Sampling Theory and Methods

Spring 2008

C. L. Williams

Syllabus and Lecture 1 for Introduction to Sampling Methods and Theory

Outline

Why use surveys?

Information on characteristics of populations is constantly needed by politicians, marketing departments of companies, public officials responsible for planning health and social services, and others. For reasons relating to timeliness and cost, this information is often obtained by use of sample surveys.

A Health Care Example

A health department in a large state is interested in determining the proportion of the state's children of elementary school age who have been immunized against the childhood infectious diseases (e.g., polio, diphtheria, tetanus, pertussis, etc.).

Our Goal in this course

Our goal in this course is to present a variety of methods for selecting a subset (a sample) from the original set of all measurements (the population) of interest to the researchers. It is the members of the sample who will be interviewed, studied, or measured. For example, in the problem stated above, the net effect of such methods will be that valid and reliable estimates of the **proportion of children** who have been immunized for these diseases could be obtained in the time frame specified and at a fraction of the cost that would have resulted if attempts were made to obtain the information concerning every child of elementary school age in the state.

We will consider
Summary Statistics

More formally, a sample survey may be defined as a study involving a subset (or sample) of individuals selected from a larger population. Variables or characteristics of interest are observed or measured on each of the sampled individuals. These measurements are then aggregated over all individuals in the sample to obtain *summary statistics (e.g., means, proportions, totals)* for the sample. It is from these summary statistics that extrapolations can be made concerning the entire population. The validity and reliability of these extrapolations depend on how well the sample was chosen and how well the measurements were made.

Why not a Census?

When all the individuals in the population are selected for measurement, the study is called a *census*. The summary statistics obtained from a census are not extrapolations, since every member of the population is measured. The validity of the resulting statistics, however, depends on how well the measurements are made. The main advantages of sample surveys over censuses lie in the reduced costs and greater speed made possible by taking measurements on a subset rather than on an entire population. In addition, studies involving complex issues requiring elaborate measurement procedures are often feasible only if a sample of the population is selected for measurement since limited resources can be allocated to getting detailed measurements if the number of individuals to be measured is not too great.

Mandated to collect data

In the United States, some government agencies are mandated to develop and maintain programs whereby sample surveys are used to collect data on the economic, social, and health status of the people, and these data are used for research purposes as well as for policy decisions.

- National Center for Health Statistics (NCHS), a center within the United States Department of Health and Human Services, is mandated by law to conduct a program of periodic and ongoing sample surveys designed to obtain information about illness, disability, and the utilization of health care services in the United States.

Other National Centers and Bureaus

- Bureau of Labor Statistics within the Department of Labor
- National Center for Educational Statistics within the Department of Education), which collect data relevant to the mission of their departments through a program of sample surveys.
- U.S. Bureau of the Census
- Centers for Disease Control.

Observational Studies

Sample surveys belong to a larger class of non-experimental studies generally given the name *"observational studies"* in the health or social sciences literature. Most sample surveys can be put in the class of observational studies known as *"cross-sectional studies."* Other types of observational studies include cohort studies and case-control studies. Cross-sectional studies are "snapshots" of a population at a single point in time, having as objectives either the estimation of the prevalence or mean level of some characteristics of the population or measurement of the relationship between two or more variables measured at the same point in time. Cohort and case-control studies are used for analytic rather than descriptive purposes.

Example Epidemiological study

Case control studies are used in epidemiology to test hypotheses about the association between exposure to suspected risk factors and the incidence of specific diseases. These study designs are widely used to gain insight into relationships.

Private sector or Business example

If you were to work for on of the large consumer data base companies you job may be to gather the delinquent accounts (e.g., the "cases") along with a sample of accounts that are not delinquent (e.g., the "controls"), and then compare the characteristics of each group for purposes of determining those factors that are associated with delinquency.

Estimation of proportion vaccinated

For example, in the hypothetical example, presented at the beginning of this chapter, the major objective is to *estimate*, through use of a sample, the proportion of all children of elementary school age who have been immunized for childhood diseases.

Descriptive surveys

In *descriptive surveys*, much attention is given to the selection of the sample since extrapolation is made from the sample to the population. Although hypotheses can be tested based on data collected from such descriptive surveys, this is generally a secondary objective in such surveys. *Estimation* is almost always the primary objective.

Four major components-Designed Sample Surveys

- Sample design,
- Survey measurements,
- Survey operations,
- Statistical analysis and report generation.

Sampling plan

In a sample survey, the major statistical components are referred to as the sample design and include both the sampling plan and the estimation procedures. The *sampling plan* is the methodology used for selecting the sample from the population. The *estimation procedures* are the algorithms or formulas used for obtaining estimates of population values from the sample, data and for estimating the reliability of these population estimates.

Collaborative effort

The choice of a particular sample design should be a collaborative effort involving input from the *statistician* who will design the survey, the persons involved in executing the survey, and those who will use the data from the survey. The data users should specify what variables should be measured, what estimates are required, what level of reliability and validity are needed for the estimates, and what restrictions are placed on the survey with respect to timeliness arid costs. Those individuals involved in executing the survey should furnish input about costs for personnel, time, and materials as well as input about the feasibility of alternative sampling and measurement procedures. Having received such input, the statistician can then propose a sample design that will meet the required specifications of the users at the lowest possible cost.

The Statistician's role

What you will learn in the course

Just as sampling and estimation are the statistician's responsibility in the design of a sample survey, the choice of measurements to be taken and the procedures for taking these measurements are the responsibility of those individuals who are experts in the subject matter of the survey and also of those individuals having expertise in the measurement sciences. The former (often called "subject matter persons") give the primary input into specifying the measurements that are needed in order to meet the objectives of the survey.

The psychologist's or sociologist's role

What you won't learn in this course

Once these measurements are specified, the the measurement experts -often **psychologists** or **sociologists** with special training and skills in survey research - begin designing the questionnaires or forms to be used in eliciting the data from the sample individuals. The design of a questionnaire or other survey instrument which is suitable for collecting valid and reliable data is often a very complex task; it requires considerable care and sometimes a preliminary study, especially if some of the variables to be measured have never been measured before by the survey process. Once the survey instruments have been drafted, the **statistician** provides input with respect to procedures to be used to evaluate and assure the quality

Pilot study

"A dry run"

Once the sample has been chosen and the measurement instruments or questionnaires drafted, pretested, and modified, the field work of the survey - including data collection - can begin. But before the data collection starts, there should be a dry run or *pilot survey* on a small sample, with the objective of testing the measurement instruments and eliminating any discernible imperfections in the survey procedures.

Valid and reliable results

What must be done

In order for the estimates from the survey to be valid and reliable, it is important that the data be collected in accordance with the survey design, and it is the task of those individuals responsible for survey operations to oversee and supervise the data collection procedures. The nature of the survey operations staff depends on the size and scope of the sample survey, the complexity of the measurements, and the nature of the survey

Our main interest

Statistical Analysis and Report Writing

After the data have been collected, coded, edited, and processed, the data can be analyzed statistically and the findings incorporated into a final report. As in all components of a sample survey, considerable care should be taken in the interpretation of the findings of the survey. These findings are in the form of estimated characteristics of the population from which the sample was taken. These estimates, however, are subject to both sampling and measurement errors, and any interpretation of the findings should take these errors into consideration.

Requirements of a good sample

Basic definitions

Some definitions are needed to make the notion of a good sample more precise.

- **Observation unit** An object on which a measurement is taken. This is the basic unit of observation, sometimes called an *element*. In studying human populations, observation units are often individuals.
- **Target population** The complete collection of observations we want to study. Defining the target population is an important and often difficult part of the study. For example, in a political poll, should the target population be all adults eligible to vote? All registered voters? All persons who voted in the last election? The choice of target population will profoundly affect the statistics that result.

Requirements of a good sample

Basic definitions, cont'd

- **Sample** A subset of a population.
- **Sampled population** The collection of all possible observation units that might have been chosen in a sample; the population from which the sample was taken.
- **Sampling unit** The unit we actually sample. We may want to study individuals but do not have a list of all individuals in the target population. Instead, households serve as the *sampling units*, and the observation units are the individuals living in the households.
- **Sampling frame** The list of sampling units. For telephone surveys, the sampling frame might be a list of all residential telephone numbers in the city; for personal interviews, a list of all street addresses; for an agricultural survey, a list of all farms or a map of areas containing farms.

Ideal Study

In an ideal survey, the sampled population will be identical to the target population, but this ideal is rarely met exactly. In surveys of people, the sampled population is usually smaller than the target population.

The Population

The population (*or universe or target population)* is the entire set of individuals to which findings of the survey are to be extrapolated. The terms universe, target population, and population are generally used interchangeably. The individual members of the population whose characteristics are to be measured are called elementary units or elements of the population.

Enumeration Units or listing units

The individual members of the population whose characteristics are to be measured are called elementary units or elements of the population. For example, if we are conducting a sample survey for purposes of estimating the number of persons living in South Carolina who have never visited a dentist, the universe consists of all persons living in South Carolina, and each person living in South Carolina is an elementary unit or element.

If we are conducting a sample survey of hospital medical records for purposes of estimating the number of hospital discharges in a given year having specified diagnoses, each hospital discharge occurring during the year is an element, and the totality of such discharges constitutes the universe or population.

An algorithmic approach

If a sample is to be drawn from a enumeration list (or list of enumeration units, it is necessary to specify by some algorithm the elementary units that are to be associated with each enumeration unit. In order to get a representative set an *enumeration* or *counting rule* must be used.

Primary purpose

The primary goal in conducting a survey is to estimate certain values relating to the distribution of characteristics of interest from a population. The most common of these are totals, means, aggregates (partial sums), proportions and or ratios. Measures of relative standing like percentiles, standard deviations or other distributional features could also be considered.

Some basic statistics-We all know

Population parameters

1. **Population Total** of a characteristic is generally denoted by T and is the sum of the values of the characteristic over all elements in the population. The population total is given by

$$X_T = \sum_{i=1}^{n} X_i$$

2. **Population Mean** with respect to a characteristic is given by

$$\mu = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{N}$$

3. **Population proportion.** When the characteristic being measured represents the presence or absence of some *dichotomous* attribute, we want to consider the proportion of units in the population having the attribute.

$$\pi = P_X = \frac{X_T}{N}$$

where

$$X_i = \begin{cases} 1 - \text{if attribute is present in element i} \\ 0 - \text{if attribute is not present in element i} \end{cases}$$

Population variance and standard deviation

. The *variance* and the *standard deviation* of the distribution of a characteristic in a population are of interest because they measure the spread of the distribution. The population variance of a characteristic is denoted by σ_X^2 and is given by

$$\sigma_X^2 = \frac{\sum_{i=1}^{N} (X_i - \mu_X)^2}{N}.$$

Standard deviation

The population standard deviation, denoted σ_X is simply the square root of the variance and is given by

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu_X)^2}{N}}.$$

If the characteristic of interest **Y** is a dichotomous random variable denoting the existence or the non-existence of some attribute, the variance is given by

$$\sigma_Y^2 = P_Y(1 - P_Y).$$

Population coefficient of variation

The *population coefficient of variation* is defined as the ratio of the standard deviation to the mean. The coefficient of variation represents the spread of the distribution relative to the mean of the distribution.

$$CV_X = \frac{\sigma X}{\mu X}$$

Illustrative Example from text.

Suppose we are interested in the distribution of household visits made by physicians in a community over a specified year. In this instance the elementary units are physicians and there are *25* of them. The *25* physicians in the community are labeled from 1 to 25 and the number of visits made by each physician is shown in the following table.

Table 2 Physicians visits in a year.

Physician	No. of Visits	Physician	No. of visits
1	5	14	4
2	0	15	8
3	1	16	0
4	4	17	7
5	7	18	0
6	0	19	37
7	12	20	0
8	0	21	8
9	0	22	0
10	22	23	0
11	0	24	1
12	5	25	0
13	6		

Population Characteristics - Parameters

In other words, $N = 25$. If we let $X =$ the number of physician visits made by physician i, the population mean, total, variance, and standard deviation are given as:

$$\mu_X = 5.08 \text{ (visits)} \quad \sigma_X^2 = 67.91 \text{ (visits)}^2$$
$$X_T = 127 \text{ (visits)} \quad \sigma_X = 8.24 \text{ (visits)}$$

If we let **Y** represent the attribute of having performed one or more household visits during the specified time period, we have

$$P_Y = \frac{14}{25} = 0.56$$

where P_Y is the proportion of physicians in the population who performed one or more household visits during the period. We also have

$$\sigma_Y^2 = (0.56) \times (1 - 0.56) = 0.246$$

$$\sigma_Y = \sqrt{0.246} = 0.496.$$

Coefficient of variation

For the distribution of household visits in the example:

$$CV_X = \frac{8.24}{5.08} = 1.62 \text{ and}$$
$$CV_Y = \frac{0.496}{0.56} = 0.886$$

The square of the coefficient of variation, V_X^2 is known as the *relative variance* or *rel-variance* and is a parameter that is widely used in sampling methodology.

Illustrative Example.

Are cholesterol levels in a population are more variable than systolic blood pressure variables in the same population?

- mean systolic blood pressure level in the population is 130 mmHg (millimeters of mercury) and the standard deviation is 15mmHg.
- mean cholesterol level is 200mg/100ml (milligrams per 100 milliliters) and that the standard deviation is 40 mg/100 ml.

Standard deviations do not tell us in any meaningful way which characteristic is more variable in the population because they are measured in different units (millimeters of mercury vs. milligrams per 100 milliliters in this instance).

Comparison of the two variables can be made, however, by examination of the respective coefficients of variation; 15/130 or .115, for systolic blood pressure vs. 40/200 or .200 for cholesterol. The coefficients of variation can be compared because they are *dimensionless* numbers. Thus, since the coefficient of variation for cholesterol level is greater than that for systolic blood pressure, we would conclude that cholesterol has more variability than systolic blood pressure in this population.

Illustrative Example.

Consider two variables that are measured in the same measurement units, for example, systolic blood pressure and diastolic blood pressure.

- mean diastolic blood pressure in the population is 60mmHg, with a standard deviation equal to 8 mmHg, and
- systolic blood pressure-has a mean and a standard deviation as given in the previous example (i.e., $\mu_X = 130$ mmHg and $\sigma_X = 15$mmHg).

Systolic blood pressure is more variable than diastolic blood pressure ($\sigma_X = 15$ mmHg vs. 8 mmHg). In relative terms,

- diastolic blood pressure has the greater variability, CV $= 8/60$ or 0.133 vs.
- systolic blood pressure CV $= 15/130$ or 0.115.

Relative variation is often of more concern in designed studies than absolute variation hence the importance of the coefficient of variation.

Sampling

Sample surveys can be categorized into two very broad classes on the basis of how the sample was selected, namely probability samples and non-probability samples. A *probability* sample has the characteristic that every element in the population has a known, nonzero probability of being included in the sample. In probability sampling, because every element has a known chance of being selected, *unbiased estimates of population parameters* that are linear functions of the observations (e.g., population means, totals, proportions) can be constructed from the sample data. Also, the standard errors of these estimates can be estimated under the condition that the second-order inclusion probabilities (i.e., joint probability of including any two enumeration units) are known.

Nonprobability Sampling

A *nonprobability* sample is one based on a sampling plan that does not have the features present in a probability sample.and the user has no firm method of evaluating either the reliability or the validity of the resulting estimates. These issues and concepts will be addressed later in Chapter 2. Nonprobability samples are used quite frequently, especially in market research and public opinion surveys. They are used because probability sampling is often a time-consuming and expensive procedure and, in fact, may not be feasible in many situations. An example of nonprobability sampling is the so-called quota survey in which interviewers are told to contact and interview a certain number of individuals from certain demographic subgroups.

Sampling Frame

In probability sampling, the probability of any element appearing in the sample must be known. So there must be a list of elements in the population available from which the sample can be selected. Such a list is called a sampling frame and should have the property that every element in the population has some chance of being selected in the sample by whatever method is used to select elements from the sampling frame. A sampling frame does not have to list all elements in the population.

Multistage sampling design

Often a particular sampling design specifies that the sampling be performed in two or three stages; this design is called a *multistage sampling* design. For example, a household survey conducted in a large state might have a sampling design specifying that a sample of counties be drawn within the state; that within each county selected in the sample, a sample of minor civil divisions (townships) be drawn; and that within each minor civil division a sample of households be drawn. In multistage sampling, a different sampling frame is used at each stage of the sampling. The units listed in the frame are generally called *sampling units*. The sampling units for the first stage are generally called *primary sampling units* (PSUs). The sampling units for the final stage of a multistage sampling design are called *enumeration units* or listing units.

Sample statistics
Total, mean and proportion

- **Sample Total:** The sample total is generally denoted by x_T and is the sum of the values over all elements in the sample:

$$x_T = \sum_{i=1}^{n} x_i$$

- **Sample Mean:** The sample mean generally denoted by $\bar{x}$ and is the sum of the values in the sample divided by the sample size:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Sample Proportion:** The sample proportion of a dichotomous characteristic is generally denoted by p_x is given by

$$p_x = \frac{x_t}{n}$$

Sample statistics

- **Sample Variance and standard deviation:** The sample variance s_x^2 is given by:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- **Dichotomous**

$$s_x^2 = \frac{np_x(1-p_x)}{n-1}$$

and for large sample sizes (> 20) an approximation can be used

$$s_x^2 = p_x(1-p_x)$$

- **Sample standard deviation**

$$s_x = \sqrt{p_x(1-p_x)}$$

Estimation of Population Characteristics

An estimate of the population total X_T can be obtained from the sample total x_T as given

$$x'_T = \left(\frac{N}{n}\right)(x)$$

An estimate σ_x^2 of the population variance σ_x^2 is given by

$$\hat{\sigma}_x^2 = \left(\frac{N-1}{N}\right)(s_x^2)$$

If the number of elements N in the population large, (N-1)/N:

$$\hat{\sigma}_x^2 \approx s_x^2$$

Table 2 Physicians visits in a year.

Physician	No. of Visits	Physician	No. of visits
1	5	17	7
6	0	19	37
7	12	21	8
12	5	25	0
13	6		

$x_T = 80$; $\bar{x} = 8.89$; $s_x^2 = 125.11$.

Population Parameter	Estimate from Sample
$X_T = 127$	$x'_T = \frac{25}{9}(80) = 222.22$
$\mu = \bar{X} = 5.08$	$\bar{x} = 8.89$
$\sigma_x^2 = 67.91$	$\hat{\sigma}^2 = \frac{24}{25}(125.22) = 120.11$
$P_Y = 0.56$	$p_y = \frac{7}{9} = 0.78$
$\sigma_y^2 = 0.246$	$\hat{\sigma}^2 = \frac{24}{25}(0.1944) = 0.1866$

Sampling Distributions-Mean of the Distribution

The *mean of the sampling distribution* of an estimated population parameter respect to a particular sampling plan that yields **T** possible samples resulting in **C** possible values of $\hat{d}$ is also known as the expected value of $\hat{d}$, denoted by $E(\hat{d})$, and is defined as

$$E(\hat{d}) = \sum_{i=1}^{C} \hat{d}_i \pi_i$$

where d_i, is a particular value of $\hat{d}$ and π_i, is the probability of obtaining that particular value of $\hat{d}$.

Sampling Distributions-Mean of the Distribution

(Note that if each sample is equally likely, then $\pi_i = \frac{f_i}{T}$ where f_i is the number of times that a particular value, $\hat{d}_i$, of $\hat{d}$ occurs).

Sampling Distributions-Variance of the Distribution

The variance $\text{Var}(\hat{d})$ of the sampling distribution of an estimated parameter $\hat{d}$ with respect to a particular sampling plan is given by

$$Var(\hat{d}) = \sum_{i=1}^{C} [\hat{d}_i - E(\hat{d})]^2 \pi_i$$

Sampling Distributions-Variance of the Distribution

The algebraic equivalent of this equation which can be used for computations is

$$Var(\hat{d}) = \sum_{i=1}^{C} \hat{d}_i^{\,2} \pi_i - [E(\hat{d})]^2$$

Sampling Distributions-Standard Error of the Distribution

The standard deviation $SE(\hat{d})$ of the sampling distribution of an estimated parameter $\hat{d}$ is more commonly known as the standard error of $\hat{d}$ and is simply the square root of the variance $Var(\hat{d})$ of the sampling distribution of $\hat{d}$:

$$SE(\hat{d}) = [Var(\hat{d})]^{1/2}$$

Sampling Distributions-Non Immunized for Measles

		Students Not Immunized for Measles	
School	No. of Students	Total	Proportion
1	59	4	.068
2	28	5	.179
3	90	3	.033
4	44	3	.068
5	36	7	.194
6	57	8	.140
Total	314	30	.096

Sampling Distributions Possible Samples

Sample(i)	Sample Schools	Total$_i$
1	1,2	27
2	1,3	21
3	1,4	21
4	1,5	33
5	1,6	36
6	2,3	24
7	2,4	24
8	2,5	36
9	2,6	39
10	3,4	18
11	3,5	30
12	3,6	33
13	4,5	30
14	4,6	33
15	5,6	45

Sampling Distributions for Sample Totals

Sample Total$_i$	Frequency	Relative frequency
18	1	1/15
21	2	2/15
24	2	2/15
27	1	1/15
30	2	2/15
33	3	3/15
36	2	2/15
39	1	1/15
45	1	1/15

Sampling Distributions-Non Immunized for Measles

$$
\begin{aligned}
E(\hat{d}) &= \sum^{alld} \hat{d}_i(f/15) \\
&= 18(1/15) + 21(2/15) + 24(2/15) + 27(1/15) \\
&+ 30(2/15) + 33(3/15) + 36(2/15) + 39(1/15) \\
&+ 45(1/15) \\
&= 30.
\end{aligned}
$$

Variance

The variance $Var(\hat{d})$ of the sampling distribution of $\hat{d}$ is

$$\begin{aligned} Var(\hat{d}) &= (18-30)^2(1/15) + (21-30)^2(2/15) + (24-30)^2(2/15) \\ &+ (27-30)^2(1/15) + (30-30)^2(2/15) + (33-30)^2(3/15) \\ &+ (36-30)^2(2/15) + (39-30)^2(1/15) + (45-30)^2(1/15) \\ &= 52.8 \end{aligned}$$

Sampling Distributions-Example with 1-10 Place Cards

Number Picked	Schools Chosen in Sample	Number Picked	Schools Chosen in Sample
1	1 and 2	6	2 and 3
2	1 and 3	7	2 and 4
3	1 and 4	8	2 and 5
4	1 and 5	9	2 and 6
5	1 and 6	10	3 and 4

Sampling Distributions-Example with 1-10 Place Cards, cont'd

Sample(i)	Sample Schools	Sample Total$_i$
1	1,2	27
2	1,3	21
3	1,4	21
4	1,5	33
5	1,6	36
6	2,3	24
7	2,4	24
8	2,5	36
9	2,6	39
10	3,4	18

Sampling Distributions-Example with 1-10 Place Cards, cont'd

Sample Total$_i$	π_i
18	1/10
21	2/10
24	2/10
27	1/10
33	1/10
36	2/10
39	1/10
Total	1

Sampling Distributions-Bias of the Estimate for a parameter of the Distribution

Mean Square Error

The *mean square error of population estimate* $\hat{d}$, denoted by $\text{MSE}(\hat{d})$, is defined as the mean of the squared differences over all possible samples between the values of the estimate and the true value **d** of the unknown parameter. In terms of the notation developed in the last section, the mean square error of **d** is defined by the relation

$$MSE(\hat{d}) = \sum_{i=1}^{C} \left(\hat{d}_i - d\right)^2 \pi_i$$

$$MSE(\hat{d}) = \sum_{i=1}^{C} \left(\hat{d}_i - d\right)^2 \pi_i$$

Notice the difference between the mean square error of an estimate and the variance of an estimate. The mean square error of an estimate is the mean value of squared deviations about the true value of the parameter being estimated; the variance of an estimate is the mean value of the squared deviations about the mean value of the sampling distribution of the estimate.

In general, the mean square error of the estimate is related to its bias and variance by the following relation:

$$MSE(\hat{d}) \quad = \quad Var\left(\hat{d}\right) + Bias^2\left(\hat{d}\right)$$

In other words, the mean square error of a population estimate is equal to the variance of the estimate plus the square of its bias.

Illustrative Example.

In the example involving the six schools, the first sampling plan discussed yielded an unbiased estimate of the population total. The mean square error of this estimate is given by

$$MSE(x_t') \quad = \quad 52.8 + 0^2 \quad = \quad 52.8$$

In other words, the MSE is equal to the variance of the estimate. In the example involving the same six schools and the sampling plan that yielded a biased estimate x_t' of X_T. the variance of x_t' is (from Equation (2.24))

$$\begin{aligned} Var(x_t') \quad = \quad & (18 - 27.9)^2(1/10) + (21 - 27.9)^2(2/10) \\ & +(24 - 27.9)^2(2/10) + (27 - 27.9)^2(1/10) \\ & +(33 - 27.9)^2(1/10) + (36 - 27.9)^2(2/10) \\ & +(39 - 27.9)^2(1/10) \quad = \quad 50.49 \end{aligned}$$

Thus, the MSE for x_t' is MSE(x_t') = 50.49 + (30 - 27.9)^2 = 54.9

Another illustrative example

Assessment of Burn victims

- Three students being trained to assess the severity of burn victims.
- Assess patients by viewing photographs of ten (10) burn victims each with "full-thickness" burns of 37%.

Recall

$$MSE(\hat{d}) = Var\left(\hat{d}\right) + Bias^2\left(\hat{d}\right)$$

Table: 2.9 Data for burn Area estimates

Student	$\hat{d}$=mean(%)	Variance$(\%)^2 = Var\left(\hat{d}\right)$
Dave	37	64
Don	42	9
Virginia	50	9

Table: 2.9 Bias and MSE for burn Area estimates

Student	Bias$=\left(\hat{d}-d\right)$	MSE$=\text{Var}\left(\hat{d}\right)+\text{Bias}^2\left(\hat{d}\right)$
Dave	37-37=0	$64+0^2=0$
Don	42-37=5	$9+5^2=34$
Virginia	50-37=13	$9+13^2=178$

Validity, Reliability, and Accuracy

Desirable Samples

- Sample designs should yield reliable and valid estimates. However, we have never defined just what the terms "reliable" and "valid" mean in terms of characteristics of estimates. We now have developed enough concepts and notation concerning estimates to be able to define these two terms as well as a third term, the "accuracy" of an estimate, which we will see is derived from the validity and reliability.

Reliability

The **reliability** of an estimated population characteristic refers to how **reproducible** the estimator is over repetitions of the process yielding the estimator. If we assume that there is no measurement error in the survey, then the reliability of an estimator can be stated in terms of its sampling variance or, equivalently, its standard error. **The smaller the standard error of an estimator, the greater is its reliability**.

Validity

The validity of an estimated population characteristic refers to how the mean of the estimator over repetitions of the process yielding the estimate, differs from the true value of the parameter being estimated. Again, if we assume that there is no measurement error, the validity of an estimator can be evaluated by examining the bias of the estimator. **The smaller the bias, the greater is the validity**.

Accuracy

The **accuracy**, of an estimator refers to how far away a particular value of the estimate is, on average, from the true value of the parameter being measured. The accuracy of an estimator is generally evaluated on the basis of its MSE or, equivalently, on the basis of the square root of its MSE (denoted by RMSE and called "root mean square error"). **The smaller the MSE of an estimate, the greater is its accuracy.**

Simple Random Sample(SRS)-Total ways

$$T = \frac{25!}{(5!)\,(25-5!)}$$

SRS-Probabilty not being included

$$\text{Probabilty not being included} = \frac{\binom{N-1}{n}}{\binom{N}{n}}$$

SRS-Probabilty being included

$$
\begin{aligned}
&= 1 - \left(\frac{N-n}{N}\right) \\
&= \left(\frac{n}{N}\right)
\end{aligned}
$$

Total Estimates

$$\begin{aligned}
\widehat{pt} &= \frac{N\sum_{i=1}^{n} x_i}{n} \\
\widehat{Var(\widehat{pt})} &= N^2\left(\frac{N-n}{N}\right)\left(\frac{s_x^2}{n}\right) \\
\widehat{SE(\widehat{pt})} &= N\sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right)
\end{aligned}$$

Mean Estimates

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
\widehat{Var(\overline{x})} &= \left(\frac{N-n}{N}\right)\left(\frac{s_x^2}{n}\right) \\
\widehat{SE(\overline{x})} &= \sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right)
\end{aligned}$$

Proportion Estimates

$$
\begin{aligned}
p_y &= \frac{\sum_{i=1}^{n} y_i}{n} \\
\widehat{Var\left(p_y\right)} &= \left(\frac{N-n}{N}\right)\frac{p_y\left(1-p_y\right)}{n-1} \\
\widehat{SE\left(p_y\right)} &= \sqrt{\left(\frac{N-n}{N}\right)}\sqrt{\frac{p_y\left(1-p_y\right)}{n-1}}
\end{aligned}
$$

Simple Random Sampling

Schools in Sample	Total_i	Schools in Sample	Total_i
1,2,3	24	2,3,4	22
1,2,4	24	2,3,5	30
1,2,5	32	2,3,6	32
1,2,6	34	2,4,5	30
1.3,4	20	2,4,6	32
1,3,5	28	2,5,6	40
1,3.6	30	3,4,5	26
1,4,5	28	3,4,6	28
1,4.6	30	3,5,6	36
1,5,6	38	4,5,6	36

Total_i	f_i	$\pi_i = \frac{f_i}{T}$
20	1	.05
22	1	.05
24	2	.10
26	1	.05
28	3	.15
30	4	.20
32	3	.15
34	1	.05
36	2	.10
38	1	.05
40	1	.05
Total	20	1.00

$$\begin{aligned} E(\widehat{pt}) &= 20(0.05) + 22(0.05) + 24(0.10) + 26(0.05) \\ &+ 28(0.15) + 30(0.20) + 32(0.15) + 34(0.05) \\ &+ 36(0.10) + 38(0.05) + 40(0.05) \\ &= 30. \end{aligned}$$

The variance $Var(\widehat{pt})$ of the sampling distribution of $\widehat{pt}$ is

$$\begin{aligned} Var(\widehat{pt}) &= (20-30)^2(0.05) + (22-30)^2(0.05) + (24-30)^2(0.10) + (26-3 \\ &+ (28-30)^2(0.15) + (30-30)^2(0.20) + (32-30)^2(0.15) + (34-3 \\ &+ (36-30)^2(0.10) + (38-30)^2(0.05) + (40-30)^2(0.05) \\ &= 26.4. \end{aligned}$$

Thus, we see that for simple random sampling, the estimated population total, $\widehat{pt}$, is an unbiased estimate of the population total pt. The standard error of $\widehat{pt}$ given by the equation above is directly proportional to σ_{pt}, the standard deviation of the distribution of PT the population total, in the population, and inversely proportional to the square root of the sample size, n.

The standard error also depends on the square root of the factor

$$\frac{(N-n)}{(N-1)},$$

which is known as the finite population correction, and is often denoted *fpc*.

We can obtain some insight into the role played by the *fpc* by examining its value for a hypothetical population containing N = 10,000 elements and for sample sizes as given in Table 3.3. From this table we see that if the sample size, n, is very much less than the population size, N, then the *fpc* is very close to unity and thus will have very little influence on the numerical value of the standard error SE($\widehat{pt}$) of the estimated total, $\widehat{pt}$. On the other hand, as n gets closer to N, the *fpc* decreases in magnitude and thus will cause a reduction in the value of SE($\widehat{pt}$)

fpc

can be re-written as

$$\begin{aligned}\sqrt{fpc} &= \sqrt{\frac{N-n}{N-1}} \\ &= \sqrt{\frac{N}{N-1}} \times \sqrt{1-\frac{n}{N}}\end{aligned}$$

so for increasing sample sizes...

Sample Size,n	fpc $= \sqrt{\frac{N-n}{N-1}}$
1	1.0000
10	.9995
100	.9950
500	.9747
1000	.9487
5000	.7071
9000	.3162

Total Estimates

$$\begin{aligned}
\widehat{pt} &= \frac{N\sum_{i=1}^{n} pt_i}{n} \\
Var(\widehat{pt}) &= N^2\left(\frac{N-n}{N-1}\right)\left(\frac{\sigma_{pt}^2}{n}\right) \\
SE(\widehat{pt}) &= N\sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma_x}{\sqrt{n}}\right)
\end{aligned}$$

Mean Estimates

$$\begin{aligned} \overline{x} &= \frac{\sum_{i=1}^{n} pt_i}{n} \\ Var(\overline{x}) &= \left(\frac{N-n}{N-1}\right)\left(\frac{\sigma_{pt}^2}{n}\right) \\ SE(\overline{x}) &= \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma_x}{\sqrt{n}}\right) \end{aligned}$$

Proportion Estimates

$$
\begin{aligned}
p_y &= \frac{\sum_{i=1}^{n} y_i}{n} \\
Var\left(p_y\right) &= \left(\frac{N-n}{N}\right)\frac{P_y\left(1-P_y\right)}{n-1} \\
SE\left(p_y\right) &= \sqrt{\left(\frac{N-n}{N-1}\right)}\sqrt{\frac{P_y\left(1-P_y\right)}{n}}
\end{aligned}
$$

Coefficients of Variation

We define the coefficient of variation $CV\left(\widehat{d}\right)$ of an estimate $\widehat{d}$ of a population parameter d as its standard error SE(d) divided by the true value d of the parameter being estimated.

$$CV\left(\widehat{d}\right) = \frac{Var\left(\widehat{d}\right)}{\widehat{d}}$$

The square of the coefficient of variation $CV^2\left(\widehat{d}\right)$ is a measure of the relative variation of a estimate. That is the variation relative to the estimate.

Define

$$CV(PT) = \frac{\sigma_{PT}}{\overline{X}}$$

then

$$CV(\widehat{pt}) = \left(\frac{CV(PT)}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}}$$

$$CV(\overline{x}) = \left(\frac{CV(PT)}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}}$$

$$CV(p_y) = \left(\frac{1-P_y}{\sqrt{nP_y}}\right)\sqrt{\frac{N-n}{N-1}}$$

reliability of Estimates

The standard error of an estimate is a measure of the sampling variability of the estimate over all possible samples. Under the assumption that measurement error is nonexistent or negligible, the reliability of an estimate can be judged by the size of the standard error; the larger the standard error, the lower is the reliability of the estimate (see Section 2.4). for reasonably large values of n (say, greater than 20), distributions that are close to the normal or Gaussian distribution, then we can use normal theory to obtain approximate confidence intervals for the unknown population parameters being estimated. For example, approximate 100(1 - α)% confidence intervals for the population total are given by

$$\widehat{pt} \quad \pm \quad z_{\alpha/2}(N)\sqrt{\frac{N-n}{N}}\left(\frac{s_{pt}}{\sqrt{n}}\right)$$

$$\overline{x} \quad \pm \quad z_{\alpha/2}\sqrt{\frac{N-n}{N}}\left(\frac{s_{pt}}{\sqrt{n}}\right)$$

Text Example

$$\begin{aligned} \text{sample total} &= 44 \\ s_{pt} &= 3.48 \\ \widehat{pt} &= \left(\frac{25}{9}\right)(44) \\ &= 122.22 \end{aligned}$$

so that the confidence interval is given by:

$$\begin{aligned} 122.22 &\pm 1.96(25)\sqrt{\frac{25-9}{25}}\left(\frac{3.48}{\sqrt{9}}\right) \\ &\rightarrow 122.22 \pm 45.47 \\ &\rightarrow (76.75, 167.69) \end{aligned}$$

Note that the true population total, $PT = 127$, is covered by this confidence interval. These 95% confidence intervals have the following usual interpretation: if we were to repeatedly sample n elements from this population according to the same sampling plan, and if, for each sample, confidence intervals were calculated, 95% of such confidence intervals would include the true unknown population parameter.

If the variable has a nearly symmetric distribution and the sample size is not small, then the confidence coefficients expressed in the confidence intervals will be approximately correct. If the data are badly skewed, however, and the sample size is small, the confidence coefficients may be misleading (Exercise 3.1 illustrates the situation using the data in Table 2.1).

Indicator functions

$$Y = \begin{cases} 1 \\ 0 \end{cases}$$

Family	Race	Out-of-Pocket Medical Expense (dollars)
1	W	500
2	B	350
3	B	430
4	W	280
5	W	170
6	B	50

Population characteristics for Z/Y

$$\frac{Z}{Y} = \frac{\$830}{3}$$

Family	Race	Out-of-Pocket Medical Expense (dollars X_i)	Y_i	Z_i
1	W	500	0	0
2	B	350	1	350
3	B	430	1	430
4	W	280	0	0
5	W	170	0	0
6	B	50	1	50

$$Y_i = \begin{cases} 1 & \text{African american family} \\ 0 & \text{Caucasian family} \end{cases}$$

Given the sampling distibution

Sample Elements	z	y	z/y^*
1,2,3,4	780	2	390
1,2,3,5	780	2	390
1,2,3,6	830	3	276.67
1,2,4,5	350	1	350
1,2,4,6	400	2	200
1,2,5,6	400	2	200
1,3,4,5	430	1	430
1,3,4,6	480	2	240
1,3.5,6	480	2	240
1.4,5,6	50	1	50
2,3,4,5	780	2	390
2,3,4,6	830	3	276.67
2,3,5,6	830	3	276.67
2,4,5,6	400	2	200
3,4,5,6	480	2	240

Using complementary indicator function

Family	Race	Out-of-Pocket Medical Expense (dollars X_i)	Y_i	Z_i
1	W	500	1	500
2	B	350	0	0
3	B	430	0	0
4	W	280	1	280
5	W	170	1	170
6	B	50	0	0

$$Y_i = \begin{cases} 1 & \text{Caucasian family} \\ 0 & \text{African american family} \end{cases}$$

$$SE\left(\frac{z}{y}\right) = \left[\frac{\sigma_z}{\sqrt{E(y)}}\right] \times \sqrt{\frac{Y - E(y)}{Y - 1}}$$

where

$$\sigma_z = \sqrt{\left[\frac{\sum_{i=1}^{Y} (Z_i - \overline{Z})^2}{Y}\right]}$$

Standard Error Estimate

$$\widehat{SE\left(\frac{z}{y}\right)} = \left[\frac{\widehat{\sigma_z}}{\sqrt{pt_y}}\right] \times \sqrt{\frac{\widehat{pt_y} - pt_y}{\widehat{pt_y} - 1}}$$

Text Example For example, suppose that from a hospital admitting 20,000 patients annually, a survey of hospital patients is to be taken for the purpose of determining the proportion of the 20,000 patients that received optimal care as defined by specified standards. The quality care review committee planning the survey may feel that some remedial action should be taken if fewer than 80% of the patients are receiving optimal care. In this instance, the committee would be concerned about overestimates of the true proportion, but would probably not be too concerned if the estimated proportion were 80% when the true propor- tion were 75%. The statistician might formulate this by saying that the user would like to be "virtually certain" that the estimated proportion differs from the true proportion by no more than l00[80-75)/75]% or 6.67% of the true proportion.

$$
\begin{aligned}
3 \times SE\left(p_y\right) &= 3 \times \sqrt{\frac{P_y\left(1-P_y\right)}{n}} \sqrt{\frac{N-n}{N-1}} \\
&\rightarrow 3 \times SE\left(p_y\right) \leq 0.0667 P_y \\
&\rightarrow 3 \times \sqrt{\frac{P_y\left(1-P_y\right)}{n}} \sqrt{\frac{N-n}{N-1}} \leq 0.0667 P_y
\end{aligned}
$$

or

$$
n \geq \frac{9 N P_y\left(1-P_y\right)}{(N-1)(0.667)^2 P_y^2+9 P_y\left(1-P_y\right)}
$$

So setting P_y=0.80 and N=20,000

$$\begin{aligned} n &\geq \frac{9(20,000)(0.80)\,(0.20)}{(19,999)\,(0.0667)^2(0.80)^2 + 9(0.80)\,(0.20)} \\ n &\geq 493.295 \text{or } 494. \end{aligned}$$

See Box 3.5 for the Exact and approximate Sample size Required under simple random sampling

Text Illustrative Example

A sample survey of retail pharmacies is to be conducted in a state that contains 2500 pharmacies. The purpose of the survey is to estimate the average retail price of 20 tablets of a commonly used vasodilator drug. An estimate is needed that is within 10 % of the true value of the average retail price in the state. A list of all pharmacies is available and a simple random sample is to be taken from the list. A phone survey of 20 of N = 1000 pharmacies in another state showed an average price of \$7.00 for 20 tablets with a standard deviation of \$1.40.

$$
\begin{aligned}
CV(\overline{x})^2 &= \frac{[(N-1)/N]s_x^2}{\overline{x}^2} \\
&= \frac{[999/1000](1.4)^2}{(7.00)^2} \\
&= 0.04
\end{aligned}
$$

with a tiny $\epsilon = 0.1$ and N=2500. Using the exact formula from Box 3.5

$$\begin{aligned} n &= \frac{9(2500)(0.04)}{9(0.04) + 2499(0.1)^2} \\ &= 35.6 \\ &\approx 36. \end{aligned}$$