

## Chapter 7

# Sample Size Issues for Conjoint Analysis

“I’m about to conduct a conjoint analysis study. How large a sample size do I need? What will be the margin of error of my estimates if I use a sample of only 100 respondents?” These are common questions. Unfortunately, they are difficult questions to answer because many issues come into play:

- What is it exactly that you are trying to measure to get a statistically significant result: a specific part-worth, preference for a product, or the difference in preference between groups of people?
- Do you expect that the differences between features/products/groups you are trying to detect are subtle or strong?
- What level of certainty do you need to be able to act upon your conclusions: 99% confidence, 90% confidence, or what?
- How large is the total population in the market for your product?
- What conjoint methodology do you plan to use? How many conjoint questions will each respondent answer?
- Do you need to compare subsets of respondents, or are you going to be looking at results only as a whole?
- How homogenous is your market? Do people tend to think alike, or are there strong differences in preferences among individuals?
- How do you plan to select your sample? Will it be a random sample or convenience sample?
- How large of a budget do you have for the project?

Answers to these questions play a role in determining the appropriate sample size for a conjoint study. This chapter provides advice and tools to help conjoint researchers make sample size decisions. It involves more statistical theory and formulas than other chapters, so please bear with me.

Though most of the principles that influence sample size determination are based on statistics, successful researchers develop heuristics for quickly determining sample sizes based on experience, rules-of-thumb, and budget constraints. Let us begin our discussion by making a distinction between sampling and measurement error. Subsequent sections will discuss each of these sources of error.

## 7.1 Sampling Error versus Measurement Error

Errors are deviations from truth. In marketing research we are always concerned with reducing error in cost-effective ways. Assuming that you have selected the appropriate modeling method, there are two main sources of error that cause preference data to deviate from truth. The first is sampling error.

Sampling error occurs when samples of respondents deviate from the underlying population. If we have drawn a random sample (each population element has an equal probability of being selected), sampling error is due to chance. If, on the other hand, our sample is not random (for example, a convenience sample), the sampling errors may be systematic. With random sampling, we reduce sampling error by simply increasing the sample size. With nonrandom sampling, however, there is no guarantee that increasing sample size will make the samples more representative of the population.

To illustrate sampling error, assume we wanted to figure out how far the average adult can throw a baseball. If we drew a random sample of thirty people, and by chance happened to include Ichiro Suzuki (outfielder for the Seattle Mariners), our estimate would likely be farther than the true distance for the average adult. It is important to note that the samples we use in marketing research are rarely random. Some respondents resist being interviewed and, by selecting themselves out of our study, are a source of nonresponse bias.

A second source of error in conjoint data is measurement error. We reduce measurement error by having more or better data from each respondent. Consider again the example of the baseball toss. Suppose you are one of the study participants. You throw the ball, but you accidentally step into an uneven spot on the ground, and the ball does not go as far as you typically could throw it. If we asked you to take another few tosses, and averaged the results, we would reduce the measurement error and get a better idea of how far you could throw a baseball.

In conjoint analysis, we reduce measurement error by including more conjoint questions. We recognize, however, that respondents get tired, and there is a limit beyond which we can no longer get reliable responses, and therefore a limit to the amount we can reduce measurement error.

## 7.2 Binary Variables and Proportions

Sampling error is expressed in terms of standard errors, confidence intervals, and margins of error. We can begin to understand what these terms mean by considering binary variables and proportions. In fact, we will spend a good deal of time talking about confidence intervals for proportions because the statistical principles can be applied to choice-based conjoint results and shares of choice in market simulations for all conjoint techniques.

A binary variable is a categorical variable with exactly two levels, such as a yes/no item on a consumer survey or a true/false checklist item. Many product attributes in conjoint studies have exactly two levels. And consumer choice itself is binary—to choose or not, to buy or not. Binary variables are usually coded as 1 for yes and 0 for no. Looking across a set of binary variables, we see a set of 1s and 0s. We can count the number of 1s, and we can compute the proportion of 1s, which is the number of 1s divided by the sample size  $n$ .

In statistical theory, the sampling distribution of the proportion is obtained by taking repeated random samples from the population and computing the proportion for each sample. The standard error of the proportion is the standard deviation of these proportions across the repeated samples. The standard error of a proportion is given by the following formula:

$$\text{standard error of a proportion} = \sqrt{\frac{pq}{(n-1)}}$$

where  $p$  is the sample estimate of the proportion in the population,  $q = (1 - p)$ , and  $n$  is the sample size.

Most of us are familiar with the practice of reporting the results of opinion polls. Typically, a report may say something like this: “If the election were held today, Mike Jackson is projected to capture 50 percent of the vote. The survey was conducted by the XYZ company and has a margin of error of  $\pm 3$  percent.” What is margin of error?

Margin of error refers to the upper and lower limits of a confidence interval. If we use what is known as the normal approximation to the binomial, we can obtain upper and lower limits of the 95% confidence interval for the proportion as

$$\text{margin of error for a proportion} = \pm 1.96 \sqrt{\frac{pq}{(n-1)}}$$

Going back to the polling report from XYZ company, we note that margin of error has a technical meaning in classical statistics. If XYZ were to repeat the poll a large number of times (with a different random sample each time), 95 percent of the confidence intervals associated with these samples would contain the true proportion in the population. But, of course, 5 percent of the confidence intervals would not contain the true proportion in the population. Confidence intervals are random intervals. Their upper and lower limits vary from one sample to the next.

Suppose we interview 500 respondents and ask whether they approve of the president's job performance, and suppose 65 percent say yes. What would be the margin of error of this statistic? We would compute the interval as follows:

$$\pm 1.96 \sqrt{\frac{(0.65)(0.35)}{(500 - 1)}} = \pm 0.042$$

The margin of error is  $\pm 4.2$  percent for a confidence interval from 60.8 to 69.2 percent. We expect 95 percent of the confidence intervals constructed in this way to contain the true value of the population proportion.

Note that the standard error of the proportion varies with the size of the population proportion. So when there is agreement among people about a yes/no question on a survey, the value of  $p$  is closer to one or zero, and the standard error of the proportion is small. When there is disagreement, the value of  $p$  is closer to 0.50, and the standard error of the proportion is large. For any given sample size  $n$ , the largest value for the standard error occurs when  $p = 0.50$ .

When computing confidence intervals for proportions, then, the most conservative approach is to assume that the value of the population proportion is 0.50. That is, for any given sample size and confidence interval type,  $p = 0.50$  will provide the largest standard error and the widest margin of error. Binary variables and proportions have this special property—for any given sample size  $n$  and confidence interval type, we know the maximum margin of error before we collect the data. The same cannot be said for continuous variables, which we discuss in the next section.

### 7.3 Continuous Variables and Means

With continuous variables (ratings-based responses to conjoint profiles), one cannot estimate the standard error before fielding a study. The standard error of the mean is directly related to the standard deviation of the continuous variable, which differs from one study to the next and from one survey question to the next. Assuming a normal distribution, the standard error of the mean is given by

$$\text{standard error of the mean} = \frac{\text{standard deviation}}{\sqrt{n}}$$

And the margin of error associated with a 95% confidence interval for the mean is given by

$$\text{margin of error for the mean} = \pm 1.96(\text{standard error of the mean})$$

Suppose we had conducted an ACA study with forty respondent interviews. We want to estimate purchase likelihood for a client's planned product introduction with a margin of error of  $\pm 3$  and a 95% confidence level. We run an ACA market simulation to estimate purchase likelihood on a 100-point scale, and the simulator reports the standard error next to the purchase likelihood estimate:

*Total Respondents = 40*

	<i>Purchase Likelihood</i>	<i>Standard Error</i>
Product A	78.34	3.06

The margin of error is  $\pm 1.96 \times 3.06 = \pm 6.00$ , so we need to cut the margin of error in half to achieve our  $\pm 3$  target level of precision. We know that the standard error of the mean is equal to the standard deviation divided by the square-root of the sample size. To decrease the standard error by a factor of two, we must increase sample size by a factor of four. Therefore, we need to interview about  $40 \times 4 = 160$  or 120 additional respondents to obtain a margin of error of  $\pm 3$  for purchase likelihood.

#### 7.4 Small Populations and the Finite Population Correction

The examples we have presented thus far have assumed infinite or very large populations. But suppose that, instead of estimating the job performance rating of the president by the United States population at large, we wanted to estimate (with a margin of error of  $\pm 3$  percent) the job performance rating of a school principal by members of the PTA. Suppose there are only 100 members of the PTA. How many PTA members do we need to interview to achieve a margin of error of  $\pm 3$  percent for our estimate?

First, we introduce a new term: finite population correction. The formula for the finite population correction is  $\frac{(N-n)}{(N-1)}$ , where  $n$  is the sample size and  $N$  is the population size. The formula for the finite population correction is often simplified to  $(1 - f)$ , where  $f = \frac{n}{N}$ , which is approximately equivalent to  $\frac{(N-n)}{(N-1)}$  for all except the smallest of populations.

After a population reaches about 5,000 individuals, one can generally ignore the finite population correction factor because it has a very small impact on sample size decisions. Using the simplified finite population correction for a finite sample, the margin of error for a proportion and a 95% confidence interval is equal to

$$\pm 1.96 \sqrt{(1 - f) \frac{pq}{(n - 1)}}$$

The finite population correction may also be used for continuous variables and means.

With a population of 100, we can solve for  $n$  assuming an expected proportion. The worst-case scenario (i.e., the one that has the largest standard error) is for a 0.50 proportion, so it is standard to let  $p = 0.50$ . Solving for  $n$ , we discover that we would need to interview 92 PTA members, or 92 percent of the population to achieve a margin of error of  $\pm 3$  percent.

The important point to be made is that with small populations, you may have to interview a significant proportion of the population to achieve stable estimates. Suppose your client produces a very expensive, highly specialized piece of machinery, for which there were only 100 total potential customers in the world. Given many people's unwillingness to complete surveys, it will likely be much more difficult to complete surveys with 92 out of 100 potential buyers of this product than to interview, say, 1,000 potential buyers of something like office chairs, for which there are so many buyers as to approximate an infinite population. Even so, in terms of estimating a proportion, both scenarios lead to the same margin of error when projecting to the population of interest.

Conjoint studies may be used for large or small populations. We can use conjoint analysis for even the smallest of populations, provided we interview enough respondents to represent the population adequately.

## 7.5 Measurement Error in Conjoint Studies

Many researchers and dozens of data sets have demonstrated that conjoint utilities do a good job of predicting individual respondents' preferences for products. Holdout choice sets (choice tasks not used to estimate utilities) are often included in conjoint questionnaires. Using the conjoint data, a respondent's holdout choices usually can be predicted with a hit rate of roughly 75 to 85 percent. These choice tasks typically include between three and five different product concepts, so by chance we expect a success rate between 20 and 33 percent.

The hit rates with conjoint are significantly greater than chance and significantly better than the marketer's best guesses—even if the marketer knows each customer very well. In fact, conjoint predictions at the individual level frequently approach or sometimes even exceed test-retest reliability, suggesting that a good set of conjoint utilities is about as reliable at predicting choices to repeated holdout tasks as the respondents' earlier choices.

If there were only one buyer of your product in the world, you could learn a great deal about that individual's preferences from a conjoint interview. The utility data would be reasonably accurate for predicting his or her preferences and weights placed upon attributes. We can learn a great deal about an individual respondent provided we ask that respondent the right questions and enough questions. Let us consider numbers of conjoint questions or tasks needed for alternative methods of conjoint analysis.

### Adaptive Conjoint Analysis

An Adaptive Conjoint Analysis (ACA) interview results in a set of utilities for each individual. We want conjoint measurements for each individual in the study to be as accurate as possible.

Of the three conjoint methods discussed in this chapter, ACA is the best at reducing measurement error. ACA's interviews adapt to the respondent, asking questions designed to be maximally relevant and efficient for refining utility estimates. The priors section helps in stabilizing the utility estimates at the individual level. One sees fewer reversals in part-worths (out-of-order utilities) for ordered attributes like price in ACA than in traditional conjoint and choice-based conjoint with individual estimation.

In ACA, one needs to decide how many pairs questions to ask. The number of pairs each respondent completes plays a significant role in reducing measurement error. The suggested number of pairs is  $3(K - k - 1) - K$ , where  $K$  is the total number of levels across all attributes and  $k$  is number of attributes. If respondents answer as many pairs as suggested, a total of three times the number of observations as parameters are available at the individual level for computing utilities (this includes information from the self-explicated priors). Sometimes the suggested number of pairs is greater than respondents can reasonably do. You should make sure not to overburden respondents because this can lead to poor results. You can ask fewer than the recommended number of pairs, though this increases the measurement error for each individual.

If your sample size is particularly small and the number of attributes to measure is large, ACA may be the best tool to use. In fact, it is possible to have an entire research study designed to learn about the preferences of one respondent, such as an important buyer of an expensive industrial product. As we discussed in chapter 5, there are many considerations for determining whether ACA is appropriate for a study. For further discussion of ACA measurement, estimation, and sample size issues, see Johnson (1987a).

### Traditional Conjoint Studies

Like ACA, traditional full-profile conjoint (such as Sawtooth Software's CVA or SPSS's conjoint module) usually leads to the estimation of individual-level part-worth utilities. Again, the minimum sample size is one. But, because the traditional conjoint methodology does not include a self-explicated priors section, its utilities tend to have greater variability (larger standard errors) at the individual level relative to ACA (holding respondent effort equal).

One should include enough conjoint questions or cards to reduce measurement error sufficiently. Sawtooth Software's CVA manual suggests asking enough questions to obtain three times the number of observations as parameters to be estimated, or a number equal to  $3(K - k + 1)$ , where  $K$  is the total number of levels across all attributes and  $k$  is the number of attributes.

Respondents sometimes lack the energy or patience to answer many questions. We need to strike a good balance between overworking the respondent (and getting noisy data) and not asking enough questions to stabilize the estimates.

### Choice-Based Conjoint

Though generally considered more realistic than traditional conjoint, choice-based questions are a relatively inefficient way to learn about preferences. As a result, sample sizes are typically larger than with ACA or traditional ratings-based conjoint, and choice-based conjoint (CBC) results have traditionally been analyzed by aggregating respondents. Lately, hierarchical Bayes has permitted individual-level estimation of part-worth utilities from CBC data. But to compute individual-level models, HB uses information from many respondents to refine the utility estimates for each individual. Therefore, one usually does not calculate utilities using a sample size of one. It should be noted, however, that logit analysis can be run at the individual level, if the number of parameters to be estimated is small, the design is highly efficient, and the number of tasks is large.

There are rules-of-thumb for determining sample sizes for CBC if we are willing to assume aggregate estimation of effects. Like proportions, choices reflect binary data, and the rules for computing confidence intervals for proportions are well defined and known prior to collecting data.

Consider a design with three brands and three prices. Assume each person completes ten tasks, and each task displays three products (i.e., each brand and price occurs once per task). If we interview 100 respondents, each brand will have been available for choice

$$(100 \text{ respondents}) \times (10 \text{ tasks}) \times \frac{(3 \text{ concepts})}{(3 \text{ brands})} = 1,000 \text{ times}$$

Johnson and Orme (1996) looked at about twenty commercial choice-based conjoint data sets and determined that having each respondent complete ten tasks is about as good at reducing error as having ten times as many respondents complete one task. Of course, in the limit this suggestion is ridiculous. It does not make sense to say that having one respondent complete 1,000 tasks is as good as having 1,000 respondents complete one task. But, according to Johnson and Orme (1996) simulation results, if a researcher obtains data from three to four hundred respondents, doubling the number of tasks they complete is about as good (in terms of reducing overall error) as doubling the sample size. It makes sense from a cost-benefit standpoint, then, to have respondents complete many choice tasks.

Johnson, who is the author of Sawtooth Software's CBC System, has recommended a rule-of-thumb when determining minimum sample sizes for aggregate-level full-profile CBC modeling: set

$$\frac{nta}{c} \geq 500$$

where  $n$  is the number of respondents,  $t$  is the number of tasks,  $a$  is number of alternatives per task (not including the *none* alternative), and  $c$  is the number of analysis cells. When considering main effects,  $c$  is equal to the largest number of levels for any one attribute. If you are also considering all two-way interactions,  $c$  is equal to the largest product of levels of any two attributes (Johnson and Orme 2003).

Over the years, we have become concerned that practitioners use Johnson's rule-of-thumb to justify sample sizes that are too small. Some feel that they will have ample stability in estimates when each main-effect level of interest is represented across the design about 500 times. But 500 was intended to be a minimum threshold when researchers cannot afford to do better. It would be better, when possible, to have 1,000 or more representations per main-effect level.

## 7.6 Typical Sample Sizes and Practical Guidelines

The recommendations below assume infinite or very large populations. They are based on the theories above and our observations of common practices in the market research community:

- Sample sizes for conjoint studies generally range from about 150 to 1,200 respondents.
- If the purpose of your research is to compare groups of respondents and detect significant differences, you should use a large enough sample size to accommodate a minimum of about 200 per group. Therefore, if you are conducting a segmentation study and plan to divide respondents into as many as four groups (i.e., through cluster analysis) it would be wise to include, at a minimum,  $4 \times 200 = 800$  respondents. This, of course, assumes your final group sizes will be about equal, so one would usually want more data. The stronger segmentation studies include about 800 or more respondents.
- For robust quantitative research where one does not intend to compare subgroups, I would recommend at least 300 respondents. For investigational work and developing hypotheses about a market, between thirty and sixty respondents may do.

These suggestions have to be weighed against research costs. There are difficult decisions to be made based on experience, the application of statistical principles, and sound judgment. If, after the fact, you find yourself questioning whether you really needed to have collected such a large sample size for a particular project, it is an interesting exercise to delete a random subset of the data to see how having fewer respondents would have affected your findings.

A thorough discussion of sampling and measurement errors would require more time and many more pages. The reader is encouraged to consult other sources in these areas. For statistics and sampling see Snedecor and Cochran (1989) and Levy and Lemeshow (1999). For measurement theory see Nunnally (1967).