Chapter 4. Sampling

Linh Pham, Trung Tran, Quy Khuc

©Vietkap team, Hanoi, Vietnam

February 3, 2021

Preprint DOI: <https://osf.io/w9qks/>

Abstract

In previous chapters, we have understood the importance of primary data in scientific research and the procedure for conducting a primary data investigation. After completing a questionnaire, the next step is to conduct a field research. In this chapter, we will clarify two questions: who are the respondents and are the number large enough? Because of limited resources and time, most studies were unable to collect data on the whole population of study subjects but on selected participants. This work is called “Sampling”. Based on the analytical results calculated from the investigated sample, researchers can extrapolate scientific information and conclusions about the whole population. The sampling process consists of 2 main tasks: selecting the target population and deciding on the number of samples to investigate (sample size). Currently, there are many scientific documents on sampling, but this content is complex because it is related to knowledge of probability, statistics, and contains many concepts, formulas. Consequently, readers, especially young researchers, often find it difficult to understand and to apply. In this chapter, we do not cover all sampling methods as well as the complex formulas, yet the main goal is to present the right, basic knowledge in an apparent and easy to apply way. For more advanced knowledge, readers can look at references.

Keywords: Study population, sample selection, sample size, sampling

Introduction

Unlike studies without the purpose of generalizing from a sample, those that try to extrapolate and generalize about characteristics of the entire population from those of a sample will depend greatly on sampling (Figure 1). Therefore, in the primary data investigation, sampling is considered the overarching step [6, 4, 1, 3]. We will start this chapter with the most basic concepts of research population, sampling frame and sample size. Next, we will delve into sampling methods and calculate sample size using examples of research in social sciences.

Figure1. Sampling. source [2]

Population, sampling frame and sample

A **study population** is a collection of all research subjects or individuals possessing characteristics that meet the criteria of the research, from which a sample is drawn. A sample is a subset of the research population including a number of research subjects, selected according to a certain rule and usually representative of the research population [1, 2].

A **sampling frame** is simply understood as a collection of individuals in the target population [2]. After identifying the research population, a sampling frame needs to be developed for data collection. In this step, scientists need to avoid three types of “false” frames. First is a frame that contains too many individuals, some of which are even not in the target population. Next is one that contains too few individuals that is either in or out of the target population. The third “false” frame contains an incorrect set of individuals or is itself not part of the target population. In order to frame a sample, scientists need to determine individuals in the target population and the sample size. For example, in the study of students’ academic performance, the frame could be the list of students in separate grades, or according to performance, another the list of teachers. Depending on the research objective or questions, scientists build a suitable framework from which an investigation could be conducted.

Sample size is the number of individuals of a sample selected from the population to be surveyed [1, 3]. Thus, a sample could be understood as a subset of the population and sample size is the size of this subset. For example, a research on income of residents in a province has the whole population of one million. The researcher decides to conduct a survey on 100 thousand people belonging to various income groups, then 100 thousand people is the sample size. The scientific basis for selecting the sample size is the Law of large numbers in probability and statistics. This rule states that as a sample size grows, its mean gets closer to the average of the whole population.

Sampling methods

There are two sampling methods used by scientists, including probability sampling (random) and non-probability (non-random) sampling.

Probability sampling involves completely random selection, irrespective of subjectivity. As a result, the probability of each observation being selected is similar, thereby likely to select a sample representative of the whole population. Since the sampling error can be calculated, we can apply statistical estimation methods, test statistical hypotheses, and process data to extrapolate results from the sample. However, this method has some limitations. First, it is difficult to apply in that a specific list of the general population could not be identified. Besides, it demands a plethora of time, cost and human resources for data collection if the subjects are scattered over a geographic area.

Non-probability sampling involves selecting a unit of observation depending on the subjectivity of the chosen. This method is often used when the random sampling method is unviable, such as in novel surveys without any information about the subjects, or subjects dispersing, volatile or belonging to disparate groups. Non-random sampling is not completely based on mathematics but requires a close combination between theoretical analysis and social practice. Hence, it, to a large extent, represents the subjectivity of the selector.

Determining sample size

In sampling, determining sample size always causes confusion for researchers, even for experienced ones. Common problems are in deciding on the size and in explaining why that number: 300 and not 350 or 500, for example. In this section, we will discuss two approaches to determining sample size. The first is to use mathematical formulas, and the second, for some social studies, we can employ a combination of "expert, experience and mathematical formulas".

Using formulas: In practice, the entire population can be divided into several groups with different characteristics, from which there are different ways to calculate the sample size. Table 1 shows two simplest and most common formulas for sample size determination. Besides them, there are many other techniques using statistical analysis models to calculate the sample size like the Exploratory Factor Analysis (EFA) or regression model. However, in this chapter, we will not go into details to avoid complication.

Table 1. Sample size calculation formulas

	Whole population size unknown	Whole population size known
Formulas	$n = \frac{Z^2 p(1-p)}{e^2}$	$n = \frac{N}{1 + Ne^2}$
	n: sample size Z: confidence level (usually at 95%), standard value of 1.96 p: probability e: error	n: sample size N: whole population size e: error

From the formulas in Table 1, it can be seen that sample size calculation depends on many factors. However, given that representative factors are the salient point, there are four main elements that influence the determination

of the sample size.

- First is reliability of the data. That is, the degree of certainty and generalization for the whole. The more generalizable the data is, the more reliable it is.
- Second is small and acceptable errors. Although errors are inevitable, we must limit and minimize them to ensure accuracy for estimate of the sample.
- Third is the types of test statistics, some of which require a minimum sample size to have meanings.
- Fourth is the size of the whole population. The ratio of sample to population must meet a certain requirement. For example, the sample size to be surveyed for a city of 5 million inhabitants would be larger than that for a commune of 5,000 inhabitants.

Determining sample size in social science research: Although sample size determination is not too complex, scientists with many years of research might face several difficulties in identifying reasonable ones for their studies. The research on factors affecting household income could be cited as an example. In social science research, scientists often use regression models to process data. Thus, the sample size is also calculated in order to serve this purpose. Drawing from experience in primary data investigation, we noted that there are three ways to determine the sample size.

- The first is that researchers can rely on the “expert” method - consult experienced scientists for advice. This method is fast and effective if the right specialist is found.
- Second, researchers can apply the rule of thumb in determining sample size. In other words, a sample size ranging from 300 to 500 is acceptable [3].
- Third, scientists can use different levels of confidence and error to construct a table of minimum sample sizes according to distinctive error and confidence scenarios. From that, a sample size should have a minimum of 271 items to ensure a 90% confidence level or more, with 5% error. When the sample number is 385 or higher, the confidence level will be 95% or more [5]. For higher reliability, scientists should determine a sample size of 400 or more. The reason is that after the survey, they have to discard a certain number of faulty observations or outliers, usually accounting for 3-5% of the total sample size.

In summary, sampling is a critical step in the primary survey “ecosystem”. A bigger sample size means that the characteristics of the population is clearly presented. However, it also means higher cost and more time. In contrast, if the sample size is too small, it will lack objectivity and show no difference between the subjects. With the goal of honing this step, scientists need to firmly grasp the research objectives, the Law of large numbers and the theory of statistical probability. A suitable sample size is neither too big nor too small, meeting the purpose and status of the study. Flexibility in determining the sample size according to financial and human resources should also be heeded. However, under any circumstances, the sample size should meet at least one criterion according to the formula for sample size calculation, or be based on the sample size of similar conducted studies.

Review questions

1. Define sample and sample size. Give examples?
2. What is the scientific basis for sampling? Which factors affect sampling? Why?
3. Present the formula for sample size determination and give examples?
4. Present sample selection methods. Give examples and describe the advantages and disadvantages of each method?

References

- [1] Kenneth Bailey. *Methods of social research*. Simon and Schuster, 2008.
- [2] Nicholas E Flores. Conceptual framework for nonmarket valuation. In *A primer on nonmarket valuation*, pages 27–54. Springer, 2017.
- [3] Uma Sekaran and Roger Bougie. *Research methods for business: A skill building approach*. John Wiley & Sons, 2016.
- [4] Quy Van Khuc, Phu Pham, and Duc-Trung Tran. Questionnaire design.
- [5] Quy Van Khuc, Tri Vu Phu, and Phuong Luu. Dataset on the hanoian suburbanites’ perception and mitigation strategies towards air pollution. *Data in Brief*, 33:106414, 2020.
- [6] Quy Van Khuc and Duc-Trung Tran. Primary data.