a response would represent a systematic bias, and huge sample sizes will not correct this bias. However, if respondents err at random when answering questions, with large sample sizes the random errors cancel each other out and our sample estimate of the proportion is unbiased, even though our variance is increased.

Unfortunately, there are no common formulae to account for this, and it is largely an academic issue. Therefore, o produce reliable research, the key areas of focus must include questionnaire design and administration in addition to sample size.

[6] There can be occasions in which after conducting this calculation to determine the sample size, it turns out that we would be talking to a large portion of the total universe of potential respondents. In these instances the finite population correction factors are employed.

[7] Some prefer to use the terms accuracy and precision, where precision represents consistency over repeated administrations and accuracy represents the difference between the actual value and the value determined by the research.