Survey Sampling: Introduction

Three Elements of Statistical Study:

- Collecting Data: Sample Surveys and Sampling
- Describing and Presenting Data: Graphical and Numerical descriptions
- Drawing Conclusions from Data: Inference, estimation and test of hypothesis.

Survey sampling: want to use sample information to make inference of the finite population.

—- The rest of statistics, $Y_1, Y_2, \cdots Y_n$ have a joint prob distribution, say $N(u, \sigma^2)$. $Y_1, Y_2, \cdots Y_n$ are random variables. Observed values of random variables are $y_1, y_2, \cdots, y_n$.

—- General probability sampling: $Y_1, Y_2, \cdots, Y_N$ is the population. We sample n of the N units, say $y_1, y_2, \cdots, y_n$ according to a pre-specified design in which we assign a probability of selection to each possible subset of the population of size n. Neither $Y_1, Y_2, \cdots Y_N$ nor $y_1, y_2, \cdots, y_n$ are random

variables. Random variables are Z_i's with

$$Z_i = \begin{cases} 1 & \text{if unit} \quad i \in S \\ 0 & \text{otherwise} \end{cases}$$

Some definitions:

- Observation Units: An object on which a measurement is taken. Sometimes called an element
- 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 votes? All persons who voted in the last election? The choice of target population will profoundly affect the statistics that result

- Sample: A subset of a population
- Sampled population: The population from which the sample was taken
- Sampling unit: The unit we actually sample

  Example: We 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

  Example: 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

- Census: When data is collected on every unit of the population, it is called a census.
- Population parameter: A number that results from measuring all the units in the population
- Statistic: A number that results from measuring all the units in the sample
- statistics derived from samples are used to estimate population parameters.

Why Sampling?

- Cost: Census is expensive
- Time: Census is very time consuming.
- Impractical: In some applications census can be impractical. Example: The government requires automakers who want to sell cars in the U.S. to demonstrate that their cars can survive certain crash tests. Obviously, the company can't be expected to crash every car to see if it survives! So the company crashes only a sample of cars.

Types of Samples:

1. Non-probability (non-random) samples: These samples focus on volunteers, easily available units, or those that just happen to be present when the research is done. Non-probability samples are useful for quick and cheap studies, for case studies, for qualitative research, for pilot studies, and for developing hypotheses for future research.

- Convenience sample: also called an ”accidental” sample or “man-in-the-street” samples. The researcher selects units that are convenient, close at hand, easy to reach, etc.

- Purposive sample: the researcher selects the units with some

purpose in mind, for example, students who live in dorms on campus, or females.

- Quota sample: the researcher constructs quotas for different types of units. For example, to interview a fixed number of shoppers at a mall, half of whom are male and half of whom are female.

2. Probability-based (random) samples: These samples are based on probability theory. Every unit of the population of interest must be identified, and all units must have a known, non-zero chance of being selected into the sample.

- Simple random sample (SRS): Randomly select a size n sample from a size N population. Each unit in the population is identified.

  —-a) the sampling unit and observation unit coincide;

  — b) Each subset of size n has same probability of being the sample;

  —-c) Each unit has an equal chance of being selected in the

sample;

—-Random number generators

—-Lottery method

- Systematic random sampling: First randomly picks the first item or subject from the population. Then, select each n'th subject from the list.

  ——The results are representative of the population unless certain characteristics of the population are repeated for every n'th individual which is highly unlikely.

  ——Systematic sampling is useful for selecting large samples, say 100 or more. It is less cumbersome than a simple ran-

dom sample using either a table of random numbers or lottery method

——If the selection interval matches some pattern in the list, for example, the list is male, female, male, female, $\cdots$, and you select No.1, No.3, No.5 $\cdots$ observations to form a systematic sample, you will introduce systematic bias into your sample

- Stratified random sampling: Divide population into H strata, take an SRS of size n_h from stratum h, $h = 1, \cdots, H$, select the sample independently.

  Example 1:


SASK.
MANITOBA
NFLD.
BRITISH COLUMBIA
ALTA.
C A N A D A
QUÉBEC
ONTARIO
WASH.
MONT.
N. DAK.
Lake Superior
Montréal
MAINE
OREGON
MINN.
Ottawa
VT.
N.H.
IDAHO
WYOMING
S. DAK.
WIS.
Toronto
N.Y.
MA.
R.I.
CT.
MICHIGAN
Chicago
New Castle
PA.
New York
Pacific
Ocean
CALIF.
NEVADA
NEBR.
IOWA
U N I T E D S T A T E S
OHIO
N.J.
MD.
DE
Washington, D.C.
UTAH
COLO.
ILL.
IND.
WV
VA.
KY.
KANSAS
MO.
Atlantic Ocean
Los Angeles
San Diego
ARIZONA
N. MEX.
TENN.
N.C.
OKLA.
ARK.
S.C.
Phoenix
MISS.
Dallas
ALA.
GA.
TEXAS
LA.
FLA.
San Antonio
Houston
Gulf of Mexico
Nassau
MEXICO
Monterrey
Havana
Santo Domingo
© 2001 Microsoft Corp. All rights reserved.

Example 2: You want to find out the attitudes of students on your campus about immigration.

—— 27,000 students: 22,000- West; 3,000 - East; 1000 - Mid-west; 600 - South; - 400 Foreign.

——Select a simple random sample of 1500 students, you might not get any from the Midwest, South, or Foreign.

——Divide the students into these five groups (Stratum), and then select the same percentage of students from each group using a simple random sampling method. This is proportional stratified random sampling.

——Divide students into the five groups and then select the

same number of students from each group using a simple random sampling method. This is disproportionate stratified random sampling.

- Cluster sampling: A cluster is a naturally-occurring grouping of the members of the population. For example, city residents are also residents of neighborhoods, blocks, and housing structures. Randomly select n clusters, then observe all the elements in the selected clusters or partial of the elements in the selected clusters.

  Example: To obtain information about the drug habits of all high school students in New Mexico.

—-obtain a list of all the school districts in NM

—-Select an SRS of school districts

—-Within each selected school district, list all the high schools and select an SRS of high schools

—-Within each selected high school, list all high school classes, and select an SRS of classes

—-The students in the selected classes are the observations in your sample

Biases

- Selection Bias: If some part of the target population is not in the sampled population, a bias called Selection Bias occurs.

  —Example: In a survey to estimate per capita income, if transient people are ignored.

  —Mis-specification of the target population

  —-Failure to include all the target population in the sampling frame, also called undercoverage

  —-Substituting a convenient member of a population for a designated member not readily available

  —-Non Response: Failure to obtain responses from all those

chosen in the sample (distorts the results of a survey, typically non-respondents differ from the respondents in some pronounced way)

—-Allowing a sample to consist entirely of volunteers (Radio, TV, or call-in polls) Note that large samples are generally considered good but if the sample is unrepresentative, it can be quite bad. The design of the survey is far more important than the absolute size of the sample.

- Measurement Bias: Measurement bias occurs when the measuring instrument has a tendency to reord in one direction more often than the other. Measurement biases are more common when dealing with people.

  —-People may not tell the truth

  —-Lack of understanding of questions

  —-Lack of proper account of events in memory

  —-Variations in responses due to interviewer

  —-Misreading questions, or miss recording responses

  —-Desire to impress the interviewer

  —-Ordering and wording of questions have effects on responses

Many of these problems can be avoided by proper questionnaire design

Questionnaire Design:

- Decide what you want to find out; this is the most important step in writing a questionnaire
- Pilot study: Test questions before sending out the questionnaire.
- Keep the questions Simple and Clear: Questions should be neither too lengthy nor too technical. They should be easily understood by non experts
- Questions should be specific and not general
- Decide whether to use open or closed questions.

—-Open Question: The respondent is not prompted with categories for responses. It allows responses to form their own response categories.

——Closed Question: A question is closed when specific response categories are provided.

—— Closed questions with well thought and researched categories elicit more accurate responses.

- Avoid questions that prompt or motivate the respondent to say what investigator wants to hear
- Use choices rather that Agree/Disagree type questions
- Ask only one concept in one question

- Pay attention to question-order effect. Ask general questions first then follow with specific questions.

Sampling and Non-Sampling Errors:

- Sampling Errors: Sampling error are results of inherent variability in the sampling process. These arise because the results vary from sample to sample. Margin of errors reported are a result of sampling error. These can only be reduced by increasing the sample size but not be eliminated.
- Non-Sampling Errors: These result from selection bias, measurement error and inaccuracies of responses. These can not be attributed to sample-to-sample variability. Such errors can be eliminated by proper precautions. Selection bias can be reduced by using probability sample. Accurate responses can

be achieved through proper and careful design of survey instrument and training of interviewers.

Probability Sampling: Inference about a population depends on probabilities with which units are included in a sample.

Simple Random Sample (SRS): Randomly select a size n sample from size N population

- Every possible subset of the population of size n is equally likely to be the sample.
- Each individual is equally likely to be in the sample.

Stratified Random Sample:

- First, partition the population into subgroups, called strata.
- Then, independently select an SRS from each stratum.

Cluster Sample:

- Some or most sampling units contain more than one observation unit. These observation units are called clusters. A cluster is a naturally-occurring grouping of the members of the population
- Take an SRS of clusters.
- Sample all observation units or partial of the observations in the sampled cluster.

Framework for Probability Sampling:

- $U = \{1; 2; \cdots ; N\}$ universe finite population
- S sample
- N population size
- n sample size.
- $\pi_i = p(\text{unit} \quad i \quad \text{in the sample})$
- t population total, $t = \sum_{i=1}^{N} y_i$
- $\bar{y}_U$ population mean, $\bar{y}_U = t/N$
- $\bar{y}_S$ sample mean, $\bar{y}_S = \sum_{i=1}^{n} y_i/n$

- $\hat{t} = N\bar{y}_S$
- $E[\hat{t}] = \sum_S \hat{t}_S p(S)$
- Bias $[\hat{t}] = E[\hat{t}] - t$
-

$$\begin{aligned} V(\hat{t}) &= E[(\hat{t} - E[\hat{t}])^2] \\ &= \sum_S p(S)(\hat{t}_S - E[\hat{t}])^2 \end{aligned}$$

- $$\begin{aligned} MSE(\hat{t}) &= E[(\hat{t}-t)^2] \\ &= E[\hat{t}-E(\hat{t})+E(\hat{t})-t]^2 \\ &= E[(\hat{t}-E(\hat{t}))^2]+(E(\hat{t})-t)^2 \\ &= V(\hat{t})+(bias[\hat{t}])^2 \end{aligned}$$

- S^2 population variance, $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y}_U)^2$

- s^2 sample variance, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y}_S)^2$

- unbiased, $E[\hat{t}] = t$

- precise, $V[\hat{t}] = E[(\hat{t} - E(\hat{t}))^2]$ is small, measures how close estimates from different samples are to each other
- accurate, $MSE[\hat{t}] = E[(\hat{t} - t)^2]$ is small, measures how close the estimate is to the true value

Example: $U = \{1, 2, 3\}, N = 3, n = 2$

i	1	2	3
y_i	1	2	4

Possible S	$p(S)$	sample values	$\bar{y}_S$	$\hat{t}_S$
$S_1 = \{1, 2\}$	1/3	1,2	$\bar{y}_{S_1} = 1.5$	$\hat{t}_1 = 9/2$
$S_2 = \{1, 3\}$	1/3	1,4	$\bar{y}_{S_2} = 5/2$	$\hat{t}_2 = 15/2$
$S_1 = \{2, 3\}$	1/3	2,4	$\bar{y}_{S_3} = 3$	$\hat{t}_3 = 9$

$t = \sum_{i=1}^{3} y_i = 1 + 2 + 4 = 7$

$$
\begin{aligned}
E[\hat{t}] &= \sum_S \hat{t}_S p(S) \\
&= \hat{t}_1 p(S_1) + \hat{t}_2 p(S_2) + \hat{t}_3 p(S_3) \\
&= \frac{9}{2} \times \frac{1}{3} + \frac{15}{2} \times \frac{1}{3} + 9 \times \frac{1}{3} \\
&= 7
\end{aligned}
$$

$\hat{t}$ is unbiased for t

$$
\begin{aligned}
V(\hat{t}) &= \sum_S p(S)(\hat{t}_S - E[\hat{t}])^2 \\
&= p(S_1)(\hat{t}_{S_1} - E[\hat{t}])^2 + p(S_2)(\hat{t}_{S_2} - E[\hat{t}])^2 \\
&+ p(S_3)(\hat{t}_{S_3} - E[\hat{t}])^2 \\
&= \frac{7}{2}
\end{aligned}
$$

$$
MSE(\hat{t}) = V(\hat{t}) + bias^2 = \frac{7}{2}
$$

Randomization Theory Results for Simple Random Sampling

Assumptions:

- The approach is nonparametric; there are no assumptions on the distribution of the y_i.
- y_i are a collection of unknown numbers.

  Want to show $V(\bar{y}_S) = (1 - \frac{n}{N})\frac{S^2}{n}$, which is estimated by $\hat{V}(\bar{y}_S) = (1 - \frac{n}{N})\frac{s^2}{n}$, where $(1 - \frac{n}{N})$ is called finite population correction (fpc). If $n = N$, variance is 0. No variability due to taking a sample, i.e, no sampling variability. If $\frac{n}{N}$ is large, variance is small.

Proof:

$$Z_i = \begin{cases} 1 & \text{if unit} \quad i \in S \\ 0 & \text{otherwise} \end{cases}$$

$$\bar{y}_S = \sum_{i \in S} \frac{y_i}{n} = \sum_{i=1}^{N} Z_i \frac{y_i}{n}$$

$$p(Z_i = 1) = p(\text{unit} \quad i \in S) = \frac{n}{N}$$

$$E(Z_i) = E(Z_i^2) = \frac{n}{N}$$

$$\begin{aligned} V(Z_i) &= E(Z_i^2) - (E[Z_i])^2 \\ &= \frac{n}{N} - (\frac{n}{N})^2 \\ &= \frac{n}{N}(1 - \frac{n}{N}) \end{aligned}$$

For $i \neq j$,

$$\begin{aligned} E[Z_i Z_j] &= P(Z_i = 1 \text{and} Z_j = 1) \\ &= P(Z_j = 1 | Z_i = 1) P(Z_i = 1) \\ &= (\frac{n-1}{N-1})(\frac{n}{N}) \end{aligned}$$

$$
\begin{aligned}
Cov(Z_i, Z_j) &= E[Z_i Z_j] - E[Z_i]E[Z_j] \\
&= (\frac{n-1}{N-1})(\frac{n}{N}) - (\frac{n}{N})^2 \\
&= -\frac{1}{N-1}(1 - \frac{n}{N})(\frac{n}{N})
\end{aligned}
$$

$$
\begin{aligned}
E(\bar{y}) &= E[\sum_{i=1}^{N} Z_i \frac{y_i}{n}] \\
&= \sum_{i=1}^{N} \frac{n}{N} \frac{y_i}{n} \\
&= \sum_{i=1}^{N} \frac{y_i}{N} \\
&= \bar{y}_U
\end{aligned}
$$

$$
\begin{aligned}
V(\bar{y}) &= \frac{1}{n^2} V(\sum_{i=1}^{N} Z_i y_i) \\
&= \frac{1}{n^2} Cov(\sum_{i=1}^{N} Z_i y_i, \sum_{j=1}^{N} Z_j y_j) \\
&= \frac{1}{n^2} [\sum_{i=1}^{N} y_i^2 V(Z_i) + \sum_{i=1}^{N} \sum_{j \neq i}^{N} y_i y_j Cov(Z_i, Z_j)] \\
&= \frac{1}{n^2} [\frac{n}{N}(1 - \frac{n}{N}) \sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N} \sum_{j \neq i}^{N} y_i y_j \frac{1}{N-1} (1 - \frac{n}{N})(\frac{n}{N})] \\
&= \frac{1}{n^2} \frac{n}{N}(1 - \frac{n}{N})[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1} \sum_{i=1}^{N} \sum_{j \neq i}^{N} y_i y_j] \\
&= \frac{1}{n}(1 - \frac{n}{N}) \frac{1}{N(N-1)} [(N-1) \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2 + \sum_{i=1}^{N} y_i^2] \\
&= \frac{1}{n}(1 - \frac{n}{N}) \frac{1}{N(N-1)} [N \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2] \\
&= (1 - \frac{n}{N}) \frac{S^2}{n}
\end{aligned}
$$

Also can show that $\hat{V}(\bar{y}_S) = (1 - \frac{n}{N})\frac{s^2}{n}$ is an unbiased estimator of $V(\bar{y}_S) = (1 - \frac{n}{N})\frac{S^2}{n}$.

Assumptions for confidence intervals

- Only have a finite population
- Want to use asymptotic results in finite population sampling
- Pretend that our population itself part of a larger superpopulation, the superpopulation is itself a subset of a larger superpopulation, and so on, until the superpopulations are as large as we could wish. Our population is embedded in a series of increasing finite populations. This embedding can give us properties such as consistency and asymptotic normality.

- Hajek (1960), CLT for SRS without replacement. It says if certain conditions hold and if n, N, and $N-n$ are all "sufficiently large," then the sampling distribution of

$$\frac{\bar{y}-\bar{y}_U}{\sqrt{(1-\frac{n}{N})}\frac{S}{\sqrt{n}}}$$

is approximately normal with mean 0 and variance 1. A large-sample $100(1-\alpha)\%$ CI for the population mean is

$$[\bar{y}-z_{\alpha/2}\sqrt{(1-\frac{n}{N})}\frac{S}{\sqrt{n}},\bar{y}+z_{\alpha/2}\sqrt{(1-\frac{n}{N})}\frac{S}{\sqrt{n}}],$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$th percentile of the standard normal distribution.

Simple Random Sampling:

- $\hat{V}(\bar{y}_S) = (1 - \frac{n}{N})\frac{s^2}{n}$
- CI $[\bar{y} - z_{\alpha/2}\sqrt{(1 - \frac{n}{N})}\frac{S}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\sqrt{(1 - \frac{n}{N})}\frac{S}{\sqrt{n}}]$

Stratified Random Sampling:

strata	1	2	$\cdots$	H	
popn size	N_1	N_2	$\cdots$	N_H	$\sum_{i=1}^{H} N_i = N$
sample size	n_1	n_2	$\cdots$	n_H	$\sum_{h=1}^{H} n_h = n$
popn total	t_1	t_2	$\cdots$	t_H	

- Independently take a SRS of size n_h from stratum H
- $t_{str} = t_1 + t_2 + \cdots + t_H$

- $$\begin{aligned} \hat{t}_{str} &= \hat{t}_1 + \hat{t}_2 + \cdots + \hat{t}_H \\ &= N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots N_H\bar{y}_H \end{aligned}$$

- $$\begin{aligned} \hat{V}(\hat{t}_{str}) &= \hat{V}(\hat{t}_1) + \hat{V}(\hat{t}_2) + \cdots + \hat{V}(\hat{t}_H) \\ &= \sum_{h=1}^{H}(1 - \frac{n_h}{N_h})\frac{N_h^2 s_h^2}{n_h} \end{aligned}$$

- $$\begin{aligned}\bar{y}_{str} &= \frac{\hat{t}_{str}}{N} \\ &= \frac{\sum_{h=1}^{H} \hat{t}_h}{N} \\ &= \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{N} \\ &= \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h\end{aligned}$$

- CI $\bar{y}_{str} \pm z_{\alpha/2}\sqrt{\hat{V}(\bar{y}_{str})}$