Sampling Methods and Sample Size in Church-Based Research

David R. Dunaetz
Azusa Pacific University, ddunaetz@apu.edu

Follow this and additional works at: https://place.asburyseminary.edu/gcrj

Part of the Christianity Commons, Practical Theology Commons, and the Religious Thought, Theology and Philosophy of Religion Commons

Recommended Citation

This Editorial is brought to you for free and open access by the Journals at ePLACE: preserving, learning, and creative exchange. It has been accepted for inclusion in Great Commission Research Journal by an authorized editor of ePLACE: preserving, learning, and creative exchange.
Sampling Methods and Sample Size in Church-Based Research

David R. Dunaetz, Editor
Azusa Pacific University

Abstract
An underlying assumption of most church-based research is that the sample from which the data is collected is representative of a wider collection of church members, church leaders, small groups, churches, denominations, or whatever is the focus of the study. To increase the likelihood that this assumption is met without making the research an impossible undertaking, various sampling strategies are used ranging from random sampling to convenience sampling, each with advantages and disadvantages. After a sampling strategy is chosen, an appropriate sample size should be pursued, depending on the nature of the study. The sample size of quantitative studies seeking to describe the frequency of a phenomenon (such as a belief or behavior) or the number in a category (such as demographics) is determined by the desired margin of error. It is more complex to determine the desired sample size of quantitative studies that test a relationship between concepts or differences between groups; several variables must be considered (statistical power, acceptable false-positive error rate, and effect size). For qualitative studies, sampling and data collection should continue until saturation is reached, that is, until little or no new information would be gained from additional data.
One of the goals of church-based research is to collect data from a limited number of participating individuals, small groups, churches, or denominations and then to make conclusions about the broader population of individuals, small groups, churches, or denominations (Dunaetz et al., 2018; Hartwig et al., 2020; Perl & Olson, 2000; Wulff, 2011). For example, data collected from 200 small groups that includes data on average member satisfaction with the group and the size of the small group may allow us to conclude what the optimal size of church small groups are in general.

The reason that academic articles are interesting and useful, whether it be the *Great Commission Research Journal* or some other academic journal, is that the results of studies presented are, to some degree, applicable to other situations, a phenomenon known as external validity. Sometimes we will read an insightful study and be convinced that its conclusions are applicable to our own context. Other times we will conclude that the study is not applicable to our context. For example, American church leaders are likely to find Moon’s (2024) article on alternative financial models of churches (pp. 43-62 of this issue) more relevant than Hildebrand’s (2024) article on the names of God in Thai (pp. 87-97 of this issue), which is likely to be much more interesting to pastors and missionaries working in Thailand.

One of the underlying reasons that some studies seem more relevant than others is the choice of data sources, or the sample, that are used to make conclusions and generalizations. If a study of churches examines churches that are similar to our own, we are more likely to see the relevance of the study than if the churches are clearly different, especially in ways that are relevant to the phenomena being studied. Researchers must therefore choose a sample that is appropriate for the conclusions that they hope to make. The goal of this article is to discuss relevant issues for selecting a sample for a study we would like to undertake (whether our unit of analysis be individuals, small groups, churches, or denominations) and how to determine how many of our unit of analysis are needed in our sample to make credible conclusions.

**Sampling Strategies**

Samples are relatively small subsets of the population that interests us...
If we are studying the people in our community, the population might be everyone in our specific town, a set of towns, or all the people within a certain radius. In any case, it is highly unlikely that we can get data from everyone to know what they are thinking or doing. Instead, we seek a sample, a relatively small number of people from whom we collect information. Using inferential statistics (Howell, 2007), we can make inferences (or conclusions) about the population that interests us if our sample is representative of the population. The following are six strategies that can be used to obtain a sample, two of which are probability strategies (everyone has an equal chance of being chosen) and two of which are non-probability strategies (some people are more likely to be chosen than others).

**Probability Strategies**

In probability strategies, every possible source of data has an equal chance of being chosen to be a source of data. If we are studying churches in a country, that means that each church would have an equal chance of being contacted by the researcher, not simply those that are the easiest to contact. Probability strategies of sampling tend to be difficult to carry out, so sometimes shortcuts are taken that facilitate data collection.

**Simple Random Sampling.** The gold standard of sampling is simple random sampling. In our example of studying churches in a country, simple random sampling would require that we obtain the name of every church in the country, put these names into some sort of “hat” and pull out names randomly from the hat until we have the desired sample size. These churches would be representative of all the churches in the country, deviating from what is typical only by chance (which can be described by inferential statistics quite well). The external validity of such a study would be high and the conclusions would be highly generalizable to the population being studied (all churches in the country). However, undertaking such a project would be very difficult. It might be possible to find all the national organizations, networks, and denominations in the country, but it is unlikely that all would readily share the contact information of their member churches. In addition, there may be many independent churches that would not be included. Moreover, churches start and die regularly (Eymann, 2012; Hayward, 1999; Moberg, 1962), making the list of all churches out of date almost immediately. To make things worse, many churches will not respond to a request for information, especially if there is no personal relationship between the researcher and the church leader providing the data (Dillman et al., 2014; Fowler, 2013). In general, simple random sampling in church-based research, and in
almost all social research, is very difficult.

**Systematic Sampling.** A slightly easier probability strategy, useful when a complete list of members of a population is known, is systematic sampling. Suppose a researcher is studying a megachurch with 4000 members, rather than randomly selecting a sample of 400 people from the membership list, the researcher might choose every tenth person on the membership list. Certainly, Abigail Aamot might not be very likely to be chosen to participate in the study, but in general, most everyone else will have close to an equal chance of being asked to participate, resulting in a nearly random sample that will be representative of the population. So, this is a good strategy when a complete list of the population is available and those selected can be counted upon to respond, but if such a list is not available, it is very difficult to carry out.

**Stratified Random Sampling.** Sometimes we are interested in comparing groups to understand how they differ. For example, a church might be interested in understanding what different groups in the church think of the current worship style (for example, high schoolers, young adults, young marrieds, parents of young children, parents of older children, empty nesters, and seniors). It is important to have a large enough sample from each group to draw conclusions about each one. In a simple random sample or systematic sampling, it is quite possible that one or more groups would be underrepresented, simply by chance. To prevent this from happening, every tenth person from the list of high schoolers, the list of young adults, the list of young marrieds, and so on could be chosen to participate in the study. This would ensure that each group would be represented. If some groups are smaller than others (for example, young marrieds), every fifth person on the list could be chosen in order to ensure a sufficient sample for each group. But again, this strategy is only feasible if such lists exist and those selected are motivated to participate.

**Cluster Sampling.** Sometimes lists of the people who interest us are partially available for some groups within the population but not for others. Suppose a church network wants to study what the members of the denomination’s churches think about a name change. The researcher could try to get a random sample drawing from all the churches in the network, but a much easier approach would be to randomly select a handful of churches and then survey all the members in this limited number of churches. This would be an example of cluster sampling. It is much easier to randomly select churches than it is individuals because the list of churches will be more readily available to the researcher, and it is easier to collect data from people grouped together in a handful of
churches than to collect a few from all the churches. If one church does not wish to participate, another can be randomly selected to replace it. A limitation to this approach is that people from small churches are likely to be overrepresented since there are far more small churches than large churches, although most people do not go to small churches (e.g., if there were 50 churches of 40 people each and 1 church of 4000 people in a town, two-thirds of the church attenders in that town would be in the large church). However, in some situations, such as classes in schools, which tend to have approximately the same size, cluster sampling would work very well.

**Non-Probability Strategies**

All the probability strategies described above tend to be difficult, if not impossible to put into practice. Even if a random sample is truly random, most church-based research is carried out through interviews or surveys. In general, it is not ethical to force randomly selected people to participate in these studies if they are randomly selected. Forcing people to reveal information that they do not want to reveal is incompatible with love and respect, central values of most churches (Matt. 7:12, Rom. 12:10, I Pet 2:17). In contrast, non-probability strategies, strategies that are non-random and in which not all members of the population are equally likely to be selected, are much easier to carry out and do not require mandatory participation.

**Convenience Sampling.** The most common sampling strategy in church-based research is convenience sampling. First, the researchers clearly determine the population being studied and then recruit whomever they can to participate in the study as long as the participant meets the selection criteria (i.e., they are part of the population being studied). For example, a study of the membership of a specific church would try to get as many of that church’s members to participate, perhaps by sending out an email invitation to the whole membership list. Or if the study were slightly broader and the population being studied included all who were associated with the church, the email invitation could be sent to everyone on the church’s email list. Since response rates in voluntary surveys tend to be low, several reminders might be needed in both cases to reach the desired sample size.

This sampling strategy does not select a representative sample of the population being studied because some people are more likely to respond to the invitation to participate (e.g., people who know the researchers, people who are especially conscientious and respond to all church-related requests, and people who are interested in the topic announced in the invitation) and some are less likely to respond (e.g., people without email,
people who struggle with reading, and people who do not enjoy revealing to others what they think). However, in many situations, this is a small price to pay to ensure reaching the desired number of participants. Nevertheless, the selection criteria should be clear (e.g., people who attend the church at least twice per month, people who are currently leading a small group in the church, or people who are salaried by the church full-time) in order to have a sample that approximately represents the population being studied.

**Quota Sampling.** In quota sampling, the researcher determines beforehand the minimum number of people (or churches, etc.) needed in each group examined in the study and then works to recruit participants who fit the criteria. For example, in recent years, the Great Commission Research Network has conducted an annual study of the challenges faced by pastors or churches in the United States (e.g., Dunaetz, 2023; Moody, 2023). In general, a convenience sample designed to represent the population that shares the organization’s values is used (i.e., all the board members of the organization reach out to their personal social networks). However, denominations and church networks are invited to participate as well so that they can compare their churches to a broader sample of evangelical churches and gain insights into the needs of the denomination or network by seeing how they are different from the broader sample. To increase the likelihood that these differences can be detected, each denomination or network is directed to keep inviting pastors and church leaders to participate until they have at least 30 responses. Once the target of 30 responses for the group is achieved, efforts taper off to collect data (but the data-collecting survey is not turned off until the analysis is undertaken, because more data is always better).

**When are Probability Strategies the Most Important?**

Random selection such as used in the probability strategies previously described is most important when comparing the size of one group to another. Political polls are the most visible type of study where random selection is needed because the number of people who support Candidate A may be very close to the number of people who select Candidate B. Convenience sampling (e.g., by Fox News or MSNBC) would produce very biased results. For this reason, political polls are very expensive and often carried out by large organizations with many resources. A panel of respondents, chosen to be representative of the nation, may be paid to participate in surveys, not just the ones that interest them.

Similarly, demographic studies need to be based on a random sample. Religious demography (Johnson & Grim, 2013; Johnson & Zurlo, 2020;
Zurlo et al., 2021) is the study of how religions and religious values are distributed among a population. This is a very important topic for mission leaders who seek to reach the least reached, and excellent resources are available (e.g., Johnson & Zurlo, 2020; joshuaproject.net; Mandryk, 2010). The most credible sources used to determine the number of Christians in a specific location are government censuses, studies carried out by professional organizations with the necessary resources (e.g., Pew Research Center), or denominations that keep track of the characteristics of their associated churches. However, when these approaches are not used, and estimates need to be made from a biased sample, the uncertainty of the figures generated can be very high (See the cases discussed in Rhodes, 2022).

Any time a study seeks to answer the question “How many?”, such as “How many people in our town attend church?” “How many people in a country self-identify as evangelicals?”, it is important to use a probability strategy (e.g., cluster sampling) to maximize unbiased sampling. In contrast, studies that seek to understand relationships, especially relationships that are grounded in human nature or in culture, are less sensitive to non-random samples (e.g., convenience sampling). For example, if we want to study the relationships between churches changing their name (e.g., from First Baptist Church to Cornerstone Church or New Hope Church) and changes in people’s commitment to their church, it is not necessary to get a representative sample from all churches that have changed names. A convenience sample of all church attenders, even if they just come from a few churches, might reveal trends that are likely to be true for the population of all churches. Certainly, there is a possibility that people from a church with a botched name change might be overrepresented in a convenience sample, but the researcher may deem this to be only a minor threat and may be willing to conduct the study in spite of it.

Typically, quantitative studies in which a hypothesis is tested (Dunaetz, 2021) only require a convenience sample from the appropriate population. The more representative the sample is of the population studied, the more credible the study will be, but even convenience samples from a very specific population may be useful, as long as the sample is large enough.

**Sample Sizes**

In quantitative research, determining the needed sample size depends on the goal of the research. If we are simply trying to collect descriptive statistics such as the number of something or the average of some variable, our sample size depends on the desired margin of error. Our desired
margin of error can be expressed as a confidence interval, that is, a range of numbers that is likely (typically 95% likely) to contain the true number, that is, the value of the variable in question if we were to measure the whole population, and not just a sample taken from the population. On the other hand, if we want to test a hypothesis, we need to choose a sample large enough to make detecting the relationship in the hypothesis very likely.

**Sample Size in Quantitative Studies: Descriptive Statistics**

Descriptive statistics are numbers that we get from measuring a sample (e.g., some of the members in our church, but not our whole church) in order to estimate what the true value is (that is, if we had data for everyone in the church). For example, suppose a pastor is trying to determine if the church would be more interested in a sermon series on apologetics (Choice A) or a sermon series on building relationships with people in order to better communicate the gospel (Choice B). To be reasonably sure what the majority desires, we need to know the confidence interval (usually the 95% confidence interval) of the information we collect from our sample. Suppose the pastor does a survey of 50 church members, and 30 say they would prefer the series on apologetics. This means that 60% of the sample (= 30/50) prefer the series on apologetics. However, the margin of error (based on the 95% confidence interval) of this estimate is ±14%, that is, there is a 95% chance that the true percentage of those in the church who prefer the series on apologetics lies in the range 60% ± 14%. This can also be written 95% Confidence Interval [46% - 74%]. So, it is very possible that a minority of the church (e.g., 47%) would prefer the series on apologetics. This means that it is too close to call.

To be more certain, we need to have a large enough sample size so that 50% is not in the confidence interval, for example, 60% ± 8% (or 95% Confidence Interval [52% - 68%]). If this were the case, we could be reasonably certain that the series on apologetics is the preferred series. In a study like this, when there are only two choices, the formula for the confidence interval is the following (De Veaux et al., 2004):

\[
95\% \text{ Confidence Interval} = p \pm 1.96 \sqrt{\frac{p(1-p)}{n}}
\]

Where \( p \) = the fraction of people who chose choice A and \( n \) is the sample size (Bulmer, 1979; Howell, 2007). In our example, 30 out of 50 people chose Choice A, so \( n = 50 \) and \( p = 30/50 = 60\% \). Similarly, 40% chose Choice B (20/50 = 40%, which is 1 - \( p \) in the above equation). Therefore, the 95%
confidence interval is $60\% \pm 1.96\sqrt{.6(1-.6)/50}$, or $60\% \pm 14\%$ as stated above. This means that we cannot be sure of the entire congregation’s preference since it is somewhere in the range of 46% to 74%.

If we want a narrower confidence interval to be more certain of the congregation’s preference, we need to collect more data so that we will have a larger $n$ (and which will most likely give us a slightly different value of $p$). If we collected data from 100 people ($n = 100$), and 61% preferred Choice A, we would have a confidence interval of $61\% \pm 9.6\%$, that is, 95% Confidence Interval [51.4%, 70.6%]. In this case, we can be reasonably sure that the true proportion of the congregation that prefers Choice A is greater than 50%. If we had surveyed 1000 people, our confidence interval would have been about ±2%, a figure that is often reported in national surveys concerning predicted election results that are based on a sample of about 1000 people.

It is important to emphasize that such estimates are only valid for random samples. If the calculations are based on data from a biased sample (e.g., all the men at a Saturday morning men’s group), the estimate is also likely to be biased. Such data would be useful, but its meaning is not as relevant as we would like it to be.

**Sample Size in Quantitative Research: Hypothesis Testing**

Quantitative research that aims to test a hypothesis (Dunaetz, 2021; Fisher, 1925) uses a different approach to determine the sample size. An important concept is statistical power (Cohen, 1988; Faul et al., 2007). Statistical power is the probability that we will be able to detect that a hypothesis is supported based on the data that we collect. It can range from 0% to 100%. Power of 0% indicates that there is no chance of detecting a hypothesized effect if it were real. Suppose for example a pastor wants to know if people in the church thought a certain sermon was better than average. His hypothesis is that the average congregant thinks that sermon was above average. If he asks only two people, he has a 0% chance of having strong evidence (e.g., having less than a 5% chance of being wrong, $p < .05$) that his hypothesis is correct. The reason for this is that if the overall audience did not think it was above average (e.g., 50% thought it was above and 50% thought it was below; this is known as the null hypothesis), there would be up to a 25% chance (50%*50%) that both people in a random sample would think it was above average. This means that he would have much more than a 5% chance of being wrong ($p > .05$).

On the contrary, if we asked 1000 people what they thought, we have
a much greater chance of detecting with reasonable certainty \( p < .05 \) that the sermon was either above or below average. The power to detect this difference approaches 100\% as the sample size approaches the population size (the number of people in the church). This means that the statistical power to support a hypothesis increases with sample size.

The needed sample size also depends on other factors. First, it depends on the criteria we use to determine what it means to be reasonably sure. In the above example, we said that we wanted to have less than 5\% chance of being wrong if we conclude that our hypothesis is supported \( p < .05 \). This 5\% is known as the alpha level, the significance level, or the rate of false positives. However, we could have said we are willing to be wrong 10\% of the time. Then we would choose an alpha level of 10\%; if there were less than a 10\% chance of getting our results randomly given there was no difference between this sermon and the average sermon, we would say our results are significant. It also means we would make erroneous conclusions (false positives) 10\% of the time, versus 5\% of the time for a 5\% significance level. Thus, we can reduce the needed sample size if we increase our alpha level or significance level. If we can tolerate a greater likelihood of a false positive, we can reduce the sample size. If we want to make a false positive less likely, we should increase our sample size.

Secondly, the needed sample size depends on whether our hypothesis is directional or non-directional. A directional hypothesis (also known as a one-tailed hypothesis) predicts the direction of a difference (or a relationship). If our hypothesis states that the sermon in question is better than average, it is a directional (one-tailed) hypothesis. However, if we do not state the direction of a difference (or relationship), we are making a non-directional hypothesis. If we make the hypothesis that the average evaluation of the sermon in question is different from the overall average, that is, either below or above the overall average, we are making a non-directional hypothesis. If we make the hypothesis that the average evaluation of the sermon in question is different from the overall average, that is, either below or above the overall average, we are making a non-directional (two-tailed) hypothesis. To reach the needed statistical power, a smaller sample is needed in a directional (one-tailed) hypothesis than in a non-directional (two-tailed) hypothesis. This is because we are only looking in one direction to find a relationship rather than two.

However, if we make a directional hypothesis (e.g., the sermon is above average) and the results indicate that reality is in the other direction (the sermon is below average), we cannot ethically change our hypothesis to a non-directional hypothesis that would have been supported. This is a phenomenon known as hypothesizing-after-the-fact (HARKing; Kerr, 1998) which is unethical because it can be misused to make almost any data set look like it supports a hypothesis.

A third factor to consider in determining sample size is the effect size.
If there is a big difference between groups or a strong correlation between two variables (i.e., a large effect size), the difference or correlation will be easier to detect than if the difference or correlation is small. The smaller the expected effect size, the larger the sample needed. But often we do not know what effect size to expect. In this case, we should aim to detect the smallest effect size that has a practical impact on the phenomenon that we are studying. For example, if a congregation’s rating of a pastor’s sermons is negatively correlated with the length of the sermon with \( r = -.02 \), such a small correlation may be significant, but it has no practical implications; it is too small to be noticed in everyday life. In church-based research, a rule of thumb is that we are especially interested in correlations of at least \(|r| \geq .15\). Correlations with a magnitude smaller than .15 may not have much of a visible or practical effect.

Another consideration in determining sample size is how much statistical power we want. If we want 95% statistical power or greater, we will need a very, very large sample. Traditionally, researchers are content with 80% power. An online sample size calculator (such as [https://sample-size.net/correlation-sample-size/](https://sample-size.net/correlation-sample-size/)) can be used to calculate the needed sample size in light of the various factors we have discussed. As a rule of thumb for church-based research, a sample of 350 will usually be sufficient. This will give us 80% power to detect a correlation of \( r = .15 \).

**Sample Size in Qualitative Research**

Qualitative research, rather than measuring specific concepts as in quantitative research (Dunaetz, 2022), seeks to gain insight into how people understand various phenomena (Creswell & Creswell, 2017; Patton, 2002). Qualitative research may focus on producing a phenomenology (a study of a specific phenomenon; Moustakas, 1994), a narrative biography (a study of an important person; Bornat, 2008), grounded theory (the development of a new theory based on the examination of people's experiences; Charmaz, 2014), a case study (Tellis, 1997), or an ethnography (a study of a specific culture, either anthropological or organizational; Brewer, 2000). These studies tend to produce more subjective interpretations of the data than quantitative research. Rather than collecting quantitative data on well-defined concepts, qualitative research collects verbal and subjective evidence relative to the topic being studied. The most common sources of data are interviews and written texts. In studies relevant to everyday disciple making, interviews may be the most common source of data.

The principle of *saturation* (Saunders et al., 2018) is often used to
determine how many interviews need to be conducted or how many source documents need to be examined in a qualitative study. Saturation occurs when additional interviews or additional readings of source documents no longer contribute something new vis-à-vis the research question. For example, in a study of how Generation Z Christians in a specific church live out their faith, the researcher might notice that by the thirtieth interview, little or no new information is being gained; the lifestyles being described have all been described by those previously interviewed. Once saturation is reached, data collection may stop, and data interpretation should begin.

This means that it is impossible to know how many people should be studied for a qualitative study. As a rule of thumb, the interviewer should set aside enough time for 30 interviews. If saturation is reached before the thirtieth interview, then data collection can end early. However, the researcher should be prepared to continue collecting data beyond the thirtieth person if saturation has not been reached.

**Summary**

For church-based research to be credible, it needs to be based on data collected from people who are representative of the population that is being studied. A sample selected randomly is theoretically the best, but ethical and practical limitations often require the researcher to use a convenience sample. Such samples might be quite appropriate when examining the relationships between concepts that interest us, but they are less appropriate when we are trying to get an accurate count for classifying the data into categories (e.g., demographic studies).

The needed sample size depends on the type of study. In demographic studies, larger sample sizes will permit us to have narrower confidence intervals. In hypothesis testing of relationships between variables, larger sample sizes will make detecting the hypothesized phenomena more likely; as a rule of thumb, 350 participants for a quantitative study (such as a survey) will usually be quite sufficient to test a hypothesis, although smaller sample sizes can still produce useful results. In qualitative studies, data should be collected until saturation is reached, that is until no more useful information relevant to the research question is gained from further data collection.

By properly designing church-based research, which includes both determining the way the sampling will be conducted and how large the sample should be, the researcher is more likely to produce useful research that will enable others to better serve the Lord in the communities where he has placed them.
References


