Sampling Basics

Principles and Procedures

The results from a survey are used to describe the characteristics of individuals in a specific finite group. In cases where the population is small, it may be prudent to survey the entire group (i.e, a census not a sample). However, when large groups are targeted, a census may not be possible and likely are not needed. Careful sampling the population allows us to make fairly accurate generalizations about the population from which the sample comes. The degree to which the generalizations being made are valid largely depends on the sampling methods used and the response rate. Basically, the validity of the conclusions arrived at will depend on whether the sample used is a representative sample, meaning those responding to the survey are a reasonable representation of the population targeted in the study.

To ensure full transparency when conducting survey research, best practice requires that you always report the sample size, the response rate, the way in which samples were selected, and if known, the population size or at least an estimate of what the population is assumed to be. In addition, you need to report issues that might have occurred in terms of any systematic response refusal patterns.

Prior to describing various techniques used to select a sample, we should first explore various issues related to sampling and how to determine an appropriate sample size for a survey research study.

Measurement Error Related to Sampling.

With regards to sampling, there are generally two ways that measurement errors can occur. The primary problem is sampling errors but response refusal issues can also increase measurement error. You will likely never know the degree to which these errors have affected a study. The primary way to alleviate these problems is to increase the size of a sample. Still, you should always assume there will be some degree of measurement error.

Sampling error occurs when those chosen to take a survey do not adequately mirror (represent) those in the population from which the sample was drawn. Fundamentally this is an issue because the results over or under represent various subgroups within the population. When this happens, the results obtained from the survey are not generalizable and are consider to be are flawed (biased). This can easily happen with small samples especially when there are several smaller subgroups within the population; however, survey error can also be the result of the sampling techniques used.

Response Refusal can be another problem. Even when efforts to adequately sample the population are made, you cannot force individuals to respond to the survey. When a large number of potential respondents choose not to complete a survey the results may not be generalizable (i.e., they may not accurately represent those in the population). The degree to which response refusal affects generalizability depends on whether the refusal pattern is random or systematic. When potential respondents do not complete the survey, you need to determine whether the response refusal is spread equally across the sample regardless of an individual's characteristics or subgroup membership. This would mean the response refusal is random in which case the sample will continue to be fairly representative of the population. However, if the response refusal is systematic, meaning one group of individuals with similar characteristics are more likely not to complete the survey compared to other groups of individuals, then you have a problem.

Oversampling as a solution. One common solution to both these problems is to increase the sample size by oversampling the population. With random response refusal and sampling error in general, oversampling may help solve the problem. However, increasing the sample size will not fix the problem of systematic response refusal. For this reason, you should always examine the characteristics of respondents who choose not to take the survey to determine if any discernable pattern can be found and to make sure there is no systematic response refusal issues. This may not be possible if you do not have access to information about non-respondents.

It is generally a good idea, at least in theory, to use as large a sample as possible. This will maximize the likelihood of obtaining good representation for the general population and subgroups within that population. In practice, if you are studying a small finite population you may need to invite all to participate. However, if the population is large, you may not be able to survey everyone and you likely don't need to. One argument for not increasing the sample size is cost. However, with the prevalence of online surveys, cost is often not an issue. Access can be a challenge and a good reason for not surveying the entire population; for example, when the population includes younger children. Perhaps the best reason for not obtaining a larger sample is that it simply is not needed. At some point, surveying additional individuals would not change the result and oversampling may lead to survey fatigue (and decreased response rates). If those in a particular population are inundated with invitations to complete surveys, they are less likely to respond. This can negatively affect everyone trying to obtain survey data from that population. Survey fatigue associated with receiving frequent invitations to complete surveys is a serious concern for researchers.

Incentives and Compensation. One way some choose to alleviate the survey error and response refusal problems, one that may also reduce systematic response refusal issues, is to offer respondents an incentive to complete the survey. This can be done by offering payment or some other form of compensation. Some surveys offer the opportunity to win a reward (i.e., offering those who complete the survey entry into a prize draw). This has been found to be effective in some cases but not all. The degree to which compensation will be effective depends on how enticing the incentive is for the participants and the integrity of the individuals. In truth, offering incentives in not always effective and may have some unintended negative consequences. Certainly, adequately compensating people for the time and effort it takes to complete a survey may increase response rates and reduce systematic error. However, while more respondents may complete the survey, several might do so without any intention of answering items thoughtfully or accurately; they simply want the compensation. Likewise, if the incentive is somehow coercive (individuals are forced to complete the survey), not only does this potentially violate ethical practice guidelines but will likely produce flawed data if respondents don't take the time to reflect and response honestly. If the research requires Institutional Review Board (IRB) approval further guidelines for offering incentives may also apply.

Determining an Appropriate Sample Size

To get a rough estimate of the number of individuals that should be included in a sample (i.e., a sample size large enough to accurately represent the population) you can use previously calculated estimates, organized and provided in statistical tables like the one below. There are also online sample size calculators. Looking at the recommended sample size table, you will note that when the population you are trying to describe is small, you will need to survey a large proportion of the population in order to have confidence in your results. On the other hand, once your population gets to a certain size, increasing the size of the sample will probably not improve the results.

Based on these data, a sample size of 300 to 400 will likely suffice in most cases. For populations

less than 1000 individuals, you will need to obtain responses from a large proportion of the population and, if feasible, you may wish to invite all those in the population to respond for response refusal reasons. For small populations of 100 or less, you will likely need to survey almost all those in the population if you wish to obtain valid results.

Recommended sample size for a given population
based on a required 95% confidence level

Estimated Population	Required Sample Size	Estimated Population	Required Sample Size
500000000	384	1000	278
10000000	384	750	254
500000	384	500	217
100000	384	400	196
75000	382	200	132
10000	370	100	80
5000	357	50	44
3000	341	25	24
1500	306	10	10

Adapted from Krejcie, and Morgan, (1970)

Factors that affect sample size requirements.

The basic sample size estimates may need to be adjusted given the purposes for the study and data analysis needs.

- The more homogeneous (i.e., similar or like minded) the population the smaller the sample size can be.
- The better the sampling procedures the smaller the sample can be. Noting that some sampling techniques require larger samples to reduce measurement error.
- The more planned comparison breakdowns (i.e., disaggregation of data based on multiple characteristics) the larger the sample needs to be. For example, you may wish to report comparisons between two groups of respondents, with an additional breakdown for each group based on gender and/or various age categories. A larger sample is neede to ensure sufficient numbers in each breakdown category.
- If the survey is to provide data that will be analyzed using sophisticated statistical procedures, additional respondents may be needed to satisfy the requirements of that specific procedure (e.g., a multiple regression or confirmatory factor analysis).
- The sample also needs to be larger if there is a greater likelihood of response refusal. Once you have determined a sample size that will likely produce a representative sample, the formula to calculate a revised estimate that accounts for response refusal requires you to estimate the proportion of invited participants likely to actually respond.

$$Adjusted\>Sample\>Size =

\frac{Estimated\>Sample\>Size}{Proportion\>Likely\>to\>Respond}$$

Calculating Required Sample Sizes

Using tables to estimate an appropriate sample size is based on the assumption that the sample needs only be sufficient to provide a representative sample for the population. These approximations also assume you know or can estimate the population size. Using tables to determine sample sizes can provide a reasonable estimate. Especially when the survey is designed to capture several variables; however, you may need to calculate a sample size for a specific variable based on a particular level of precision, confidence, and variability. In these cases, the sample size calculation will depend on a few considerations (criteria).

Level of precision (e). Sometimes called sampling error, the level of precision indicates how accurate you want the result to be. For example, the results may need to be within a specific confidence interval (e.g., within ± 5% or within a 95% confidence interval). In this case we would expect the true value (population parameter) to be within the specified range around the statistic obtained from the survey. In this case, if the precision needed to be ± 5% (a common value as it represents a 95% confidence interval), we would set e = 0.05 to indicate that level of precision.

Risk Level or Confidence Level (z). The confidence level is an indication of the risk you are willing to accept that the statistic (i.e., the mean or proportion being measured in the survey) is within a specific distance from the actual population parameter. The risk level (z) is based on the Central Limit Theorem which proves that the mean values of repeated samples drawn from a population are normally distribution. In a normal distribution, we know that 95% of the sample mean values, obtain from repeated samplings, will fall within 1.96 standard deviations of the population's actual mean value (i.e., the population parameter). With this level of confidence there is only a 5% probablility that the sample mean values you obtained will be extreme (out of the ordinary, far from the actual population parameter). In order to lower the risk of getting an extremely erroneous estimate of the population parameter you would need to chose a higher confidence level (z). For example, if you wished to lower your risk you could set the confidence level to 99% and set the z = 2.58 to indicate that level of risk. Using a larger confidence level will result in a larger sample size estimate which lowers the risk that the sample will produce an extreme outlier.

Common Risk Level Values

CONFIDENCE LEVEL	CRITICAL VALUE (z)
80%	1.29
90%	1.65
95%	1.96
99%	2.58

Degree of Variability (p). The variability level is an indication of the expected difference in response values, or prevalence of individuals with a specific characteristic within the population. Variability will range from 0 to 1 with a p = 0.50 representing maximum variability. The more homogenous (similar) the population the less variability there will be in responses. The more heterogeneous (dissimilar) the population the more variability exists which requires a larger sample size to obtain a generalizable result. Often we don't know the amount of variability that is likely to exist. Other times we might anticipate the variability from previous studies or antidotal observations. If the variability is unknown, a common practice is to set p = 0.5 (i.e., maximum variability). This is not necessary best practice, however, using p = 0.5 will produce a conservative (larger) sample size

estimate due to the expected dissimilarity of individuals in the population.

Determining the degree of variability is complicated when proportions are not a dichotomous condition. For example, the degree of variability for individuals with red hair can be estimated from previous approximations of those in the population. If 20% of those in the population tend to have red hair then p = 0.20. However, when the condition being studied is not a dichotomous choice, determining variability is more difficult. For example in the case where you ask people whether they agree with some statement. Often surveys use likert scales (e.g., strongly agree, agree, disagree, strongly disagree) to capture information. Variability in this case is not a dichotomous condition. You can set the variability based on a single condition (e.g., those who strongly agree) but that leaves out those who agree but not strongly. You might need to collapse categories to combine similar conditions (e.g., those who agree or strongly agree). In either case, p should represents the expected prevalence of the condition in question.

When an item on a survey is used to measure something that is not a proportion but rather a scale or continuous value, the variability estimates require we use estimates for the variability of the mean. For example, a survey may be used to ascertain the ages or height of individuals in a population; both continuous variables. In these cases p represents the expected variability of the mean not the number (i.e., proportion) of individuals exhibiting a specific condition or characteristic. This is a challenge because we often do not have a good estimate of the population variance. Furthermore, there are often multiple values being obtained from a single survey; depending on which is used, the sample size determinations can vary widely. If variance is unknown (and cannot be easily estimated), or if there are several continuous and proportion-based variables being measured in the survey, then the sample size calculation obtained is often simply a guess. For this reason, basing sample size requirement on these types of variables is a challenge and often avoided. Cochran (1977) however suggest you might estimate the variability of a continuous value using one of the follow methods.

- Pre-sample the population to obtain an estimate.
- Use values obtained when pilot testing the instrument.
- Use variance obtained from previous studies.
- Make an educated guess based on what you know about the population.

Definitions

Before we consider various ways to calculate potential sample sizes, we should review a few definitions.

Statistic – a value obtained from a sample.

Parameter – a value obtained from a population.

For example, a survey may be used to determine the proportion of individuals who indicate agreement with a specific stance. Individuals either agree or disagree. If the result is obtained from a sample it is called a statistic. If the result is obtained from the population it is called a parameter (sometime referred to as the true value assuming no measurement error).

Definitions for Variables Required to Complete Sample Size Calculations

Sample Size (n) – the estimated size of the sample required to obtain an adequate estimation of the population parameters.

Population Size (N) – the size (or estimated size) of the population.

Confidence Level (z) – indicates the risk you are willing to accept that the statistic obtained from a sample will not be very close to the actual population parameter. Using z = 1.96 would indicate a 95% level of confidence is required.

Degree of Variability (p) – indicates the response variance you expect to obtain from individuals in the population. Using p = 0.50 would indicate the maximum amount of variability is likely to occur.

Level of Precision (e) – indicates the amount of sampling error that would be acceptable. Using e = 0.05 indicates you expect the true value (parameter) for the population to be within ± 5% of the statistic obtain from the selected sample.

Basic Sample Size Formula (known population size)

The allure of this approach is that you need only two pieces of information – the population size and the desired level of precision (see Yamane, 1967).

$$n= \frac{N}{1+N(e^2)}$$

Calculating Sample Size (proportions)

This calculation is used in situations where a single item from a survey is intended to provide context. The formula uses an estimate for the proportion of individuals who exhibit a specific characteristic or attribute as the basis for the degree of variability expected (see Cochran, 1977; Daniels, 2018; Isreal, 1992). For this calculation, the expected degree of variability is considered in addition to the confidence level and level of precision required but does not require the population size to be known.

$$n= \frac{z^2 p(1-p)}{e^2}$$

An adjustment for smaller finite populations is possible if the size of the population is known.

$$n_o= \frac{n}{1+\frac{(n-1)}{N}}$$

Calculating Sample Size (scale/continuous variables)

One formula that can be used for scale or continuous values employs variance of the mean (σ^2) instead of variability based on proportions, p(1-p). Because a good estimate of the population variance is often unavailable, determining sample size using estimates of variability that are based on proportions is frequently preferred (Cochran, 1977).

$$n= \frac{z^2\sigma^2}{e^2}$$

Determining Sample Size Example

Returning to the counselling service example, now that we have conceptualized the study, suppose you now want to decide how many first-year undergraduate university students should be surveyed. It is likely that you know how many students are enrolled (say N=5000). Using a table of suggested sample sizes you would get your answer, n = 357. This number is based on a 95% confidence interval. Using a simplified forumula based on the population mean and level of risk (Yamane, 1967) we would get a similar answer, n = 371.

$$n= \frac{N}{1+N(e^2)}=\frac{5000}{1+5000(.05^2)}=371$$

Alternatively, you may wish to get a second opinion because you want to determine the proportion of students who exhibit symptoms of depression. With the expectation of a 95% confidence interval, 5% risk level, and based on previous estimates that suggest 40% of students typically suffer from depression, you would get an n = 369.

$$n= \frac{z^2 p(1-p)}{e^2}=\frac{1.96^2 (.4)(.6)}{.5^2}=369$$

Since you know how many students are enrolled, we can adjust this estimate to reflect the population size. Noting however that this may not be appropriate as the population size isn't really that small.

$$n_o= \frac{n}{1+\frac{(n-1)}{N}}= \frac{369}{1+\frac{(369-1)}{5000}}=344$$

These values are all fairly similar, however it is likely that not all those invited to participate will complete the survey. In fact, from previous experience suppose you believe that only about 25% of those invited to participate will actually complete the survey. Accounting for this it is reasonable to assume that you will need to send the survey out to over 1376 students if you hope to obtain the number of responses you need.

$$Adjusted\>Sample\>Size = \frac{Estimated\>Sample\>Size}{Proportion\>Likely\>to\>Respond} = \frac{344}{.25}=1376$$

Reflection Exercise

Reflect and be prepared to discuss the following quetsions after reviewing the sample size example presented above noting the topic being addressed (prevalence of depression amoung first year undergraduates).

How likely do you feel a systematic response refusal pattern might develop? Explain.

Suppose it is likely that the refusal to complete the survey pattern is systematic. Would increasing the sample size will help? Why or why not?

What are the benefits and potential limitations for incentivizing participation?

Chapter Summary

- Not all survey research requires sampling. With smaller populations a census (surveying the entire population) is required. With large populations, a properly selected sample will negate the need for a census.
- When sampling is the best course of action, those in the sample need to adequately represent those in the population. We call this a representative sample.
- Regardless of the samplng techniques used, some sampling error will inevitably occur.
- Sampling error occurs when the sample does not adequately represent the population.
- In addition to inadequate sampling procedures, response refusal can affect whether a sample is a representative sample.
- There are several ways to estimate the sample sized needed to obtain a represenative sample; however, several additonal factors will influence the required sample size includeing characteristics of population and data analysis needs.
- Increasing the size of the sample is almost always prefered to alliviate issues of sampling error and response refusal; however, there are times when getting a larger sample may not be feasible or cost effective.

Discussion Questions

1. Explain the benefits and disadvantages of using a sample.
2. Explan how survey fatigue affects response refusal. How do these issues affect sampling?
3. Explain why oversampling likely will not solve systematic response refusal issues.
4. Does setting a lower risk level guarentee the statistic you obtain will equal the population parameter? Explain.

References

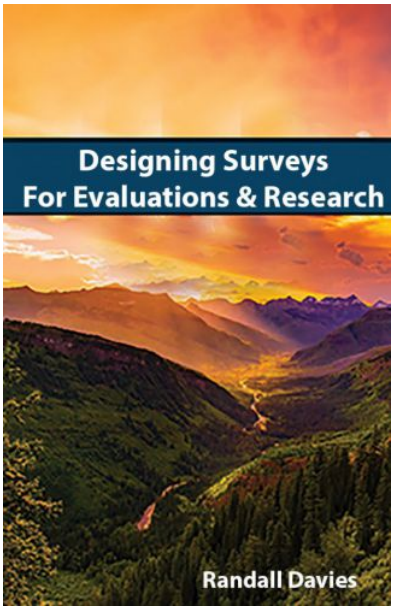
Cochran WG (1977). Sampling Techniques, 3rd edition. New York: John Wiley & Sons Link to ebook

Daniel, W. W., & Cross, C. L. (2018). *Biostatistics: a foundation for analysis in the health sciences*. Wiley.

Israel, G. D. (1992). Determining sample size. Link to Article

Krejcie, R.V. and Morgan, D.W. (1970) Determining Sample Size for Research Activities. *Educational and Psychological Measurement, 30*, 607-610.

Yamane, T. (1967). Statistics, An Introductory Analysis, 2nd Ed., New York: Harper and Row.



Davies, R. S. (2020). *Designing Surveys for Evaluations and Research*. EdTech Books. https://edtechbooks.org/designing_surveys