Outline

- Sampling process and sampling design
- Sampling Formulas

Part I

Sampling process and sampling design

Introduction

For small population, a census study is used because data can be gathered on all.

For large population, sample surveys will be used to obtain information from a small (but carefully chosen) sample of the pop'n. The sample should reflect the characteristics of the population from which it is drawn.

Sampling…….

- 3 factors that influence sample representative-ness
  - Sampling procedure
  - Sample size
  - Participation (response)

- When might you collect the entire population?
  - When your population is very small
  - When you have extensive resources
  - When you don't expect a very high response

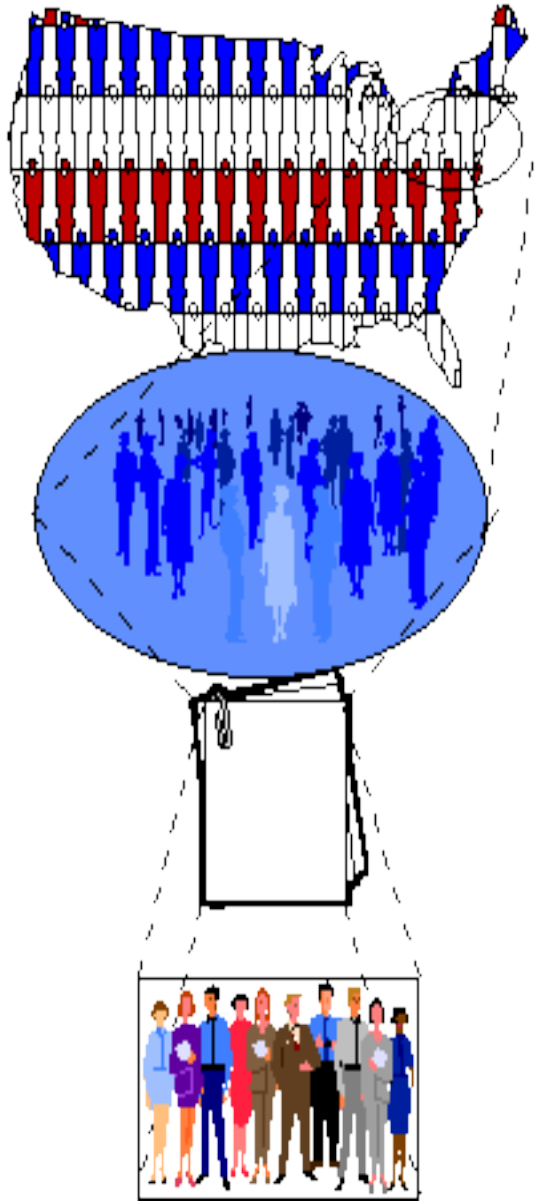
4

Sampling Process

- The sampling process comprises several stages:
  - Defining the population of concern
  - Specifying a **sampling frame**, a set of items or events possible to measure
  - Specifying a **sampling method** for selecting items or events from the frame
  - Determining the sample size (not presented)
  - Implementing the sampling plan "
  - Sampling and data collecting "
  - Reviewing the sampling process "

Who do you want to generalize to?

The Theoretical Population



What population can you get access to?

The Study Population

How can you get access to them?

The Sampling Frame

Who is in your study?

The Sample

Sampling Frame -1-

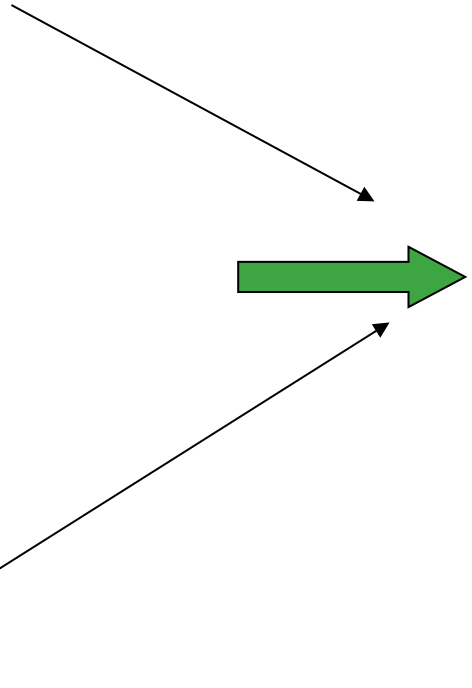
- In the most straightforward case, such as the sentencing of a batch of material from production (acceptance sampling by lots), it is possible to identify and measure every single item in the population and to include any one of them in our sample. However, in the more general case this is not possible. There is no way to identify all rats in the set of all rats. Where voting is not compulsory, there is no way to identify which people will actually vote at a forthcoming election (in advance of the election)
- As a remedy, we seek a *sampling frame* which has the property that we can identify every single element and include any in our sample .
- The sampling frame must be representative of the population

Sampling Frames -2-

- Sources
  - Administrative records-eg
    - Hospital records
    - Birth and Death Registers
    - LC lists
    - Voters' register
    - School registers
    - etc
  - Construct your own

Types of Samples and Sample design

- Probability (Random) Samples
- **Simple random sample (SRS)**
  - Systematic random sample
  - Stratified random sample
  - Multistage sample (not presented)
  - Multiphase sample "
  - Cluster sample "
- Non-Probability Samples
  - Convenience sample
  - Purposive sample
  - Quota



Probability Sampling

- A **probability sampling** scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.

- When every element in the population *does* have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

Non-Probability Sampling

- Any sampling method where some elements of population have *no* chance of selection (these are sometimes referred to as 'out of coverage'/'undercovered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, nonprobability sampling not allows the estimation of sampling errors..

- *Example: We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a nonprobability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities*.

Probability Sampling Design

- Probability sampling includes:
  - Simple Random Sampling (**srs**),
  - Systematic Sampling,
  - Stratified Random Sampling (SRS),
  - Cluster Sampling (not presented)
  - Multistage Sampling (not presented)
  - Multiphase sampling (not presented)

Simple Random Sampling (srs)

- Applicable when population is small, homogeneous & readily available
- All subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection.
- It provides for greatest number of possible samples. This is done by assigning a number to each unit in the sampling frame.
- A table of random number or lottery system is used to determine which units are to be selected.

Simple Random Sampling……..

- Advantages
- Estimates are easy to calculate.
- Simple random sampling is always an EPS design, but not all EPS designs are simple random sampling.

- Disadvantages
- If sampling frame large, this method impracticable.
- Minority subgroups of interest in population may not be present in sample in sufficient numbers for study.

srs: Replacement of Selected Units

- Sampling schemes may be ***without replacement*** ('WOR' - no element can be selected more than once in the same sample) or ***with replacement*** ('WR' - an element may appear multiple times in the one sample).

- For **example**, if we catch fish, measure them, and immediately return them to the water before continuing with the sample, this is a WR design, because we might end up catching and measuring the same fish more than once. However, if we do not return the fish to the water (e.g. if we eat the fish), this becomes a WOR design.

Systematic Sampling

- **Systematic sampling** relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list.

- Systematic sampling involves a random start and then proceeds with the selection of every *k*th element from then onwards. In this case, *k*=(population size/sample size).

- It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the *k*th element in the list.

- A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10').

Systematic Sampling……

As described above, systematic sampling is an EPS method, because all elements have the same probability of selection (in the example given, one in ten). It is *not* 'simple random sampling' because different subsets of the same size have different selection probabilities - e.g. the set {4,14,24,...,994} has a one-in-ten probability of selection, but the set {4,13,24,34,...} has zero probability of selection.

Systematic Sampling……

- **ADVANTAGES:**
- Sample easy to select
- Suitable sampling frame can be identified easily
- Sample evenly spread over entire reference population

- **DISADVANTAGES:**
- Sample may be biased if hidden periodicity in population coincides with that of selection.
- Difficult to assess precision of estimate from one survey.

Stratified Sampling

Where population embraces a number of distinct categories, the frame can be organized into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.

- Every unit in a stratum has same chance of being selected.
- Using same sampling fraction for all strata ensures proportionate representation in the sample.
- Adequate representation of minority subgroups of interest can be ensured by stratification & varying sampling fraction between strata as required.

Stratified Sampling……

- Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata.

- **Drawbacks** to using stratified sampling

- First, sampling frame of entire population has to be prepared separately for each stratum
- Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata.
- Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods

Stratified Sampling…….

Draw a sample from each stratum

Post Stratification

- Stratification is sometimes introduced after the sampling phase in a process called "poststratification“.

- This approach is typically implemented due to a lack of prior knowledge of an appropriate stratifying variable or when the experimenter lacks the necessary information to create a stratifying variable during the sampling phase. Although the method is susceptible to the pitfalls of post hoc approaches, it can provide several benefits in the right situation.

- Implementation usually follows a simple random sample. In addition to allowing for stratification on an ancillary variable, poststratification can be used to implement weighting, which can improve the precision of a sample's estimates.

Non-Probability Sampling…….

- Nonprobability Sampling includes:
  Accidental Sampling, Convenience Sampling,
  Quota Sampling and
  Purposive (or Judgement) Sampling.
- In addition, nonresponse effects may turn *any* probability design into a nonprobability design if the characteristics of nonresponse are not well understood, since nonresponse effectively modifies each element's probability of being sampled.

Part II

Sampling Formulas and estimation:

Sampling Random Sampling and Stratified Random Sampling

Many Sampling methods….

Probability Sampling:

- Random Sampling
- Systematic Sampling
- Stratified Sampling

NonProbability Sampling

- Convenience Sampling
- Judgment Sampling
- Quota Sampling
- Snowball Sampling

…….and many others……

The winner is: Probability Sampling.

In non-probability sampling, members are selected from the pop'n in some non-random manner and the sampling error (=degree to which a sample might differ from the pop'n) is unknown.

In probability sampling, each member of the pop'n has a specified probability of being included in the sample. Its advantage is that sampling error can be calculated.

Population parameters

Definition: Parameters are those numerical characteristics of the pop'n that we will estimate from a sample.

Notations:

$$\underline{Pop'n\ element}: \quad 1 \quad 2 \quad \ldots \quad N$$

$$\underline{Sample}: s\ is\ a\ subset\ of \{\ 1, 2, \ldots, N\ \}$$

$$\underline{Variable\ of\ Interest\ X}: x_1\ x_2\ \ldots\ x_N$$

$$\underline{Example}: x_i = age\ or\ weight\ or = \begin{cases} 1\ presence \\ 0\ absence \end{cases}$$

Population mean, total, variance

Pop'n mean: $$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Pop'n total: $$\tau = \sum_{i=1}^{N} x_i = N\mu$$

Pop'n variance: its square root is the StdDev

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

Simple Random Sampling (s.r.s)

The most elementary form of sampling is **s.r.s.** where each member of the pop'n has an equal and known chance of being selected at most once.

There are $\binom{N}{n}$ possible samples of size n taken without replacement.

In this section, some statistical properties of the sample mean will be derived.

The Sample Mean

The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

estimates the pop'n mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

where

$\underline{Pop'n\ element}$: $1 \quad 2 \quad \ldots \quad N$

$\underline{Sample\ element}$: $1 \quad 2 \quad \ldots \quad n \quad (n \le N)$

$\underline{Values\ of\ the\ sample\ members}$: $X_1 \; X_2 \; \ldots\ldots \; X_n$

$\underline{\underline{Huge\ Difference}}$: $\begin{cases} X_i = Random\ Value\ (sample) \\ x_i = Fixed\ Value\ (pop'n) \end{cases}$

so that $T = N\,\bar{X}$ will estimate the pop'n total $\tau = N\mu$

Distribution of Sample Means

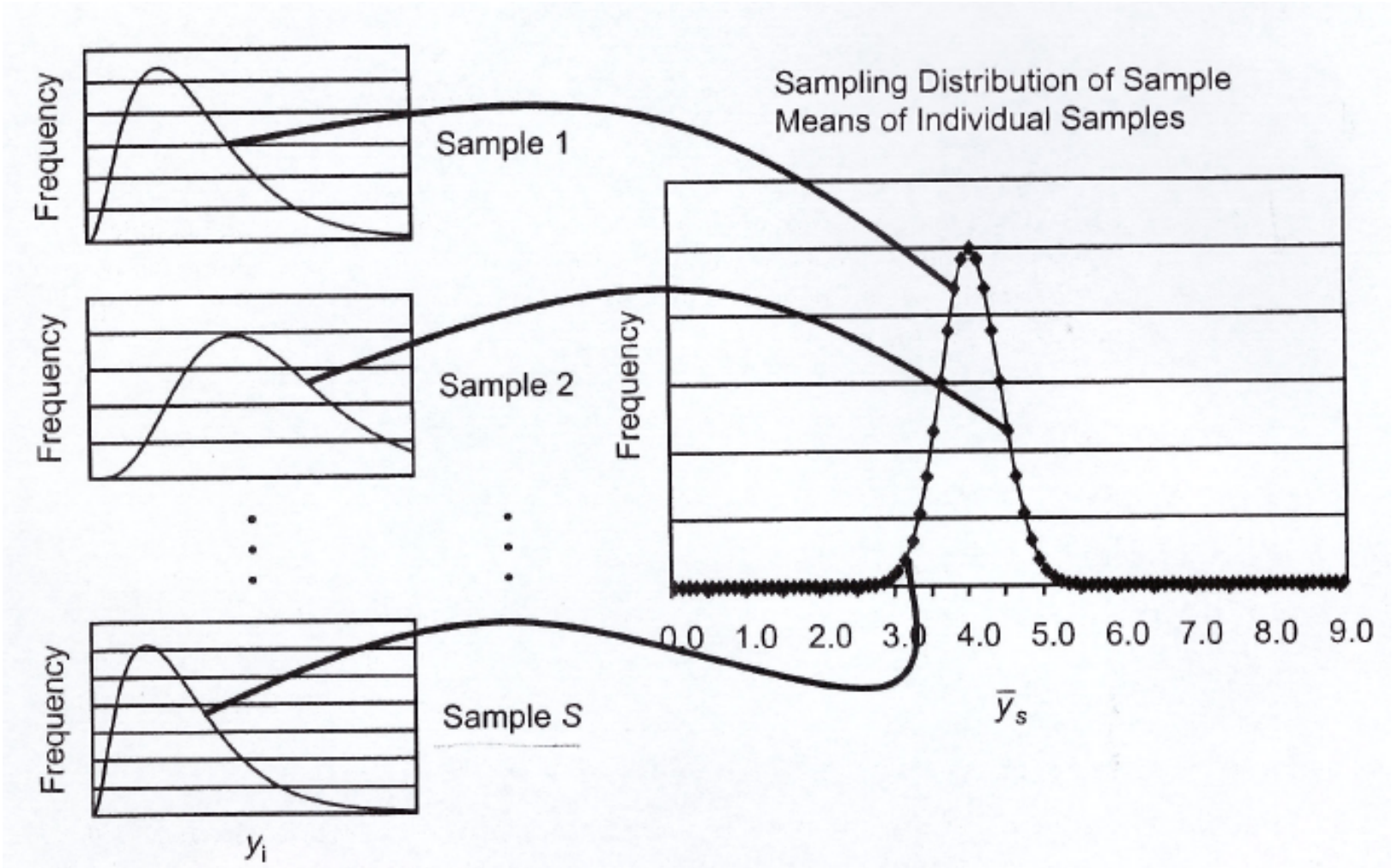


Figure 2.7 Samples and the sampling distribution of the mean.

Expectation & Variance of the Sample Mean

Theorem A (UNBIASEDNESS) :

under **s.r.s.** , $E(\bar{X}) = \mu$

Theorem B: under **s.r.s.** , $Var(\bar{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$

Recall: The variance of $\bar{X}$ in sampling without replacement differs from that in sampling with replacement by the factor $\left(1 - \frac{n-1}{N-1}\right)$ which is called the finite population correction.

The ratio $\frac{n}{N}$ is called the sampling fraction.

Estimation of the Population Variance

Theorem A: under **s.r.s.**,

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

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

Corollary A: An unbiased estimator of $Var\left(\bar{X}\right)$ is

given by $$s_{\bar{X}}^2 = \frac{\hat{\sigma}^2}{n}\left(\frac{n}{n-1}\right)\left(\frac{N-1}{N}\right)\left(\frac{N-n}{N-1}\right) = \frac{s^2}{n}\left(1-\frac{n}{N}\right)$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

Normal approximation to the sampling distribution of the sample mean

We will be using the CLT (**Central Limit Theorem**) in order to find the probabilistic bounds for the estimation error.

Application 1:

- the probability that the error made in estimating μ by $\overline{X}$ is $P(|\overline{X}-\mu|\le\delta)\cong 2\Phi\left(\frac{\delta}{\sigma_{\overline{X}}}\right)-1$ using the CLT.

Application 2:

- a $100(1-\alpha)\%$ CI (**Confidence Interval**) for the pop'n mean μ is given by $\overline{X}\pm\sigma_{\overline{X}} * z_{\alpha/2}$ using the CLT.

Estimating a Ratio

Ratio arises frequently in Survey Sampling.

If a bivariate sample (X_i, Y_i) is drawn, then the ratio

$$r = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i}{\frac{1}{N}\sum_{i=1}^{N} x_i}$$

$$R = \bar{Y} / \bar{X}$$

Estimating a Ratio (cont'd)

Theorem A:

With **s.r.s.**, the approximate variance of R is

$$Var(R) \cong \frac{1}{\mu_x^2}\left(r^2\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 - 2r\sigma_{\bar{X}\bar{Y}}\right)$$

$$= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy}\right)$$

Since the population correlation pop'n is $\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$ then

$$Var(R) \cong \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\rho\sigma_x\sigma_y\right)$$

Estimating a Ratio (cont'd)

Theorem B:

With **s.r.s.**, the approximate expectation of R is

$$E(R) \cong r + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 - \rho\sigma_x\sigma_y\right)$$

Standard Error estimate of R

The estimate variance of R is

$$s_R^2 \cong \frac{1}{n}\left(1-\frac{n-1}{N-1}\right)\frac{1}{\overline{X}^2}\left(R^2 s_x^2 + s_y^2 - 2Rs_{xy}\right)$$

where and the pop'n covariance

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i-\mu_x\right)\left(y_i-\mu_y\right)$$

is estimated by

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)$$

$$and \quad \rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \Rightarrow \hat{\rho} = \frac{s_{xy}}{s_x s_y}$$

Confidence Interval for R

An approximate $100(1-\alpha)\%$ CI (Confidence Interval) for the ratio of interest r is given by

$$R \pm s_R * z_{\alpha/2}$$

Stratified Random Sampling

Introduction (S.R.S.)

The population is partitioned into sub-pop's or strata (stratum, singular) that are then independently sampled and are combined to estimate population parameters.

A stratum is a subset of the population that shares at least one common characteristic

Example: males & females; age groups;…

Why is Stratified Sampling superior to Simple Random Sampling?

- **S.R.S.** reduces the sampling error
- **S.R.S.** guarantees a prescribed number of observations from each stratum while **s.r.s.** can't
- The mean of a **S.R.S.** can be considerably more precise than the mean of a **s.r.s.**, if the population members within each stratum are relatively homogeneous and if there is enough variation between strata.

Properties of Stratified Estimates

Notation: Let $N = N_1 + N_2 + \ldots + N_L$ be the total pop'n size if $N_l \ for \ \ l = 1,2,\ldots,L$ denote the pop'n sizes in the L strata.

The overall population mean

$$\mu = \frac{\sum_{l=1}^{L} W_l \ \mu_l}{\sum_{l=1}^{L} W_l} = \sum_{l=1}^{L} W_l \mu_l \ \underline{\underline{because}} \sum_{l=1}^{L} W_l = 1$$

is a weighted average of the pop'n means μ_l of the strata, where $W_l = N_l / N$ denotes the fraction of the l^{th} population in the stratum.

Properties of Stratified Estimates (cont'd)

Stratified sampling requires two steps:

- Identify the relevant strata in the population
- Use **s.r.s.** to get subject from each stratum

Within each stratum, a **s.r.s.** of size n_l is taken to obtain the sample mean in the l^{th} stratum will be denoted by $\bar{X}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} X_{il}$

where denotes the i^{th} observation in the l^{th} stratum.

Properties of Stratified Estimates (cont'd)

Theorem A: The stratified estimate, $\bar{X}_s = \sum_{l=1}^{L} W_l \bar{X}_l$, of the overall population mean μ is UNBIASED.

- *Since we assume that the samples from different strata are independent of one another and that within each stratum a* ***s.r.s.*** *is taken, then the variance of* $\bar{X}_s$ *can be easily calculated*

Theorem B: The variance of the stratified sample is

$$Var(\bar{X}_s) = \sum_{l=1}^{L} W_l^2 \left(\frac{1}{n_l} \right) \left(1 - \frac{n_l - 1}{N_l - 1} \right) \sigma_l^2$$

Neglecting / Ignoring the finite population correction

Approximation:

If the sampling fractions $W_l = \frac{N_l}{N}$ within all strata

were small, Theorem B will then reduce to:

$$Var\left(\overline{X}_s\right) \cong \sum_{l=1}^{L} \frac{W_l^2}{n_l} \sigma_l^2$$

Expectation and Variance of the stratified estimate of the population total

This is a corollary of Theorems A & B.

$$E(T_s) = \tau$$

$$Var(T_s) = N^2 Var(\bar{X}_s) = \sum_{l=1}^{L} N_l^2 \left(\frac{1}{n_l} \right) \left(1 - \frac{n_l - 1}{N_l - 1} \right) \sigma_l^2$$

$$where\ T_s = N\bar{X}_s$$

Methods of units' allocation in strata

For small sampling fractions within strata i.e. when neglecting/ignoring the finite pop'n correction,

$$Var(\bar{X}_s) \cong \sum_{l=1}^{L} \frac{W_l^2}{n_l} \sigma_l^2$$

Question: How to choose $n_1, n_2, \ldots, n_L$ to minimize $Var(\bar{X}_s)$ subject to the constraint $n = n_1 + n_2 + \ldots + n_L$ when resources of a survey allowed only a total of n units to be sampled?

Note: We could include finite population corrections but the results will be more complicated. Try it!

Neyman allocation

Theorem A: The samples sizes $n_1, n_2, \ldots, n_L$ that minimize $Var(\bar{X}_s)$ subject to the constraint

$$n = n_1 + n_2 + \ldots + n_L$$

are given by

$$n_l = n \frac{W_l \sigma_l}{\sum_{k=1}^{L} W_k \sigma_k} \quad where\ l = 1,2,\ldots,L$$

Corollary A: stratified estimate & optimal allocations

$$Var(\bar{X}_{so}) = \frac{\left(\sum_{l=1}^{L} W_l \sigma_l \right)^2}{n}$$

Proportional allocation

If a survey measures several attributes for each population member, it will be difficult to find an allocation that is simultaneously optimal for each of those variables.

Using the same sampling fraction

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots. = \frac{n_L}{N_L}$$

$$\Rightarrow n_l = n\frac{N_1}{N} = nW_l \;\; for\; l = 1,2,\ldots,L$$

in each stratum will provide a simple and popular alternative method of allocation.

Proportional allocation (cont')

Theorem B: With stratified sampling based on proportional allocation, ignoring the finite population correction,

$$Var\left(\bar{X}_{sp}\right)=\frac{1}{n}\sum_{l=1}^{L}W_l\sigma_l^2$$

Theorem C: With stratified sampling based on both allocation methods, ignoring the finite pop'n correction,

$$Var\left(\bar{X}_{sp}\right)-Var\left(\bar{X}_{so}\right)=\frac{1}{n}\sum_{l=1}^{L}W_l\left(\sigma_l-\bar{\sigma}\right)^2$$

$$where\ \ \bar{\sigma}=\sum_{l=1}^{L}W_l\sigma_l$$

Conclusions

A mathematical model for Survey Sampling was built using **s.r.s.** and probabilistic error bounds for the estimates derived.

The theory and techniques of survey *probability* sampling include also Systematic Sampling, Cluster Sampling, etc…as well as *non-probability* sampling methods such as Quota Sampling, Snowball Sampling,…