Principles of Survey Research
Part 5: Populations and Samples

Barbara Kitchenham
Dept. Computer Science
Keele University, Staffs, UK
barbara@cs.keele.ac.uk

Shari Lawrence Pfleeger
RAND Corporation
Arlington VA 22202-5050
shari_pfleeger@rand.org

Abstract

This article is the fifth installment of our series of articles on survey research. In it, we discuss what we mean by a population and a sample and the implications of each for survey research. We provide examples of correct and incorrect sampling techniques used in software engineering surveys.

Keywords: survey methods, populations, sampling

Introduction

In this article, we describe how to obtain a valid survey sample from a target population. We discuss why a proper approach to sampling is necessary and how to obtain a valid sample. We also identify some of the sampling problems that affect software engineering surveys.

The overriding key to understanding sampling is to acknowledge that a valid sample is not simply the set of responses we get when we administer a questionnaire. A set of responses is only a valid sample, in statistical terms, if has been obtained by a random sampling process.

As with previous articles in this series, we will use three existing software engineering surveys to illustrate common sampling errors:

1. Two related surveys undertaken by Timothy Lethbridge [1] and [2], both aiming to compare what software engineers learned at university with what they needed to know in their current jobs.
2. A survey we ourselves undertook, to investigate what evidence organizations use to assist technology adoption.
3. A Finnish survey [4] aimed at investigating project risk and risk management strategies.

In the first case, Lethbridge solicited participation via the Web and industrial contacts. In the second case, we included a questionnaire with a mailing of *Applied Software Development*. In the third case, Ropponen and Lyytinen mailed a questionnaire to each of a pre-selected sample of members of the Finnish Information Processing Association whose job title was "manager" or equivalent. Furthermore, they sent the questionnaire to at most two managers in the same company. We will show that only the Finnish study can claim that the set of responses to their questionnaire represents a sample of a defined population.

Samples and Populations

To obtain a sample, you must begin by defining a *target population*. The target population is the group or the individuals to whom the survey applies. In other words, you seek those groups or individuals who are in a position to answer the questions and to whom the results of the survey apply. Ideally, a target population should be represented as a finite list of all its members. For example, when pollsters survey members of the public about their voting preferences, they use the electoral list as their target population list.

A valid sample is a *representative subset* of the target population. The critical word in our definition of a sample is the word "representative." If we do not have a representative sample, we cannot claim that our results generalize to the target population. If our results do not generalize, they have little more value than a personal anecdote. Thus, a major concern when we sample a population is to ensure that our sample is representative.

Before we discuss how to obtain a valid sample, let us consider our three survey examples. In Lethbridge's case, he had no defined target population. He might have meant his target population to be every working software developer in the world, but this is simply another way of saying the population was undefined. Furthermore, he had no concept of sampling even his notional population. He merely obtained a set of responses from the group of people motivated to respond. Thus, Lethbridge's target population was vague and his sampling method non-existent. So although he described the demographic properties of his respondents (age, highest education qualification, nationality etc.), no generalization of his results is possible.

With respect to our survey, we have already noted in our article on formulating questions that we were probably targeting the wrong population because we were asking individuals to answer questions on behalf of their companies. However, even if our target population was all readers of *Applied Software Development*, we did not have any sampling method, so our responses could not be said to constitute a valid sample. At this point, you may begin to disagree with us. If we sent questionnaires to our entire population, why aren't the responses a valid sample? Furthermore, why should we bother sampling when we can attempt to contact the entire population? These are important questions, and we will address them later in this article.

In the meantime, consider the Finnish survey. Ropponen and Lyytinen had a list of all members of the Finnish Information Processing Association whose title was manager. Thus, they had a defined target population. Then, they sent their questionnaires to a pre-selected subset of the target population. If their subset was obtained by a valid sampling method (surprisingly, no sampling method is reported in their article), then their subset constituted a valid sample. As we will see later, this situation is not sufficient to claim that the actual responses were a valid sample, but it is a good starting point.

Obtaining a Valid Sample

We begin by understanding the target population. We cannot sample a population if we cannot specify what that population is. Our initial assessment of the target population should arise from the survey objectives, not from a sense of who is available to answer our questions. The more precisely the objectives are stated, the easier it will be to define the target population.

The specific target population may itself be a subset of a larger population. It may be specified by the use of *inclusion* or *exclusion* criteria. For example, if we are interested in the extent to which software engineering education meets the needs of industry, we should exclude from our target population software engineers who did not major in software engineering, computer science or a related discipline.

It is often instructive to consider the target population and sampling procedure from the viewpoint of data analysis. We can do this during questionnaire design but we should also re-assess the situation after any pretests or pilot tests of the survey instrument. At this point we will have some actual responses, so we can try out our analysis procedures. We need to consider whether the analyses will lead to any meaningful conclusions, in particular:

- Will the analysis results address the study objectives?
- Can the target population answer our research questions?

Considering the first question, Lethbridge's objectives were to provide information to educational institutions and companies as they plan curricula and training programs. This goal raises obvious questions: which educational institutions and which companies? Lethbridge's target population was poorly defined but can be characterized as any practising software engineer. Thus, we must ask ourselves whether replies from software engineers who would have attended different education institutions, worked in different companies or had different roles and responsibilities would indicate clearly how curricula and training courses could be improved. At the very least, general conclusions may be difficult. The results would need to be interpreted by people responsible for curricula or training courses in the light of their specific situation.

The next question concerns the target population. Will the target population provide useful answers? Lethbridge did not apply any inclusion or exclusion criteria to his respondents. Thus, the respondents may include people who graduated a very long time ago or graduated in non-computer science-related disciplines and migrated to software engineering. It seems unlikely that such respondents could offer useful information about current computer science- related curricula or training programs.

Consider now our own survey of technology adoption practices. We have already pointed that our target population was the set of organizations (or organizational decision-makers) making decisions about technology adoption. However, our sample population solicits information from individuals. Thus, our *sampling unit* (i.e. an individual) did not match our *experimental unit* (i.e. an organization). This mismatch between the population sampled and the true target population is a common problem in many surveys, not just in software engineering. If the problem is not spotted, it can result in spurious positive results, since the number of responses may be unfairly inflated by having many responses from organizations instead of one per organization. Furthermore if there are disproportionate number of responses from one company or one type of company, results will also be biased.

The general target population of the Finnish survey of project risk was Finnish IT project managers. The actual target population was specified as members of Finnish Information Processing Association whose job title was "manager" or equivalent. People were asked about their personal experiences as project managers. In general, it would seem that the sample adequately represents the target population, and the target population should be in a position to answer the survey's questions.

The only weakness is that the Finnish survey did not have any experience-related exclusion criteria. For instance, respondents were asked questions about how frequently they faced different types of project problems. It may be that respondents with very limited management experience cannot give very reliable answers to such questions. Ropponen and Lyytinen did consider experience (in terms of the number of projects managed) in their analysis of the how well different risks were managed. However, they did not consider the effect of lack of experience on the initial analysis of risk factors.

Sampling Methods

Once we are confident that our target population is appropriate, we must use a rigorous sampling method. If we want to make strong inferences to the target population, we need a probabilistic sampling method.

We describe below a variety of sampling methods, both probabilistic and non-probabilistic.

Probabilistic Sampling Methods

A probabilistic sample is one in which every member of a target population has a *known, non-zero probability* of being included in the sample. The aim of a probabilistic sample is to eliminate subjectivity and obtain a sample that is both unbiased and representative of the target population. It is important to remember that we cannot make any statistical inferences from our data unless we have a probabilistic sample.

Simple random sample

A simple random sample is one in which every member of the target population had the *same* probability of being included in the sample. There are a variety of ways of selecting a random sample from a population list. One way is to use a random number generator to assign a random number to each member of the target population, order the members on the list according to the random number and choose the first n members on the list, where n is the required sample size.

Stratified random sample

In this case, the target population is divided into subgroups called *strata*. Each stratum is sampled separately. Strata are used when we expect different sections of the target population to respond differently to our questions, or when we expect different sections of the target population to be of different sizes. For example, we may stratify a target population on the basis of sex, because men

and women often respond differently to questionnaires.

The number of members selected from each stratum is usually proportional to the size of the stratum. In a software engineering survey, we often have far fewer women than men in our target population, so we may want to sample within strata to ensure we have an appropriate number of responses from women.

Stratified random samples are useful for non-homogeneous populations, but they are more complicated to analyze than simple random samples.

Systematic Sampling
Systematic sampling involves selecting every *n*th member from a population list. If the list is random, then selecting every *n*th member is another method of obtaining a simple random sample. However, if the list is not random, this procedure can introduce bias. Non-random order would include alphabetical order or date of birth order.

Cluster-based sampling
Cluster–based sampling is the term given to surveying individuals that belong to defined groups. For example, we may want to survey all members of a family group, or all patients at specific hospitals. Randomization procedures are based on the cluster, not the individual. We would expect members of each cluster to give more similar answers than we would expect from members of different clusters. That is, answers are expected to be correlated within a cluster. There are well-defined methods for analyzing cluster data, but the analysis is more complex than that of a simple random sample. (For an example, see [3].)

Non-Probabilistic Sampling Methods

Non-probability samples are created when respondents are chosen because the are easily accessible or the researchers have some justification for believing that they are representative of the population. This type of sample runs the risk of being biased (that is, not being representative of the target population), so it is dangerous to draw any strong inferences from them. Certainly it is not possible to draw any statistical inferences from such samples.

Nevertheless, there are three reasons for using non-probability samples:

1. The target population is hard to identify. For example, if we want to survey software hackers, they may be difficult to find.
2. The target population is very specific and of limited availability. For example if we want to survey senior executives in companies employing more than 5000 software engineers, it may not be possible to rely on a random sample. We may be forced to survey only those executives who are willing to participate.
3. The sample is a pilot study, not the final survey, and a non-random group is readily available. For example, participants in a training program might be surveyed to investigate whether a formal trial of the training program is worthwhile.

Convenience Sampling
Convenience sampling involves obtaining responses from those people who are available and willing to take part. The main problem with this approach is that the people who are willing to participate may differ in important ways from those who are not willing. We see this kind of sampling particularly on web sites, where people who have complaints are more likely to provide feedback than those who are satisfied with a product or service.

Snowball sampling
This involves asking people who have participated in a survey to nominate other people they believe would be willing to take part. Sampling continues until the required number of responses is obtained. This technique is often used when the population is difficult for the researchers to identify. For example, we might expect software hackers to be known to one another, so if we found one to take part in our survey, we could ask him/her to identify other possible participants.

Quota sampling
Quota sampling is the non-probabilistic version of stratified random sampling. The target population is spit into appropriate strata based on know subgroups (e.g. sex, educational achievement, company size etc.). Each stratum is sampled (using convenience or snowball techniques) so that number of respondents in each subgroup is proportional to the proportion in the population.

Focus groups
Focus groups are usually formed by the researchers from their personal contacts. They usually consist of 10 to 20 people who are intended to represent some population. Focus groups are commonly used in pre-survey pilot studies.

Sample size

A major issue of concern when sampling is determining the appropriate sample size. There are two reasons why sample size is important. First, an inadequate sample size may lead to results that are not significant statistically. In other words, if the sample size is not big enough, we cannot come to a reasonable conclusion, and we cannot generalize to the target population. An extreme example of this problem is the receipt of a single response; we cannot draw any conclusions from a single respondent. Second, inadequate sampling of clusters or strata disables our ability to compare and contrast different subsets of the population.

To determine an adequate or minimum sample size, we need to know four things about our study:

1. The *alpha level* we intend to use, where alpha is the probability of a Type I error (that is, the probability of falsely rejecting the null hypothesis). Alpha is usually set at 0.05 or 0.01.
2. The *beta level* we intend to use, where beta is the probability of a Type II error (that is, the probability of falsely accepting the null hypothesis). Beta is usually set at 0.20. We often talk about the *power* of a test or experiment; power is calculated as 1-beta. The power of a test is the probability of correctly accepting the alternative hypothesis.
3. The *effect size,* which is the difference in outcomes between two groups. For example, suppose we want to investigate whether there are pay differences between male and female software engineers. We might survey men and women who

graduated in 1998 and ask them what their base salary is. The effect size is the difference between the average male salary and the average female salary.

4. The *variance* of the effect, which is the degree to which the data vary within a group. In our salary example, we can look at the variance of salary values for men and women.

Of course, the effect size and variance are what we expect to obtain as a result of doing our survey, so we need prior information in order to determine an appropriate sample size. We can obtain such information from previous surveys, pilot surveys, or expert opinion.

In the simple case of assessing the sample size, assuming a Normal distribution for the response variables, two groups with equal numbers in each group and equal within-group variances, the sample size (per group) is:

$$\left[\frac{(z_\alpha - z_b)\times\sigma}{\mu_1 - \mu_2}\right]^2$$

where

μ_1-μ_2 is the effect size.

σ is the common standard deviation.

z_a is the upper tail in the standard normal distribution corresponding to α. For example, $z_a = 1.96$ if $\alpha = 0.05$.

z_b is the lower tail in the standard normal distribution corresponding β. $z_b = -0.84$ if $\beta = 0.20$.

We noted in an earlier article that we should try to anticipate non-response when we set our sample size. For instance, if the formula tells us that the theoretical optimum sample size is 50 but we expect only an 80% response, we would increase our sample size to 63.

Why sample?

Let us return to the issue of why we should sample at all, rather than try to get responses from the entire population. Indeed, if the population is small (usually defined as less than 50), we probably should attempt to obtain responses from all in the population. However, we should still apply the same follow-up procedures that we would have used had we employed a sample.

If we have a large population, we need to sample the population for the following reasons.

1. Lower and more appropriate administrative costs. That is, we usually need to make sure that our survey will not cost more than it needs to cost. We use sampling to obtain sufficient responses to answer our questions but no more.
2. The ability to administer controlled follow-up procedures. We can follow-up non-respondents, both to encourage them to complete the survey and also to try to understand the reason for non-response. If we have a population of several thousands, send questionnaires to all of them and achieve a response rate of 20%, it is hard to systematically follow-up all non-responses. However, if we do not confirm that there is no bias due to non-response, we cannot confirm we have a representative sample

The second point has ramifications for survey administration. In order to follow-up non-response, we need to know who has replied and who has not. This requirement means that our questionnaires must be individually coded, so we can match replies to questionnaires. At the same time, we need to put in place procedures to protect the anonymity of respondents.

The Finnish survey provides a good example of follow-up procedures. The researchers identified a sample of 25 non-respondents and phoned them to ask why they had not participated. They found that 25% of the non-respondents were not in fact managers, 13% of addresses were out of date and 55% had no time or never responded to surveys. Thus, the researchers were able to claim that there was no evidence of systematic bias among non-respondents, so their sample can be considered representative of the target population.

We hope this article has convinced you of the need for a better approach to sampling in software engineering surveys. In our experience, invalid samples are the most common problem in such surveys.

In the next article and final article in this series, we discuss how to analyze survey data.

References

[1] Timothy Lethbridge, A survey of the relevance of computer science and software engineering education, *Proceedings of the 11th International Conference on Software Engineering*, IEEE Computer Society Press, 1998.

[2] Timothy Lethbridge, What knowledge is important to a software professional, *IEEE Computer*, May 2000.

[3] Levy, P.S. and Lemeshow, S. Sampling of Populations: Methods and Applications. *Wiley Series in Probability and Statistics, John Wiley and Sons Inc., Third Edition,* 1999.

[4] J. Ropponen and K. Lyytinen, Components of software development risk: How to address them. A project manager survey, *IEEE Transactions on Software Engineering* 26(2), February 2000.