1 INTRODUCTION

- The courses notes for this section cover Sections 1.2, 1.3, 1.4 and 1.6 of Chapter 1 from the Lohr textbook plus additional notes of my own. The numbering of sections in the course notes do not, however, match the numbering in the text.

Additional References

(C&C) *Statistics: Concepts and Controversies*, Moore and Notz (2009), W.H. Freeman.

(SREL) *Statistical Reasoning for Everyday Life*, Bennett, Briggs, and Triola (2001), Addison-Wesley.

(STT) *Statistical Tricks and Traps*, E.C. Almer (2000), Pyrczak Publishing.

1.1 Requirements of a Good Sample

- Sampling is a process that selects a part of a population (via some mechanism or plan) for observation. Typically, the goal is to estimate one or more characteristics about the population based on information contained in the sample.

- Two important sampling issues: (i) What is the best way to obtain a sample? (ii) How do we use the information contained in the sample to estimate the population characteristic(s) of interest? Question (i) is the design problem and question (ii) is the analysis problem.

- The design problem involves questions about the method of selecting sampling units, the sample size, the information (measurements) to be recorded for each sampling unit, and the observational methods used to collect data from a sampling unit.

- Example: Suppose the goal is to find the average age of students enrolled at Montana State University – Bozeman.

  - *What sampling plan will be used?* The researcher decides to use simple random sampling (SRS).
  - *How many sampling units will be sampled?* The researcher has enough time and money to sample 500 students.
  - *What information will be collected?* The researcher will collect the date of birth of the student.
  - *What method will be used to collect the date of birth?* The researcher will ask the student to write his or her date of birth on a questionnaire form.

- By properly addressing the design problem, the analysis problem will often not be difficult. That is, given the design, the researcher should know the format of the analysis (and possibly an alternative analysis) prior to data collection.

- In a survey, a sample from a finite population of interest is selected to represent the whole population.

- In a census, every member of a finite population is observed.

1.2 Populations and Sampling Units

- The **target population** is the set of individuals about which information is desired (i.e., the set of individuals the researcher would like to study).
- The **sampled population** or **study population** is the set of individuals a researcher *intends* to study (i.e., the set of individuals that could possibly be included in the sample).
  - Suppose the goal is to find the average age of all students who are currently enrolled at MSU-Bozeman. Then the target population is the set of all students who are currently enrolled at that university.
  - A random sample of students attending the university and who are currently on the MSU-Bozeman campus is taken. If there are students who are temporarily studying at another university (e.g. as part of an exchange program in another country), then the study (sampled) population of students does not match the target population. There will be students enrolled at the university that had no chance of being in the sample.
  - Thus, the target and study populations do not match. Also, what is the formal definition of "currently enrolled"?
- It is certainly desirable for the target and study populations to match. However, this is often not the case. When they do not match, it may not be possible to make statements about the target population from data collected from the study population. That is, the scope of inferences or conclusions will be restricted to only the study population.
  - In the MSU-Bozeman example, the *scope of inference* is limited to the study population.
- Unless otherwise specified, we will assume the target and study populations are the same.
- The potential members or units of a sample are the **sampling units**. A **sampling frame** is a complete specification of the sampling units from the population of potential sampling units.
- Thus, a **sample** is a collection of sampling units drawn from a sampling frame.
- Note: At this point I need to clarify the distinction between 'individuals in a population' and 'sampling units in the sampling frame'.
  - If the sampling frame consists of individuals in the study population then the sampling units and the individuals in the study population are the same.
  - However, a sampling plan could consist of sampling subgroups of individuals. In this case, the sampling units are subgroups of individuals. For example, it is common to sample households (which often contain more than one person per household). The household is the sampling unit and the set of all households forms the sampling frame. This is an example of 'cluster sampling'.
- For many populations, the sampling unit is obvious. It is necessary to conceptually form the sampling frame of population units. Often it is necessary to record additional information that allows different classification of sampling units (e.g, for 'stratification').

- For other populations, the sampling unit may not be obvious. For example, when surveying a geographical region, you may have to use a map to identify what is the basic sampling unit. This introduces a number of problems, in particular:
  - The numerous alternative size and shape combinations for a sampling unit.
  - A discrepancy between a sampling frame and the researcher's inability to access certain sampling units (e.g., too costly, remote, or dangerous to access).
  - A complete list of units or classification information is not available.

1.3 Estimates vs Estimators

- The goal of sampling is to make conclusions about some characteristics of interest for one or more populations of interest based on the data collected. This process of making conclusions is called **statistical inference**.
- A **parameter** is a value which describes some characteristic of a population (or possibly describes the entire population). Examples: the population mean $\overline{y}_U$ (or μ) or the population variance S^2 (σ^2). A **statistic** is a value that can be computed from the data without knowing the values of any parameters.
- In general, the value of a population parameter is unknown. Statistics computed from data can provide information about the unknown parameter.
- The process of estimating a population parameter by a statistic derived from survey or experimental data is called **point estimation**.
  - Prior to data collection, a sample statistic is a random variable and is called a **point estimator** of a parameter. For example, $\overline{Y} = \dfrac{Y_1 + Y_2 + \cdots + Y_n}{n}$ is a *point estimator* for $\overline{y}_U$ (or μ) where $Y_1, Y_2, \ldots, Y_n$ are random variables.
  - After collecting a sample, a sample statistic is no longer a random variable but is a realization of the point estimator and is called a **point estimate** of the population parameter. For example, $\overline{y} = \dfrac{y_1 + y_2 + \cdots + y_n}{n}$ is a *point estimate* for $\overline{y}_U$ (μ) where $y_1, y_2, \ldots, y_n$ are observed data values.
- Later in the course we will discuss *interval estimation* in the form of confidence intervals.
- The researcher's sampling goal is to collect a sample that is **representative**. Specifically, parameters of interest can be estimated from the sample data with a known degree of *accuracy* and *precision*.
- Accuracy is related to 'bias' and precision is related to 'variability'. These concepts will be discussed later in more detail.

1.4 Sampling and Estimation Concepts

- In one of the most common sampling situations, we assume the population consists of a finite number N of sampling units. The units in the population are identifiable and can be labeled $1, 2, \ldots, N$.

- Associated with each of the N units is a measurable value related to the population characteristic of interest (often referred to as the y-values of the units).
- Each y-value is considered a fixed quantity representing that unit. In other words, we assume the sequence of population y-values $(y_1, y_2, \ldots, y_N)$ is fixed.
- For each sampled unit there will be a unit label, its y-value, and any other recorded and potentially useful auxiliary variables (e.g., elevation, temperature, and precipitation when studying environmental or ecological problems, or, age, income, and gender when studying human populations).
- A **sampling design** is the procedure by which a sample of units is selected from the population.
- The classical sampling designs (e.g., simple random, stratified, cluster, systematic) require that randomness be built into the sampling design so that its estimators can be assessed probabilistically. For example, we can make statements like "Our estimate is unbiased." or "We are 95% confident that our estimate will be within 2 percent of the true proportion.".
- Sampling designs that are based on planned randomness are called **probability samples**. More formally, the design is determined by assigning to every possible sample $\mathcal{S}$ a sample probability $P(\mathcal{S})$ that equals the probability of actually selecting that sample.
  - When taking a simple random sample (SRS) of size n, the possible samples consist of n distinct units selected from the population of N units, and $P(\mathcal{S})$ is the same for every possible sample $\mathcal{S}$. Thus,

    $P(\mathcal{S}) = 1/(\text{the total number of unique samples of size } n)$
- The typical inference problems in sampling are (i) the estimation of some population characteristic based only on the sample data (point estimation) and (ii) an assessment of the variability associated with estimates. This variability assessment is often an **interval estimate** expressed in terms of a confidence interval.
- Ideally, we would like a sampling strategy which will yield samples that produce estimates with small variability that are centered around the true value). In other words, we want 'high precision' and 'high accuracy' (or little or no bias).
- Thus, by choosing an appropriate sampling design and estimation method, the researcher can often obtain unbiased estimates without making additional assumptions about the population.
- Selection by use of probability samples removes intentional or unintentional human sources of bias (such as a tendency to select units with larger or smaller than average values). Use of probability samples to generate a representative sample is especially desirable when there are parties with conflicting interests (e.g., a fish population study that will be used by fishery management, commercial users, and environmental groups).

1.5 Selection Bias

- A sampling scheme displays **bias** if the sample systematically favors certain parts of the population over other parts. If certain parts of the population are overrepresented in the sample, while other parts are underrepresented, then sampling bias exists.

  - (STT) There's a saying that "much of what we know about psychology is based on the behavior of college sophomores". This is because students who take introductory psychology are often required to participate in psychological studies. Various types of studies are announced, and students self-select a study in which to participate. Often an entire line of investigation is based solely on such students. For example, a team of researchers conducted a study on lying in various types of relationships, and in their introduction, they pointed out that their "community sample" was the first "in the literature on lying in everyday life that is not a group consisting solely of college students". When the goal of psychologists is to study broader issues with reference to the general population (which is beyond the scope of inference), their data is contaminated by a bias in favor of college students. (DePaulo, B.M. & Kashy, D.A. (1998) Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology* 74: 63-79)
  - (STT) Unfortunately, there are also cases of deliberate bias. That is, someone deliberately selects respondents who are most likely to support a certain outcome. For example, researchers at the University of Minnesota found that some schools try to keep students with learning disabilities from participating in statewide testing. In some cases, these students are sent on field trips or are told to stay home on the testing day. To the extent this is true, the sample is biased against lower achievers, which results in higher overall scores. Higher scores, in turn, make the school look better. ("Why Johnny stayed home." *Newsweek*, October 6, 1997, p.60)

- **Selection bias** occurs whenever the researchers select their sample in a way that makes it unrepresentative of the population. Lohr (page 5) states:

  > Selection bias occurs when some part of the target population is not in the sampled population, or, more generally, when some population units are sampled at a different rate than intended by the investigator.

- Selection of whichever individuals are easiest to reach is called **convenience sampling**. The resulting sample (not surprisingly) is called a **sample of convenience**. This will often lead to selection bias.

  *** Discussion: Manufacturers and advertisers often use interviews at shopping malls to gather information about the habits of consumers and the effectiveness of advertisements. The main benefit is that taking a sample of mall shoppers is quick and inexpensive. Why would the sample from a mall provide biased results regarding consumer habits?

- **Participation bias** often occurs in **voluntary response surveys** (or **self-selected surveys**) in which people decide for themselves whether or not to be included in the survey. That is, the participants are not selected by the researchers conducting the study.

- Voluntary (self-selected) responses tend to over-represent certain portions of the population, and are commonly from people with very strong opinions (often negative opinions) about an issue or they desire a change from the status quo. The opinions of the respondents are unlikely to accurately represent the opinions of the majority which is less emotionally attached to the issue. Thus, a self-selected survey is subject to participation bias.

  *** Discussion: (SREL) The television show *Nightline* conducted a poll in which viewers were asked whether or not the United Nations (UN) headquarters should be kept in the United States (US). Viewers could respond to the poll by paying 50 cents to call a '900' phone number with their opinions. The poll drew 186,000 responses, of which 67% favored moving the UN out of the US. Around the same time, a poll using a simple random sample of 500 people found that 72% wanted the UN to *stay* in the US. Which poll is more likely to be representative of the general opinions of Americans?

- Selection bias can also occur when a *judgement sample* is taken. A **judgement sample** occurs when the researcher uses her/his so-called "expert" judgement to select what she/he judges to be a representative sample.

  – For example, a wildlife biologist decides to collect information about a bird species from only those areas that she considers to be typical or representative of the habitat type the bird would be found.

- *Undercoverage* and *overcoverage* are two other forms of selection bias.

- A sample suffers from **undercoverage** if the sample fails to represent the target population because it is not possible for specific members of the population to be included in the sample.

  – Example: You want to estimate the size of the elk population in a study area in Yellowstone. Certain areas are too remote to access and so are not in the sample by choice (not by random chance). From the data that is collected, you provide an estimate for the entire study area (which includes the inaccessible parts).

- A sample suffers from **overcoverage** when sampling units not in the target population are included in the sample.

  – Example: A local TV station wants to estimate the proportion of Gallatin County voters who support a referendum to increase property taxes to be used to improve the Gallatin County Fairgrounds. They have viewers vote online. In addition to the issues related to voluntary response, there is no way to guarantee the online participants are eligible to vote in the local referendum. Thus, there will be sampling units not in the target population of eligible voters that will be in the sample.

- Both undercoverage and overcoverage could lead to a serious bias in the survey results.

1.6 Nonresponse and Measurement Errors

- **Nonresponse errors** occur when the sampling unit does not yield a response of interest.

  – A questionnaire respondent refuses to provide certain information (e.g., regarding age, income, gender, etc.) or refuses to participate at all (e.g., phone surveys).

  - The measurement device breaks down. It gets too dark to take measurements. You run out of money. You fail to find the sampling unit.

- The problem of nonresponse is not solved by starting with an excess number of cases to allow for a certain level of nonresponse. A sample is no longer a probability sample if it is affected by nonresponse. There is no perfect substitute for response.

- However, if certain cases are strongly correlated with a tendency for nonresponse, it would not be surprising for nonresponse to seriously bias results and inferences about the population.

  - Suppose the question on a survey is 'Have you ever driven while intoxicated?'. I would expect that those who have actually driven while intoxicated are less likely to respond than those who have not.

- Note that bias due to nonresponse is not a form of selection bias. That is, the bias does not arise from the procedure used for collecting the sample. It occurs *after sampling units are selected*.

- Underoverage, overcoverage, and nonresponse errors are examples of **nonobservational errors**.

- **Measurement errors** or **observational errors** occur when the information from a sampling unit is faulty (does not yield the true response).

  - In a study area, the presence of a bird's nest is missed and 'absence' is recorded. This also happens when studying elusive populations.
  - The measuring device is uncalibrated or its user is improperly trained.
  - People lie when asked about sensitive issues (e.g., drug use, abortion, criminal history, income, eating disorders, sexual orientation, etc.).

- **Response bias** occurs when the responses taken from the sampling units tend to differ from the truth in one particular direction. That is, there is a tendency to have either more overstated values or more understated values in the sample in comparison to the true values.

  - A field ecologist records an estimate of the distance between a snowmobile and a group of elk. This ecologist tends to underestimate the true distance, so the recorded distances are understated values of the true distances.

- Measurement bias occurs in surveys when the response given by the respondent may have been influenced by improperly or poorly worded survey questions or by the behavior of the interviewer. Response bias can occur because respondents may lie when presented with questions of a personal nature or questions regarding illegal activity.

*** Discussion: (SREL) Two surveys asked Catholics in the Boston area whether contraceptives should be made available to unmarried women. The first survey involved in-person interviews and 44% of the respondents answered yes. The second survey as conducted by mail and by telephone, and 75% of the respondents answered yes. Which survey estimate was more likely to be reliable?

Discussion Comments

+++ **Mall surveys**: People at malls, however, are not representative of the entire U.S. population. Mall samples are inherently biased: they systematically overrepresent some parts of the population. For example, those in malls will (on average) have greater financial resources than those not at the mall and will also be more likely to be teenagers or retirees. Moreover, the interviewers tend to select neat, safe-looking individuals from the stream of customers. Thus, the more affluent, teens, and retirees tend to be overrepresented while others (like the 'shady-looking' or unkempt) are underrepresented. Thus, the opinions of such a convenience sample may be very different than the population as a whole.

+++ **Nightline**: The *Nightline* poll was severely biased. It had selection bias because its sample was drawn only from the show's viewers, rather than from all Americans. The poll itself was a self-selected (or voluntary) survey in which viewers not only chose whether or not to respond, but also had to pay 50 cents to participate. This cost made it even more likely that respondents would be those who felt the need for a change. Thus, despite the large number of respondents, the *Nightline* survey was highly unlikely to give meaningful results. In contrast, a simple random sample of 500 people is quite likely to be representative, so the finding of this small survey has a better chance of representing the true opinions of Americans. Note also that the *Nightline* poll would also permit non-Americans to participate and allow for people to participate multiple times if they so desired.

+++ **Contraceptives**: Contraception is a sensitive topic, particularly among Catholics because the Catholic Church officially opposes the use of contraceptives. Thus, the first survey with in-person interviews may have encouraged dishonest responses. The second survey made responses more confidential, and therefore was more likely to reflect the respondents' true opinions. Thus, the second survey was probably more accurate, but still would suffer from some response bias.

1.7 Sampling Errors

- In the ideal sampling setting, it is assumed that the variable of interest is measured without error on every sampling unit. That is, the recorded y-values are the true values associated with that sampling unit.
- We also hope that the sampled y-values are representative of the population of y-values. The natural variation in the sampled y-values is referred to as **sampling error** or **sampling variation**.
- It is unfortunate that most texts use the term "sampling error" because it misleads people into believing some mistake has been made. The intrinsic and unavoidable variability that exists in a population should not be interpreted as "error" in its common usage.
- In this ideal setting, any deviation between the estimate and the parameter value occurred solely because our sample contains only partial information about the population.
- A good sampling plan is necessary for good prediction, but there still is no guarantee of good prediction. Anyone can misuse good data. Prediction of future population characteristics, even on the basis of a complete and perfect census, can fail for many reasons (e.g., unreliable prediction methods and unforeseen future events).