Chapter 5: Data Collection and Sampling

Chapter 5: Data Collection and Sampling

1

Introduction

In real life, calculating parameters is virtually impossible because populations tend to be very large. As a result, most population parameters are not only unknown, but also unknowable.

The problem of statistical inference is to obtain information about the value of population parameters needed to make decisions.

Rather than investigating an entire population, we select a sample, collect information on a variable of interest, and calculate the sample statistic.

We expect the sample statistic to be close to the population parameter. In Chapters 10-11 we learn how to quantify the accuracy of the estimate.

In this chapter, we consider various sources for collecting data, before discussing the basic concepts and techniques of sampling.

2

Methods of Collecting

Recollect how the role of statistics is to convert data into quantifiable information, and that data are the observed values of a variable of interest.

There are many methods used to collect or obtain data for statistical analysis. Three of the most popular methods are:

a) **Direct observation** leads to the collection of observational data. The advantage of direct observation is that it is relatively inexpensive, but we cannot assess causation because of the presence of spurious variables.

b) **Experiments** are more expensive but produce better data because they can be designed to limit the influence of spurious variables through control and the manipulation of the variables of interest.

c) **Surveys** are a very popular method for collecting data, and they are discussed in more detail in the next slide.

3

Survey

- A **survey** solicits information from people; e.g., Gallup polls, pre-election polls, and marketing surveys.
- The **response rate** (i.e., the proportion of all people selected who complete the survey) is a key survey parameter.
- Surveys may be administered in a variety of ways:
  - A **personal interview** is the best approach because it has a higher response rate and fewer incorrect answers, but it requires a trained interviewer to avoid response bias.
  - A **telephone interview** is less expensive, but also less personal and has a lower response rate. Cell phones are also more difficult than land lines to include in survey.
  - A **self-administered survey** is usually mailed to a sample of people. It is inexpensive and attractive when a large sample is required, but it suffers from low response rates and high rates of incorrect answers.

4

Chapter 5: Data Collection and Sampling

Questionnaire Design

Whether a questionnaire is self-administered or completed by an interviewer, it must be well designed. Proper questionnaire design takes knowledge, experience, time, and money. Some basic points to consider regarding questionnaire design follow:

1. Keep the questionnaire as short as possible.
2. Ask short, simple, and clearly worded questions.
3. Start with demographic questions to help respondents get started comfortably.
4. Use dichotomous (yes/no) and multiple-choice questions.
5. Use open-ended questions cautiously.
6. Avoid using leading-questions.
7. Pretest a questionnaire on a small number of people.
8. Think about how you intend to use the collected data when preparing the questionnaire.

5

Sampling

- Statistical inference permits us to draw conclusions about a parameter based on a sample that is quite small in comparison to the size of the population.
- A sample statistic is then used as an **estimate** of the population parameter.
- For the inference to be accurate, the **sampled population** should be very close to the **target population**.
- **Sampling**, the process of selecting a subset of the sampled population, is often done for reasons of **cost** and **practicality**.
- **Self-selected samples** are almost always biased because the individuals who participate in them are more keenly interested in the issue than are the other members of the population.

6

Sampling Plans

- A sampling plan is a procedure for specifying how a sample will be drawn from a population.
- We will focus our attention on three sampling plans:
  - Simple random sampling,
  - Stratified random sampling
  - Cluster sampling
- Determining the appropriate sample size will be addressed in detail in Chapters 10 and 12. Until then, let us just say that the larger the sample size is the more accurate we can expect the sample estimates to be.

7

Simple Random Sampling

- A **simple random sample** is a sample selected in such a way that every possible sample of the same size is equally likely to be chosen.

Example:

Drawing three names from a hat containing all the names of the students in the class is an example of a simple random sample because any group of three names is as equally likely as picking any other group of three names.

8

Chapter 5: Data Collection and Sampling

Stratified Random Sampling

- A **stratified random sample** is obtained by separating the population into mutually exclusive sets, or strata, and then drawing simple random samples from each stratum.
- Example: we can draw random samples from each of the four income groups according to their proportions in the population.

Stratum	Income category ($)	Population proportion (%)	Sample size n = 400
1	Less than 25,000	25	100
2	25,000 – 39,999	40	160
3	40,000 – 60,000	30	120
4	More than 60,000	5	20

- The only problem with this approach is that if the sample size is too small, the smallest stratum may be underrepresented.

9

Stratified Random Sampling Once Again

The population is first divided into groups of elements called strata.

Each element in the population belongs to one and only one stratum.

Best results are obtained when the elements within each stratum are as much alike as possible (i.e. a homogeneous group).

A simple random sample is taken from each stratum.

Formulas are available for combining the stratum sample results into one population parameter estimate.

Advantage: If strata are homogeneous, this method is as "precise" as simple random sampling but with a smaller total sample size.

Example: The basis for forming the strata might be department, location, age, industry type, and so on.

10

Cluster Sampling

- A **cluster sample** is a simple random sample of groups or clusters of elements.

Example:

- To estimate the average annual household income in a large city, we let each block within the city represent a cluster. A sample of clusters could then be randomly selected, and every household within these clusters could be questioned to determine income.
- Cluster sampling may increase sampling error (see Section 5-4) because elements belonging to the same cluster are likely to be similar in many respects, including the variable of interest.

11

Cluster Sampling Once Again

The population is first divided into separate groups of elements called clusters.

Ideally, each cluster is a representative small-scale version of the population (i.e. heterogeneous group).

A simple random sample of the clusters is then taken.

All elements within each sampled (chosen) cluster form the sample.

Example: A primary application is area sampling, where clusters are city blocks or other well-defined areas.

Advantage: The close proximity of elements can be cost effective (i.e. many sample observations can be obtained in a short time).

Disadvantage: This method generally requires a larger total sample size than simple or stratified random sampling.

12

Chapter 5: Data Collection and Sampling

Sampling Errors

- **Sampling error** refers to difference existing between a given sample statistic and the population parameter because of the particular set of observations selected for the sample.
- The sampling error decreases as the sample size becomes larger.

Example:

- Suppose that we wish to estimate the mean annual income of North American blue-collar workers making an inference from the annual incomes of a sample of blue-collar workers drawn from the North American population.
- The sampling error is the difference between the sample mean annual income and the population mean annual income.

13

Systematic Sampling

If a sample size of n is desired from a population containing N elements, we might sample one element for every n/N elements in the population.

This method has the properties of a simple random sample, especially if the list of the population elements is a random ordering.

We randomly select one of the first n/N elements from the population list.

We then select every n/Nth element that follows in the population list.

Advantage: The sample usually will be easier to identify than it would be if simple random sampling were used.

Example: Selecting every 100th listing in a telephone book after the first randomly selected listing

14

Convenience Sampling

It is a nonprobability sampling technique. Items are included in the sample without known probabilities of being selected.

The sample is identified primarily by convenience.

Example: A professor conducting research might use student volunteers to constitute a sample.

Advantage: Sample selection and data collection are relatively easy.

Disadvantage: It is impossible to determine how representative of the population the sample is.

15

Judgment Sampling

The person most knowledgeable on the subject of the study selects elements of the population that he or she feels are most representative of the population.

It is a nonprobability sampling technique.

Example: A reporter might sample three or four senators, judging them as reflecting the general opinion of the senate.

Advantage: It is a relatively easy way of selecting a sample.

Disadvantage: The quality of the sample results depends on the judgment of the person selecting the sample.

16

Chapter 5: Data Collection and Sampling

Nonsampling Errors

Nonsampling errors are more serious and are due to mistakes made in the acquisition of data or due to improperly selected sample observations.

Increasing the sample size will not reduce a nonsampling error.

There are three types of nonsampling errors:

- **Errors in data acquisition** arise from the recording of incorrect responses caused by inaccurate measurements, transcription mistakes, inaccurate data recording, or inaccurate responses to questions.
- **Nonresponse errors** happen when the collected sample observations is not representative of the target population, resulting in biased results.
- **Selection bias** occurs when part of the target population is not included in the sampling plan.

17

Recommendation

It is recommended that probability sampling methods (simple random, stratified, cluster, or systematic) be used.

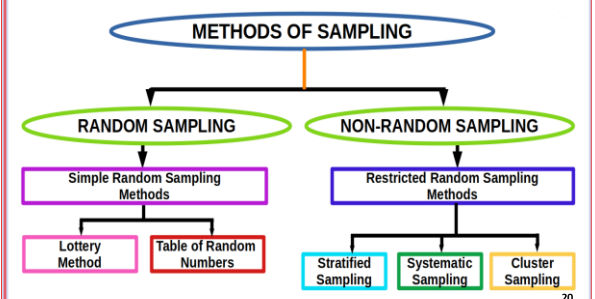
For these methods, formulas are available for evaluating the "goodness" of the sample results in terms of the closeness of the results to the population parameters being estimated.

An evaluation of the goodness cannot be made with non-probability (convenience or judgment) sampling methods.

18

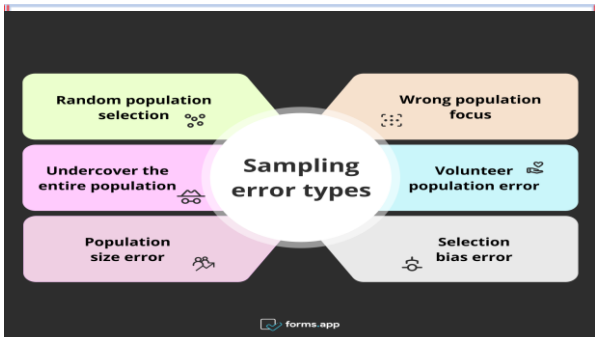
- It is extremely costly and impractical to investigate each member of the population to determine the values of the parameters.
- **Sampling** consists of collecting a sample from the population and use the sample statistics to draw inferences about the parameters.
- Care should be taken to ensure that the **sampled population** is the same as the **target population**, or bias such as nonresponse error may arise.
- There are different sampling plans, such as **simple random sampling**, **stratified random sampling**, and **cluster sampling**.
- Regardless the sampling plan, it is important to realize that both **sampling error** and **nonsampling error** may occur and to understand what the sources of these errors are.

19



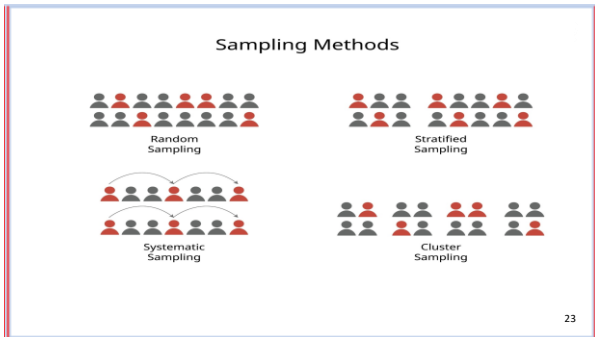
20

Chapter 5: Data Collection and Sampling

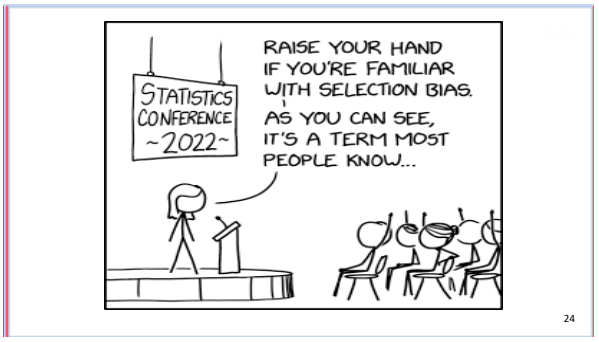


21

22



23



24