CHAPTER 1

Introduction

Sampling consists of selecting some part of a population to observe so that one may estimate something about the whole population. Thus, to estimate the amount of lichen available as food for caribou in Alaska, a biologist collects lichen from selected small plots within the study area. Based on the dry weight of these specimens, the available biomass for the whole region is estimated. Similarly, to estimate the amount of recoverable oil in a region, a few (highly expensive) sample holes are drilled. The situation is similar in a national opinion survey, in which only a sample of the people in the population is contacted, and the opinions in the sample are used to estimate the proportions with the various opinions in the whole population. To estimate the prevalence of a rare disease, the sample might consist of a number of medical institutions, each of which has records of patients treated. To estimate the abundance of a rare and endangered bird species, the abundance of birds in the population is estimated based on the pattern of detections from a sample of sites in the study region. In a study of risk behaviors associated with the transmission of the human immunodeficiency virus (HIV), a sample of injecting drug users is obtained by following social links from one member of the population to another.

Some obvious questions for such studies are how best to obtain the sample and make the observations and, once the sample data are in hand, how best to use them to estimate the characteristic of the whole population. Obtaining the observations involves questions of sample size, how to select the sample, what observational methods to use, and what measurements to record. Getting good estimates with observations means picking out the relevant aspects of the data, deciding whether to use auxiliary information in estimation, and choosing the form of the estimator.

Sampling is usually distinguished from the closely related field of *experimental design*, in that in experiments one deliberately perturbs some part of a population in order to see what the effect of that action is. In sampling, more often one likes to find out what the population is like without perturbing or disturbing it. Thus, one hopes that the wording of a questionnaire will not influence the respondents'

Sampling, Third Edition. Steven K. Thompson.

opinions or that observing animals in a population will not significantly affect the distribution or behavior of the population.

Sampling is also usually distinguished from *observational studies*, in which one has little or no control over how the observations on the population were obtained. In sampling one has the opportunity to deliberately select the sample, thus avoiding many of the factors that make data observed by happenstance, convenience, or other uncontrolled means "unrepresentative."

More broadly, the field of sampling concerns every aspect of how data are selected, out of all the possibilities that might have been observed, whether the selection process has been under the control of investigators or has been determined by nature or happenstance, and how to use such data to make inferences about the larger population of interest. Surveys in which there is some control over the procedure by which the sample is selected turn out to have considerable advantages for purposes of inference about the population from which the sample comes.

1.1. BASIC IDEAS OF SAMPLING AND ESTIMATION

In the basic sampling setup, the population consists of a known, finite number N of units—such as people or plots of ground. With each unit is associated a value of a variable of interest, sometimes referred to as the *y-value* of that unit. The y-value of each unit in the population is viewed as a fixed, if unknown quantity—not a random variable. The units in the population are identifiable and may be labeled with numbers $1, 2, \ldots, N$.

Only a sample of the units in the population are selected and observed. The data collected consist of the y-value for each unit in the sample, together with the unit's label. Thus, for each hole drilled in the oil reserve, the data not only record how much oil was found but also identify, through the label, the location of the hole. In addition to the variable of interest, any number of auxiliary variables, such as depth and substrate types, may be recorded. In a lichen survey, auxiliary variables recorded could include elevation, presence of other vegetation, or even "eyeball" estimates of the lichen biomass. In an opinion poll, auxiliary variables such as gender, age, or income class may be recorded along with the opinions.

The procedure by which the sample of units is selected from the population is called the *sampling design*. With most well-known sampling designs, the design is determined by assigning to each possible sample s the probability $\mathrm{P}(s)$ of selecting that sample. For example, in a simple random sampling design with sample size n, a possible sample s consists of a set of n distinct units from the population, and the probability $\mathrm{P}(s)$ is the same for every possible sample s. In practice, the design may equivalently be described as a step-by-step procedure for selecting units rather than the resulting probabilities for selecting whole samples. In the case of simple random sampling, a step-by-step procedure consists of selecting a unit label at random from $\{1, 2, \ldots, N\}$, selecting the next unit label at random from the remaining numbers between 1 and N, and so on until n distinct sample units are selected.

The entire sequence $y_1, y_2, \ldots, y_N$ of y-values in the population is considered a fixed characteristic or parameter of the population in the basic sampling view. The usual inference problem in sampling is to estimate some summary characteristic of the population, such as the mean or the total of the y-values, after observing only the sample. Additionally, in most sampling and estimation situations, one would like to be able to assess the accuracy or confidence associated with estimates; this assessment is most often expressed with a confidence interval.

In the basic sampling view, if the sample size were expanded until all N units of the population were included in the sample, the population characteristic of interest would be known exactly. The uncertainty in estimates obtained by sampling thus stems from the fact that only part of the population is observed. While the population characteristic remains fixed, the estimate of it depends on which sample is selected. If for every possible sample the estimate is quite close to the true value of the population characteristic, there is little uncertainty associated with the sampling strategy; such a strategy is considered desirable. If, on the other hand, the value of the estimate varies greatly from one possible sample to another, uncertainty is associated with the method. A trick performed with many of the most useful sampling designs—cleverer than it may appear at first glance—is that this variability from sample to sample is estimated using only the single sample selected.

With careful attention to the sampling design and using a suitable estimation method, one can obtain estimates that are unbiased for population quantities, such as the population mean or total, without relying on any assumptions about the population itself. The estimate is unbiased in that its expected value over all possible samples that might be selected with the design equals the actual population value. Thus, through the design and estimation procedure, an unbiased estimate of lichen biomass is obtained whether lichens are evenly distributed throughout the study area or are clumped into a few patches. Additionally, the random or probability selection of samples removes recognized and unrecognized human sources of bias, such as conscious or unconscious tendencies to select units with larger (or smaller) than average values of the variable of interest. Such a procedure is especially desirable when survey results are relied on by persons with conflicting sets of interests—a fish population survey that will be used by fishery managers, commercial fishermen, and environmentalists, for instance. In such cases, it is unlikely that all parties concerned could agree on the purposive selection of a "representative" sample.

A probability design such as simple random sampling thus can provide unbiased estimates of the population mean or total and also an unbiased estimate of variability, which is used to assess the reliability of the survey result. Unbiased estimates and estimates of variance can also be obtained from unequal probability designs, provided that the probability of inclusion in the sample is known for each unit and for pairs of units.

Along with the goal of unbiased or nearly unbiased estimates from the survey come goals of precise or low-variance estimates and procedures that are convenient or cost-effective to carry out. The desire to satisfy as many of these goals as possible under a variety of circumstances has led to the development of widely used sampling designs and estimation methods, including simple random and unequal probability sampling; the use of auxiliary information; stratified, systematic, cluster, multistage, and double sampling; and other techniques.

1.2. SAMPLING UNITS

With many populations of people and institutions, it is straightforward to identify the type of units to be sampled and to conceive of a list or frame of the units in the population, whatever the practical problems of obtaining the frame or observing the selected sample. The units may be people, households, hospitals, or businesses. A complete list of the people, households, medical institutions, or firms in the target population would provide an ideal frame from which the sample units could be selected. In practice, it is often difficult to obtain a list that corresponds exactly to the population of interest. A telephone directory does not list people without telephones or with unlisted numbers. The set of all possible telephone numbers, which may be sampled by random dialing, still does not include households without telephones. A list of public or private institutions may not be up-to-date.

With many other populations, it is not so clear what the units should be. In a survey of a natural resource or agricultural crop in a region, the region may be divided into a set of geographic units (*plots* or *segments*) and a sample of units may be selected using a map. However, one is free to choose alternative sizes and shapes of units, and such choices may affect the cost of the survey and the precision of estimators. Further, with a sampling procedure in which a point location is chosen at random in a study region and sample units are then centered around the selected points, the sample units can potentially overlap, and hence the number of units in the population from which the sample is selected is not finite.

For an elusive population with detectability problems, the role of units or plots may be superseded by that of detectability functions, which are associated with the methods by which the population is observed and the locations are selected for making the observations. For example, in selecting the locations of line transects in a bird survey and choosing the speed at which they are traversed, one determines the “effective areas” observed within the study area in place of traditional sampling units or plots.

In some sampling situations the variable of interest may vary continuously over a region. For example, in a survey to assess the oil reserves in a region, the variable measured may be the depth or core volume of oil at a location. The value of such a variable is not necessarily associated with any of a finite set of units in the region, but rather, may be measured or estimated either at a point or as a total over a subregion of any size or shape.

Although the foregoing sampling situations go beyond the framework of a population divided uniquely into a finite collection of units from which the sample is selected, basic sampling design considerations regarding random sampling, stratified sampling, and other designs, and estimation results on design-unbiased estimation, ratio estimation, and other methods still apply.

1.3. SAMPLING AND NONSAMPLING ERRORS

The basic sampling view assumes that the variable of interest is measured on every unit in the sample without error, so that errors in the estimates occur only because just part of the population is included in the sample. Such errors are referred to as *sampling errors*. But in real survey situations, nonsampling errors may arise also. Some people in a sample may be away from home when phoned or may refuse to answer a question on a questionnaire, and such nonrespondents may not be typical of the population as a whole, so that the sample tends to be unrepresentative of the population and the estimates are biased. In a fish survey, some selected sites may not be observed due to rough weather conditions; sites farthest from shore, which may not be typical of the study region as a whole, are the most likely to have such weather problems.

The problem of nonresponse is particularly pronounced in a survey with a very low response rate, in which the probability of responding is related to the characteristic to be measured—magazine readership surveys of sexual practices exemplify the problem. The effect of the nonresponse problem may be reduced through additional sampling effort to estimate the characteristics of the nonresponse stratum of the population, by judicious use of auxiliary information available on both responding and nonresponding units, or by modeling of the nonresponse situation. But perhaps the best advice is to strive to keep nonresponse rates as low as possible.

Errors in measuring or recording the variable of interest may also occur. Quality-control effort throughout every stage of a survey is needed to keep errors to a minimum. In some situations, it may be possible to model measurement errors separately from sampling issues in order to relate the observations to population characteristics.

Detectability problems are a type of nonsampling error that occurs with a wide range of elusive populations. On a bird survey, the observer is typically unable to detect every individual of the species in the vicinity of a sampling site. In a trawl survey of fish, not every fish in the path of the net is caught. Nor is every homeless person in a society counted in a census. A number of special techniques, including line transect, capture–recapture, and related methods, have been developed for estimating population quantities when detectability problems are a central issue.

1.4. MODELS IN SAMPLING

In the basic sampling view the population is a finite set of units, each with a fixed value of the variable of interest, and probability enters only through the design, that is, the procedure by which the sample of units is selected. But for some populations it may be realistic and of practical advantage to consider a probability model for the population itself. The model might be based on knowledge of the natural phenomena influencing the distribution of the type of population or on a pragmatic statistical model summarizing some basic characteristics of such populations.

For example, a regression model may empirically describe a relationship between a variable of interest, the yield of a horticultural crop, say, with an auxiliary variable, such as the median level of an air pollutant. The model relating the variable of interest with the auxiliary variable has implications both for how to design the survey and how to make estimates.

In spatial sampling situations, the existence of correlations between values of the variable of interest at different sites, depending on the distance between the sites, has implications for choices regarding sampling design, estimation or prediction, and observational method. A model-based approach utilizing such correlation patterns has been particularly influential in geological surveys of mineral and fossil-fuel resources. In ecological surveys, such correlation patterns have implications not only for the spatial selection of observational sites, but for the observational methods (including plot shapes) used.

Ideally, one would like to be able to use a model of the population without having all conclusions of the survey depend on the model's being exactly true. A "robust" approach to sampling uses models to suggest efficient procedures while using the design to protect against departures from the model.

1.5. ADAPTIVE AND NONADAPTIVE DESIGNS

Surveys of rare, clustered populations motivate a further advance beyond the basic view of a sampling design. In adaptive sampling designs, the procedure for selecting sites or units on which to make observations may depend on observed values of the variable of interest. For example, in a survey for estimating the abundance of a natural resource, additional sites may be added to the sample during the survey in the vicinity of high observed abundance. Such designs have important applications to surveys of animal, plant, mineral, and fossil-fuel resources and may also have applications to other fields such as epidemiology and quality control.

The main purpose of adaptive procedures is to achieve gains in precision or efficiency, compared to conventional designs of equivalent sample size, by taking advantage of observed characteristics of the population. Adaptive procedures include such procedures as sequential stopping rules and sequential allocation among strata—procedures that have been rather heavily studied outside the finite-population context in the field of sequential analysis. With the population units identifiable as in the sampling situation, the possibilities for adaptive procedures are even greater, since it is possible to decide during a survey not just how many units to sample next but exactly which units or group of units to sample next.

In adaptive cluster sampling, whenever an observed value of the variable of interest satisfies a given criterion—for example, high abundance of animals observed at a site—units in the neighborhood of that unit (site) are added to the sample. A number of variations on this type of design are described in the final chapters of this book. For some populations, the designs produce remarkable increases in efficiency and appear to be particularly effective for sampling rare, clustered populations.

The sampling design is given for a conventional or nonadaptive design by a probability $\mathrm{P}(s)$ of selecting any particular sample s. For an adaptive design, the probability of selecting a given sample of units is $\mathbf{P}(s|\mathbf{y})$, that is, the probability of selecting sample s is conditional on the set $\mathbf{y}$ of values of the variable of interest in the population. Of course, in practice, the selection procedure can depend only on those values already observed.

Many natural populations tend to aggregate into fairly small portions of the study region, but the locations of these concentrations cannot be predicted prior to the survey. An effective adaptive design for such a population can result in higher selection probabilities assigned to samples that have a preponderance of units in those concentration areas. While the primary purpose of such a design may be to obtain a more precise estimate of the population total, a secondary benefit can be a dramatic increase in the yield of interesting observations—for example, more animals seen or more of a mineral obtained. Once adaptive designs are considered, the scope and potential of sampling methodology widens considerably.

1.6. SOME SAMPLING HISTORY

In the earliest known European nonfiction book, *The Histories* (ca. 440 B.C.), the author Herodotus describes a sampling method used by a Persian king to estimate the number of his troops during an invasion of Greece. A sample group of a fixed number of soldiers was instructed to stand as close together as possible and the area in which they had stood was enclosed by a short fence. Then the entire army was marched through, filling the enclosure group by group, and the number of groups required was tabulated. Multiplying the number of groups by the number in the sample group gave the estimated size of the whole force. No attempt was made to assess the accuracy of the estimate, and no description is given of how the initial sample group was selected. In fact, historians believe that the estimate reported, 1,700,000, was a gross overestimate based on present knowledge regarding feasible sizes of populations and armies at that time. Even so, the sampling strategy appears to be a fairly sensible use of an expansion estimator, and the recorded overestimate may have more to do with military propagandizing or to Herodotus's enthusiasm for large numbers than to sampling variability or bias.

> This place seemed to Xerxes a convenient spot for reviewing and numbering his soldiers; which things accordingly he proceeded to do... .What the exact number of the troops of each nation was I cannot say with certainty—for it is not mentioned by any one—but the whole land army together was found to amount to one million seven hundred thousand men. The manner in which the numbering took place was the following. A body of ten thousand men was brought to a certain place, and the men were made to stand as close together as possible; after which a circle was drawn around them, and the men were let go: then where the circle had been, a fence was built about the height of a man's middle; and the enclosure was filled continually with fresh troops, till the whole army had in this way been numbered. When the numbering was over, the troops were drawn up according to their several nations. (*The History of Herodotus, Book VII*, translated by George Rawlingson, The Internet Classics Archive by Daniel C. Stevenson, Web Atomics, 1994–2000, http:classics.mit.edu/Herodotus/history.html)

Many of the specific sampling designs and estimation methods in wide use today were developed in the twentieth century. Early in the twentieth century there was considerable debate among survey practitioners on the merits of random sampling versus purposively trying to select the most "representative" sample possible. The basic methods and formulas of simple random sampling were worked out in the first two decades of the century. An article by Neyman (1934) compared the two methods and laid out the conceptual basis for probability sampling, in which the sample is selected at random from a known distribution. Most standard sampling designs—stratified sampling, systematic sampling, cluster sampling, multistage sampling, and double or multiphase sampling—had been introduced by the end of the 1930s. The U.S. Census introduced probability sampling methods when it took over the sample survey of unemployment in the early 1940s. Unequal probability designs were introduced in the 1940s and 1950s.

The theory and methods of sampling have continued to develop and expand throughout the second half of the twentieth and the early twenty-first centuries. Studies in the theory of sampling by Godambe and others from the early 1950s forward have helped clarify the inference issues in sampling and have opened the way for subsequent development of new methods. A number of new designs and inference methods have been introduced in response to difficult problems in studies of natural and human populations, with contributing developments coming from many fields. Differences of opinion over design-based versus model-based approaches in sampling have led to the development of methods that combine both approaches. Recent developments in the field of missing data analysis have opened up new analysis methods and underscored the importance of how observed data are selected from the potential observations.

More detailed notes on the history of sampling are found in Bellhouse (1988b), Hansen et al. (1985), and Kruskal and Mosteller (1980). Some general references to sampling or specific aspects of sampling include Barnett (1991), Bart et al. (1998), Bolfarine and Zacks (1992), Chaudhuri and Stenger (1992), Cochran (1977), Foreman (1991), Ghosh and Meeden (1997), Govindarajulu (1999), Hansen et al. (1953), Hedayat and Sinha (1991), Kish (1965), Lohr (1999), Orton (2000), Raj (1968), Rubin (1987), Sampath (2001), Särndal et al. (1992), Schreuder et al. (1993), Sukhatme and Sukhatme (1970), M. E. Thompson (1997), Thompson and Seber (1996), Tryfos (1996), and Yates (1981).