Sampling Theory and Practice

STAT 454[1]

Winter 2022 (1221)[2]

Cameron Roopnarine[3] Changbao Wu[4]

3rd March 2022

[1] STAT 454≡ STAT 854

[2] Online Course until 7th February 2022

[3] LaTeXer

[4] Instructor

Contents

Chapter 1

Review of Basic Concepts in Survey Sampling

Lecture 1
5th January

Survey sampling as a scientific discipline:

- Started from Jerzy Neyman's 1934 paper (1894–1981).
- Fast development since the 1940s and 1950s.
- Became an important area of statistics and social science.
- (Used to be) the primary tool of data collection for official statistics and researchers in social science and health studies.
- Face challenges in the big data and internet era.

Some ongoing well-known surveys:

- The Current Population Survey of the US (CPS).
- The National Health and Nutrition Examination Survey of the US (NHANES).
- The General Social Survey of Canada (GSS).
- The Canadian Community Health Survey (CCHS).
- The International Tobacco Control Policy Evaluation Surveys (The ITC Surveys, headquartered at UWaterloo).
- The Canadian Longitudinal Study of Aging (CLSA, McMaster, McGill, and Dalhousie).

Statistics Canada:
One of the most respected survey organizations in the world.

Some Canadian survey statisticians:

- J.N.K. Rao (Carleton University, retired).
- David Bellhouse (University of Western Ontario, retired).
- Jiahua Chen (University of British Columbia).
- David Haziza (University of Ottawa).
- Carl E. Särndal (University of Montreal, retired).
- Louis-Paul Rivest (Laval University).

- David Binder (Statistics Canada, 1949–2012).
- Carl Schwarz (Simon Fraser University, retired).
- Steve Thompson (Simon Fraser University).
- Randy Sitter (Simon Fraser University, 1961–2007).
- V. P. Godambe (University of Waterloo, 1926–2016).
- Mary E. Thompson (University of Waterloo, retired).
- Matthias Schonlau (University of Waterloo).
- Changbao Wu (University of Waterloo).

Example 1.1. The Math Faculty plans to conduct a survey to study the well-being of recent graduates from the faculty.

- What is exactly the group to be studied?
  (The target population)
- What information is to be collected?
  (Variables to be measured; sample data)
- From what can we select individuals to be surveyed?
  (Sampling frame(s))
- How to select individuals to be surveyed?
  (Sampling methods; sampling procedures)
- What method to use to collect data?
  (The mode of data collection: Face-to-face? Telephone? Mailed questionnaire?)
- How to use the data to draw conclusions?
  (Statistical analysis)

Three versions of survey populations (with reference to Example 1.1):

- *The target population*: The set of all units covered by the main objective of the study.
  (All students who received a formal degree from Waterloo between 2016 and 2019)
- *The frame population*: The set of all units covered by the sampling frame(s).
  (Sampling frame: The list of personal email addresses of students who graduated between 2016 and 2019)
- *The sampled population* (*the study population*): The population represented by the sample. Under probability sampling, the sampled population is the set of all units which have a non-zero probability to be selected in the sample.
  - The sampled population is not the set of sampled units!
  - Units which cannot be reached or do not respond to surveys (non-response) are not part of the sampled population.

Population structures and sampling frames:

$$U = \{1, 2, \ldots, N\},$$

where N is the population size, and the labels $1, 2, \ldots, N$ represent the N units.

- **Unstructured population**: There exists a single complete list of all N units, which can be used as the sampling frame.
- **Stratified population**: The population U has a stratified structure if it is divided into H non-overlapping subpopulations:
$$U = U_1 \cup U_2 \cup \cdots \cup U_H,$$

where the subpopulation U_h is called stratum h, with stratum population size N_h, $h = 1, 2, \ldots, H$. It follows that

$$N = \sum_{h=1}^{H} N_h.$$

Sampling frames for stratified sampling: H separate lists, each list consists of all units in one stratum.

- **Clustered population**: If the survey population can be divided into groups, called *clusters*, such that every unit in the population belongs to one and only one group, we say the population is clustered.

  First stage sampling frame for cluster sampling: A complete list of clusters (but not all the units within each cluster).

- Stratified sampling versus cluster sampling:
  - Under stratified sampling, sample data are collected from every stratum.
  - Under cluster sampling, only a portion of the clusters has members in the final sample.

Example 1.2. Survey of the population of high school students in the Waterloo region. There are a total of 15 high schools. Take a sample of 300 students from the population.

- **Plan A**. Randomly select 20 students from each high school.
  (Stratified sampling)
- **Plan B**. Randomly select 5 high schools from the list of 15 schools, and then randomly select 60 students from each of the 5 selected schools.
  (Two-stage cluster sampling)
- **Plan C**. The Waterloo region can be divided into KW area (8 high schools) and non-KW area (7 high schools). First, randomly select 3 schools from the KW area and 2 schools from the non-KW area, then randomly select 60 students from each of the 5 selected schools.
  (Stratified two-stage cluster sampling)

Sampling units and observational units:

- *Sampling units*: Units used to select the survey sample.
  - Under clustering sampling, sampling units are the clusters.
  - Under non-clustering sampling, sampling units are the individual units.
- *PSU and SSU*: Under two-stage cluster sampling, the first stage sampling units are clusters, called the *primary sampling unit* (PSU); the second stage sampling units are individual units, called the *secondary sampling unit* (SSU).
- *Observational units*: Observational units are always the individual units from which measurements are taken.

Example 1.3. An educational worker wanted to find out the average number of hours each week (of a certain month and year) spent on watching television by four and five-year-old children in the Waterloo Region. She conducted a survey using the list of 123 pre-school kindergartens administered by the Waterloo Region District School Board. She first randomly selected 10 kindergartens from the list. Within each selected kindergarten, she was able to obtain a complete list of all four and five-year-old children, with contact information for their parents/guardians. She then randomly selected 50 children from the list and mailed the survey questionnaire to their parents/guardians. The planned sample size is $10 \times 50 = 500$ and the sample data were compiled from those who completed and returned the questionnaires.

- *The target population*: All four and five-year-old children in the Region of Waterloo at the time of the survey. This is defined by the overall objective of the study.
- *Sampling frames*: Two-stage cluster sampling methods were used (further details to follow). The first stage sampling frame is the list of 123 kindergartens administered by the school board. The second

stage sampling frames are the complete lists of all four and five-year-old children for the 10 selected kindergartens.

- *Sampling units and observational units*: The first stage sampling units are the kindergartens; the second stage sampling units are the individual children (or equivalently, their parents); observational units are individual children.
- *The frame population*: All four and five-year-old children who attend one of the 123 kindergartens in the Region of Waterloo. It is apparent that children who are homeschooled are not covered by the frame population. Thus, as is frequently the case, the frame population is not the same as the target population.
- *The sampled population*: All four and five-year-old children who attend one of the 123 kindergartens in the Region of Waterloo and whose parents/guardians would complete and return the survey questionnaire if the child was selected for the survey.

LECTURE 2
10th January

Survey samples: A survey sample, denoted as S, is a subset of the population $U = \{1, 2, \ldots, N\}$.

Sample size $n = |S|$ is the number of units in the sample:

$$S = \{i_1, i_2, \ldots, i_n\}\text{a set of } n \text{ “unordered” units.}$$

We could simply use $S = \{1, 2, \ldots, n\}$.

$N = 10$, $n = 3$:

$$\begin{aligned} U &= \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \\ S &= \{7, 4, 9\} \\ &= \{i_1, i_2, i_3\}. \\ S &= \{1, 2, 3\}. \end{aligned}$$

Non-probability samples versus probability samples:
Non-probability samples are selected by subjective or any convenient methods. Examples include

- *Quota sampling*: The sample is obtained by a number of interviewers, each of whom is required to sample certain numbers of units with certain types or characteristics. How to select the units is completely left in the hands of the interviewers.
- *Judgement or purposive sampling*: The sample is selected based on what the sampler believes to be “typical” or “most representative” of the population.
- *Restricted sampling*: The sample is restricted to certain parts of the population which are readily accessible.
- *Sample of convenience*: The sample is taken from those who are easy to reach.
- *Sample of volunteers*: The sample consists of those who volunteer to participate.
- *Web panels*: The sample is selected from a panel of people who signed up to do surveys in order to receive cash or other incentives.

The most serious issue with non-probability survey samples:
Biased sample with unknown inclusion probabilities.

Non-probability survey samples are not the focus of this course. But the topic is becoming important in recent years, since data from non-probability survey samples become useful sources.

Yilin Chen’s PhD thesis research is on statistical analysis with non-probability survey samples, to be introduced in the last lecture.

Probability samples, theoretically speaking, are selected through a probability measure over a pool of candidate samples. Let

$$\Omega = \{S : S \subseteq U\}$$

be the set of all possible subsets of the survey population U. Let $\mathcal{P}$ be a probability measure over Ω such that

$$\mathcal{P}(S) \geq 0 \text{ for any } S \in \Omega \text{ and } \sum_{S:S\in\Omega} \mathcal{P}(S) = 1.$$

A probability sample S is selected based on the **probability sampling design**, $\mathcal{P}$.

$\mathcal{P}(\,\cdot\,)$ is a discrete probability measure.

Example 1.4. $N = 3$; $U = \{1, 2, 3\}$, $n = 1$ or 2.

- $n = 1$: $S_1 = \{1\}$, $S_2 = \{2\}$, $S_3 = \{3\}$.
- $n = 2$: $S_4 = \{1, 2\}$, $S_5 = \{1, 3\}$, $S_6 = \{2, 3\}$.
- $n = 3$: $S_7 = \{1, 2, 3\}$ (census).

S	S_1	S_2	S_3	S_4	S_5	S_6	S_7
$\mathcal{P}(S)$	1/6	1/6	1/6	1/6	1/6	1/6	0
$\mathcal{P}(S)$	0	0	0	1/3	1/3	1/3	0
$\mathcal{P}(S)$	0	0	0	1/2	1/4	1/4	0

Note that we have $\mathcal{P}(S) \geq 0$ for any $S \in \Omega$ and $\sum_{\{S:S\in\Omega\}} \mathcal{P}(S) = 1$.

Sampling design $\mathcal{P}$ with fixed sample size: $\mathcal{P}(S) = 0$ if $|S| \neq n$. The probability measure is defined over

$$\Omega_n = \{S : S \subseteq U \text{ and } |S| = n\}.$$

The cumulative sum method for generating a discrete random variable:

$$X \sim f(x): \quad p_i = f(x_i) = \mathsf{P}(X = x_i),\ i = 1, 2, \ldots.$$

- *Step 1*. Probability cumulation.

$$\begin{aligned} b_0 &= 0 \\ b_1 &= p_1 \\ b_2 &= p_1 + p_2 \\ b_3 &= p_1 + p_2 + p_3 \\ &\vdots \\ b_j &= \sum_{i=1}^{j} p_i \\ &\vdots \end{aligned}$$

- *Step 2*. Generate $r \sim U(0, 1)$.
- *Step 3*. Let $X^\star = x_j$ if $b_{j-1} < r \leq b_j$.

Can show $X \sim f(x)$.

Survey variables and population parameters:

- y: the response variable; $\boldsymbol{x}$ the vector of auxiliary variables.
- $(y_i; \boldsymbol{x}_i)$: the values of $(y, \boldsymbol{x})$ associated with unit i, $i = 1, 2, \ldots, N$.
- A common assumption in survey sampling: the values $(y_i, \boldsymbol{x}_i)$ can be measured without error if i is selected in the sample.

- Population totals:
$$T_y = \sum_{i=1}^{N} y_i \text{ and } T_{\boldsymbol{x}} = \sum_{i=1}^{N} \boldsymbol{x}_i.$$
- Population means:
$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i \text{ and } \mu_{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i.$$
- Population variance of y:
$$\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - N\mu_y^2\right).$$

An important special case: y is a binary variable:

$$y_i = \begin{cases} 1, & \text{if unit } i \text{ has attribute "}A\text{"}, \\ 0, & \text{otherwise.} \end{cases}$$

N: the total number of units in the population (population size). M: the total number of units in the population having attribute "A."

- Population total:
$$T_y = \sum_{i=1}^{N} y_i = M.$$
- Population mean:
$$\mu_y = \frac{T_y}{N} = \frac{M}{N} = P,$$
where P is the population proportion of units with attribute "A."
- Population variance:
$$\begin{aligned} \sigma_y^2 &= \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - N\mu_y^2\right) \\ &= \frac{1}{N-1}(M - NP^2) \\ &= \frac{N}{N-1}P(1-P) \\ &\approx P(1-P) && \text{if } N \text{ is large.} \end{aligned}$$

Probability sampling and design-based inference:

- The survey population $U = \{1, 2, \ldots, N\}$ is viewed as fixed.
- The values y_i and $\boldsymbol{x}_i$ attached to unit i and the population parameters such as T_y and μ_y are also viewed as fixed.
- The values of the population parameters can be determined without error by conducting a census.
- The sample S is selected according to a probability sampling design $\mathcal{P}$.
- The sample S is a random set under $\mathcal{P}$.
- Each unit in the population has a probability to be included in the sample.

- Randomization is induced by the probability sampling design for the selection of the survey sample.

Basic sampling techniques and advanced topics:

- Basic sampling techniques and theory are developed for the estimation of the population total T_y and the population mean μ_y.
  (Chapters 1–5 in the textbook)
- The basic methods and theory can be extended to handle more advanced topics, such as design-based regression analysis using survey data.
  (Chapters 6–11 in the textbook)

Chapter 2

Review of Simple Random Sampling

2.1 Simple Random Sampling Without Replacement (SRSWOR)

The task: Select a sample of size n from a population of size N with equal probability among all candidate samples.

The total number of candidate samples: $\binom{N}{n} = \frac{N(N-1)\cdots(N-n+1)}{n!}$.

The probability measure for the sampling design:

$$\mathcal{P}(S) = \begin{cases} \frac{1}{\binom{N}{n}}, & \text{if } |S| = n \\ 0, & \text{if } |S| \neq n. \end{cases}$$

$\mathcal{P}(S)$ cannot be used to select a sample in practice. $N = 1000$, $n = 3$:

$$\binom{N}{n} = \frac{1000 \times 999 \times 998}{3}.$$

$\mathcal{P}(S)$ is a theoretical tool.

Sampling scheme or sampling procedure: Select the survey sample through a sequential draw-by-draw method; select units from the sampling frame, one-at-a-time, until the final sample is chosen.

SRSWOR is a sampling procedure to select a sample of size n with equal probability among all candidate samples.

The sampling frame for SRSWOR: A complete list of N units in the population.

The SRSWOR sampling procedure:

1. Select the first unit from the N units on the sampling frame with equal probabilities $1/N$; denote the selected unit as i_1;
2. Select the second unit from the remaining $N-1$ units on the sampling frame with equal probabilities $1/(N-1)$; denote the selected unit as i_2;
3. Continue the process and select the n^{th} unit from the remaining $N-n+1$ units on the sampling frame with equal probabilities $1/(N-n+1)$; denote the selected unit as i_n.

Theorem 2.1. Under simple random sampling without replacement, the selected sample satisfies the probability measure $\mathcal{P}$ specified as

$$\mathcal{P}(S) = \begin{cases} 1/\binom{N}{n}, & \text{if } |S| = n, \\ 0, & \text{otherwise.} \end{cases}$$

Let $S = \{i_1, i_2, \ldots, i_n\}$ be the final sample.

$$\mathcal{P}(S) = \frac{n(n-1)\cdots(2)(1)}{N(N-1)\cdots(N-n+1)} = \frac{1}{\binom{N}{n}}.$$

- Survey sample selection always focuses on units, that is, the labels.
- Survey sample data: $\big\{(y_i, x_i), i \in S\big\}$.

Lecture 3
12th October

Sample mean and sample variance:

$$\bar{y} = \frac{1}{n}\sum_{i\in S} y_i.$$

$$s_y^2 = \frac{1}{n-1}\sum_{i\in S}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i\in S} y_i^2 - n\bar{y}^2\right).$$

Remarks:

- The sample mean $\bar{y}$ and s_y^2 are useful statistics under simple random sampling, but not necessarily under other sampling methods.
- The notation $\sum_{i\in S}$ is preferred over $\sum_{i=1}^{n}$.
- The form of estimators for population parameters depends on the sampling methods.
- The combination of "sampling design" and "estimation method" is called a "sampling strategy" (Thompson, 1997; Rao, 2005).

Expectation and variance under design-based inferences:

In classic statistics: $X_1, X_2, \ldots, X_n$ are iid with $\mathsf{E}[X_i] = \mu$, $\mathsf{V}(X_i) = \sigma^2$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

- **Sample mean**:
$$\mathsf{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$
- **Sample variance**:
$$\mathsf{V}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

Under SRSWOR:

$$\mathsf{E}[\bar{y}] = \mathsf{E}\left[\frac{1}{n}\sum_{i\in S} y_i\right] \neq \frac{1}{n}\sum_{i\in S}\mathsf{E}[y_i].$$

- S: a random set.
- $\sum_{i\in S}$: a random "sum."
- y_i: a fixed quantity for the given i.

Three fundamental results in survey sampling under SRSWOR:

(a) The sample mean $\bar{y} = n^{-1}\sum_{i\in S} y_i$ is a design-unbiased estimator for the population mean $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$: $\boxed{\mathsf{E}[\bar{y}] = \mu_y}$.

There are three possible ways to prove (a), depending on how the randomization under SRSWOR is handled.

Method 1. Use the probability measure $\mathcal{P}(S)$ for the survey design, that is, $\mathcal{P}(S) = \frac{1}{\binom{N}{n}}$ for $|S| = n$.

Also, $\bar{y}$ depends only on S.

$$\bar{y} = \frac{1}{n}\sum_{i\in S} y_i = \bar{y}(S),$$

that is, $\bar{y}$ is a function of S.

$$\begin{aligned}
\mathsf{E}[\bar{y}] &= \sum(\text{value})(\text{prob}) \\
&= \sum_S \bar{y}(S)\mathcal{P}(S) \\
&= \sum_{S:|S|=n} \frac{1}{n}\sum_{i\in S} y_i \frac{1}{\binom{N}{n}} \\
&= \frac{1}{n}\frac{1}{\binom{N}{n}} \sum_{\{S:|S|=n\}} \sum_{i\in S} y_i \\
&= \frac{1}{n}\frac{1}{\binom{N}{n}} \sum_{i=1}^{N} t_i y_i \\
&= \frac{1}{N}\sum_{i=1}^{N} y_i \\
&= \mu_y,
\end{aligned}$$

where $t_i =$ number of S which includes the unit i:

$$t_i = \binom{N-1}{n-1}.$$

$N = 3$, $n = 2$: $S_1 = \{1,2\}$, $S_2 = \{1,3\}$, $S_3 = \{2,3\}$.

$$\begin{aligned}
\sum_{\{S:|S|=2\}} \sum_{i\in S} y_i &= (y_1 + y_2) + (y_1 + y_3) + (y_2 + y_3) \\
&= 2y_1 + 2y_2 + 2y_3.
\end{aligned}$$

Method 2. Use the sampling scheme, the sequential draw-by-draw procedure. Let Z_k be the y-value from the k^{th} draw:

- $S = \{i_1, i_2, \ldots, i_n\}$.
- $Z_k = y_{ik}$ for $k = 1, 2, \ldots, n$.
- $\bar{y} = \frac{1}{n}\sum_{i\in S} y_i = \frac{1}{n}\sum_{k=1}^{n} Z_k$.

Hence,

$$\mathsf{E}[\bar{y}] = \mathsf{E}\left[\frac{1}{n}\sum_{k=1}^{n} Z_k\right] = \frac{1}{n}\sum_{k=1}^{n}\mathsf{E}[Z_k].$$

What's the probability function of Z_k?

Z_k	y_1	y_2	$\cdots$	y_N
$f(\,\cdot\,)$	$1/N$	$1/N$	$\cdots$	$1/N$

Therefore,

$$\mathsf{E}[Z_k] = \sum_{i=1}^{N} y_i \frac{1}{N} = \mu_y.$$

Method 3. Use the sample inclusion indicator variables.

$$A_i = \begin{cases} 1, & \text{if } i \in S, \\ 0, & \text{if } i \notin S. \end{cases} \qquad i = 1, 2, \ldots, N.$$

The A_i's are random variables.

$$\begin{aligned}
\mathsf{P}(A_i = 1) &= p = \mathsf{P}(i \in S) = \frac{1 \times \binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}. \\
\mathsf{P}(A_i = 0) &= 1 - p. \\
\mathsf{E}[A_i] &= p = \frac{n}{N}. \\
\mathsf{V}(A_i) &= p(1-p) = \frac{n}{N}\left(1 - \frac{n}{N}\right). \\
\mathsf{E}[\bar{y}] &= \mathsf{E}\left[\frac{1}{n}\sum_{i \in S} y_i\right] \\
&= \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^{N} A_i y_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i \,\mathsf{E}[A_i] \\
&= \frac{1}{N}\sum_{i=1}^{N} y_i \\
&= \mu_y.
\end{aligned}$$

(b) The design-based variance of $\bar{y}$ under SRSWOR is given by

$$\mathsf{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_y^2}{n},$$

where σ_y^2 is the population variance. The term $(1 - n/N)$ is called the *finite population correction* (fpc) factor; The ratio n/N is called the *sampling fraction*.

This result can be proved using different methods. Use the indicator variables:

$$\begin{aligned}
\mathsf{V}(\bar{y}) &= \mathsf{V}\left(\frac{1}{n}\sum_{i=1}^{N} A_i y_i\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{N} y_i^2 \,\mathsf{V}(A_i) + \sum\sum_{i \neq j} y_i y_j \,\mathsf{Cov}(A_i, A_j)\right).
\end{aligned}$$

$$\begin{aligned}
\mathsf{V}(A_i) &= \frac{n}{N}\left(1 - \frac{n}{N}\right). \\
\mathsf{Cov}(A_i, A_j) &= \mathsf{E}[A_i A_j] - \underbrace{\mathsf{E}[A_i]}_{n/N}\,\underbrace{\mathsf{E}[A_j]}_{n/N}.
\end{aligned}$$

$$\begin{aligned}
\mathsf{E}[A_iA_j] &= \sum_i \sum_j a_i a_j \,\mathsf{P}(A_i = a_i)\,\mathsf{P}(A_j = a_j) \\
&= \mathsf{P}(A_i = 1, A_j = 1) \\
&= \mathsf{P}(i \in S, j \in S) \\
&= \frac{1 \times 1 \times \binom{N-2}{n-2}}{\binom{N}{n}} \\
&= \frac{n(n-1)}{N(N-1)}.
\end{aligned}$$

$$\begin{aligned}
\mu_y^2 &= \frac{1}{N^2}\left(\sum_{i=1}^N y_i\right)^2 \\
&= \frac{1}{N^2}\sum_{i=1}^N \sum_{j=1}^N y_i y_j \\
&= \frac{1}{N^2}\left(\sum_{i=1}^N y_i^2 + \sum_{i \neq j}\sum y_i y_j\right).
\end{aligned}$$

(c) The sample variance s_y^2 is an unbiased estimator for the population variance σ_y^2 under SRSWOR, i.e., $\boxed{\mathsf{E}[s_y^2] = \sigma_y^2}$.

(c)* An unbiased variance estimator for $\bar{y}$ is given by

$$\mathsf{v}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n},$$

which satisfies

$$\mathsf{E}[\mathsf{v}(\bar{y})] = \mathsf{V}(\bar{y}),$$

where

$$\mathsf{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_y^2}{n}.$$

$$\begin{aligned}
s_y^2 &= \frac{1}{n-1}\left(\sum_{i \in S} y_i^2 - n\bar{y}^2\right) \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i \in S} y_i^2\right) - \frac{n}{n-1}\bar{y}^2.
\end{aligned}$$

$\mathsf{E}[\bar{y}] = \mu_y$ implies

$$\mathsf{E}\left[\frac{1}{n}\sum_{i \in S} y_i^2\right] = \frac{1}{N}\sum_{i=1}^N y_i^2.$$

$$\begin{aligned}
\mathsf{E}[\bar{y}^2] &= \mathsf{V}(\bar{y}) + \big(\mathsf{E}[\bar{y}]\big)^2 \\
&= \left(1 - \frac{n}{N}\right)\frac{\sigma_y^2}{n} + \mu_y^2.
\end{aligned}$$

Homework: Show that

$$\mathsf{E}[s_y^2] = \sigma_y^2.$$

Summary of the main theoretical results under SRSWOR:

- The population mean μ_y and the population variance σ_y^2 are fixed (but unknown) population parameters.
- The sample mean $\bar{y}$ and the sample variance s_y^2 are random variables under the survey design.
- The $\bar{y}$ is an unbiased estimator μ_y: $\mathsf{E}[\bar{y}] = \mu_y$.
- $\mathsf{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_y^2}{n}$ is the theoretical variance of $\bar{y}$ and is a fixed, but unknown quantity depending on the population variance σ_y^2.
- $\mathsf{v}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}$ is unbiased estimator for $\bar{y}$ (computable with the given sample data).
- The population size N is known under SRSWOR. (As part of the sampling frame information).

The R function for SRSWOR and SRSWR (next section) with specified N and n: `sample(N,n)`

```
N=10
n=4
sam=sample(N,n)
> sam
[1] 7 1 4 2
sam=sample(N,n,replace=T)
> sam
[1] 6 6 6 1
N=100
n=4
sam=sample(N,n)
> sam
[1] 57 67 62 91
sam=sample(N,n,replace=T)
> sam
[1] 88 73 9 63
```

Lecture 4

17th January

Summary of theoretical results under SRSWOR:

1 $\mathsf{E}[\bar{y}] = \mu_y$.

2 $\mathsf{V}(\bar{y}) = (1 - \frac{n}{N})\frac{\sigma_y^2}{n} = (\frac{1}{n} - \frac{1}{N})\sigma_y^2$.

3 $\mathsf{E}[s_y^2] = \sigma_y^2$.

4 $\mathsf{E}[\mathsf{v}(\bar{y})] = \mathsf{V}(\bar{y})$, where $\mathsf{v}(\bar{y}) = (1 - \frac{n}{N})\frac{s_y^2}{n}$.

Special cases where the response variable y is binary:

- $\mu_y = \frac{M}{N} = P$; $\sigma_y^2 = \frac{N}{N-1}P(1-P) \approx P(1-P)$ if N is large.
- $\bar{y} = \frac{1}{n}\sum_{i \in S} y_i = \frac{m}{n} = p$.
  - $m = \#$ units in S with attribute A.
  - $p = \frac{m}{n}$ = sample proportion.
- $s_y^2 = \frac{n}{n-1}p(1-p) \approx p(1-p)$ if n is large.
- $\mathsf{E}[p] = P$; $\mathsf{v}(p) = (1 - \frac{n}{N})\frac{1}{n}\frac{n}{n-1}p(1-p) = (1 - \frac{n}{N})\frac{1}{n-1}p(1-p)$.

2.2 Simple Random Sampling With Replacement (SRSWR)

The required sampling frame:

A complete list of all N units in the population.

The sampling procedure:

1. Select the first unit from the N units on the sampling frame with equal probabilities $1/N$; denote the selected unit as i_1;
2. Select the second unit from the N units on the sampling frame with equal probabilities $1/N$; denote the selected unit as i_2;
3. Continue the process and select the nth unit from the N units on the sampling frame with equal probabilities $1/N$; denote the selected unit as i_n.

Note: SRSWR is not very useful in survey practice but has theoretical values due to its connection to iid samples.

Two possible treatments for SRSWR:

(1) Keep duplicated units

Let $S^* = \{i_1, i_2, \ldots, i_n\}$. Under SRSWR, certain units might be included in S^* more than once (S^* may include duplicated units).

Let $Z_k = y_{ik}$ be the y value from the kth selection, $k = 1, 2, \ldots, n$. Let the sample mean be computed as

$$\bar{Z} = \frac{1}{n}\sum_{k=1}^{n} Z_k.$$

We have

$$\mathsf{E}[\bar{Z}] = \mu_y \quad \text{and} \quad \mathsf{V}(\bar{Z}) = \left(1 - \frac{1}{N}\right)\frac{\sigma_y^2}{n}.$$

(i) The $Z_1, Z_2, \ldots, Z_n$ are iid random variables.

(ii) The common probability function for $Z_1, Z_2, \ldots, Z_n$:

Z_k	y_1	y_2	$\cdots$	y_N
$f(\,\cdot\,)$	$1/N$	$1/N$	$\cdots$	$1/N$

$\overset{\text{iid}}{\sim}$ Discrete Uniform

(iii) The mean and variance of Z_k:

$$\begin{aligned}
\mathsf{E}[Z_k] &= \sum_{i=1}^{N} y_i \times \frac{1}{N} = \mu_y. \\
\mathsf{V}(Z_k) &= \mathsf{E}\big[(Z_k - \mathsf{E}[Z_k])^2\big] \\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2 \\
&= \frac{N-1}{N}\sigma_y^2 \\
&= \left(1 - \frac{1}{N}\right)\sigma_y^2.
\end{aligned}$$

$$\mathsf{E}[\bar{Z}] = \mu_y, \qquad \mathsf{V}(\bar{Z}) = \frac{1}{n}\mathsf{V}(Z_1) = \left(1 - \frac{1}{N}\right)\frac{\sigma_y^2}{n}.$$

(2) Remove duplicated units

Let S be the set of distinct units from SRSWR; let $m = |S|$ be the number of distinct units.

Note: m is a random number under SRSWR.

The sample mean based on the m distinct units is computed as

$$\bar{y}_m = \frac{1}{m}\sum_{i\in S} y_i.$$

It can be shown that (Problem 2.2 of Chapter 2)

$$\mathsf{E}[\bar{y}_m] = \mu_y \quad \text{and} \quad \mathsf{V}(\bar{y}_m) = \left[\mathsf{E}\left[\frac{1}{m}\right] - \frac{1}{N}\right]\sigma_y^2.$$

(Proof is required for Stat 854).

Efficiency comparisons between SRSWOR and SRSWR:

Three estimators of the population mean μ_y (assume $n \ge 2$):

1. $\bar{y}$ under SRSWOR.
2. $\bar{Z}$ under SRSWR.
3. $\bar{y}_m$ under SRSWR.

- All three estimators are unbiased (first-order equivalence).
- $\bar{y}$ is more efficient than the other two in terms of variance:

$$\mathsf{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_y^2}{n} < \left(1 - \frac{1}{N}\right)\frac{\sigma_y^2}{n} = \mathsf{V}(\bar{Z}).$$
$$\mathsf{V}(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma_y^2 < \left[\mathsf{E}\left[\frac{1}{m}\right] - \frac{1}{N}\right]\sigma_y^2 = \mathsf{V}(\bar{y}_m).$$

Efficiency comparisons through Monte Carlo simulation studies:

1. Generate a finite population of size N, $\{y_1, y_2, \ldots, y_N\}$ (from any distribution), and compute μ_y: This is the “unknown” population mean.
2. Take a sample S of size n, and obtain the sample data $\{y_i, i \in S\}$; compute the estimate $\hat{\mu}_1$ for the estimator $\hat{\mu}_y$.
3. Repeat (2) a large number K (≥ 1000) times, independently, to obtain $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_K$.
4. Evaluate the performance of the estimator $\hat{\mu}_y$ using the relative bias (RB, in %) and the mean squared error (MSE) from the simulation:
   - RB (in %):
   $$RB = \frac{1}{K}\sum_{k=1}^{K}\frac{\hat{\mu}_k - \mu_y}{\mu_y} \times 100,$$
   if $\mu_y = 0$ then we use the regular bias.
   - MSE:
   $$MSE = \frac{1}{K}\sum_{k=1}^{K}(\hat{\mu}_k - \mu_y)^2 = \text{Var} + \text{Bias}^2.$$

   (The MSE $\approx$ the variance if the RB is very small (i.e., $< 1\%$)).

A simulation example in R comparing $\bar{y}$ and $\bar{Z}$:

```
set.seed(1234567,kind=NULL) #Results duplicable!
N=1000
n=200
```

```
Y=rexp(N)
muy=mean(Y)
RB=c(0,0)
MSE=c(0,0)
K=1000
for(k in 1:K){
sam1=sample(N,n)
ysam1=Y[sam1]
sam2=sample(N,n,replace=T)
ysam2=Y[sam2]
mu1=mean(ysam1)
mu2=mean(ysam2)
RB[1]=RB[1]+mu1-muy
RB[2]=RB[2]+mu2-muy
MSE[1]=MSE[1]+(mu1-muy)^2
MSE[2]=MSE[2]+(mu2-muy)^2
}
RB=(RB/(K*muy))*100
MSE=MSE/K
> RB
[1] 0.09909944 -0.45677068
> MSE
[1] 0.003621282 0.004583952
```

Re-do the simulation with $N = 20000$ **and** $n = 200$:

```
> RB
[1] 0.2398581 0.1099942
> MSE
[1] 0.005152047 0.005227150
```

Note: The rule of thumb on how many decimal points to be reported

- For RB in percentages, two decimal points, i.e., 0.10% and -0.46% from the 1st example.
- For MSE, use two or three nearest decimal points to reflect the difference, i.e., 0.0036 and 0.0046 from the 1st example.

Homework:

- Install the R package on your laptop https://www.r-project.org.
- Re-run the simulation study with a different seed for the random number generator and compare the results.
- Re-run the simulation study with fixed $n = 200$ and different sampling fractions $n/N = 1\%, 2\%, 5\%, 10\%$ and compare the results.
- **Challenge part**: Include $\bar{y}_m$ in the simulation and compare the results.

LECTURE 5
19th January

2.3 Central Limit Theorem and Confidence Intervals

Asymptotic framework for finite populations
(A frame to allow $n \to \infty$):

We assume there is a sequence of finite populations (indexed by ν) and an associated sequence of survey samples. Both the population size N_ν and the sample size n_ν go to infinite as $\nu \to \infty$. The particular finite population and the survey sample are part of the sequence.

We use $n \to \infty$ or $N \to \infty$, but the limiting process is under $\nu \to \infty$.

For stratified populations, there are two versions of the asymptotic framework:

- The total number of strata is bounded, but the stratum population sizes grow to infinity for the sequence of populations.
- The stratum population sizes are bounded, but the total number of strata goes to infinity for the sequence of populations.

The Hájek Theorem (1960)

Suppose that the sampling fraction $n/N \to f \in (0,1)$ as $n \to \infty$.

Suppose also that the population values of the response variable y satisfy

$$\lim_{N\to\infty} \frac{\max_{1\le i\le N}(y_i-\mu_y)^2}{\sum_{i=1}^N (y_i-\mu_y)^2} = 0.$$

Then under SRSWOR, the Wald-type statistic

$$\frac{\bar{y}-\mu_y}{\sqrt{\mathsf{v}(\bar{y})}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, where $\mathsf{v}(\bar{y}) = \big(1-\frac{n}{N}\big)\frac{s_y^2}{n}$ is the estimated variance of $\bar{y}$.

We also have

$$\frac{\bar{y}-\mu_y}{\sqrt{\mathsf{V}(\bar{y})}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, where $\mathsf{V}(\bar{y}) = \big(1-\frac{n}{N}\big)\frac{\sigma_y^2}{n}$ is the theoretical variance of $\bar{y}$.

Note that

$$\frac{\bar{y}-\mu_y}{\sqrt{\mathsf{V}(\bar{y})}} = \frac{\bar{y}-\mu_y}{\sqrt{\mathsf{v}(\bar{y})}} \cdot \frac{\sqrt{\mathsf{v}(\bar{y})}}{\sqrt{\mathsf{V}(\bar{y})}} \quad \text{and} \quad \frac{\sqrt{\mathsf{v}(\bar{y})}}{\sqrt{\mathsf{V}(\bar{y})}} \xrightarrow{p} 1.$$

2.4 Sample Size Calculation

One of the major questions for survey design and planning: How large should the sample size n be? The answer depends on three factors:

- The total budget for the survey.
- The cost for surveying one unit and taking all required measurements.
- The accuracy required for the main statistical inference problem from the survey data.

The answer also depends on the sampling methods: More efficient sampling methods require a smaller sample size to achieve the same goal. We discuss sample size calculation under the simple scenario where

- The sampling method is SRSWOR.
- The accuracy requirements are for estimating the population mean.

(1) Accuracy specified by the absolute tolerable error

We want the estimator $\bar{y}$ for estimating the parameter μ_y to satisfy

$$\mathsf{P}\big(|\bar{y}-\mu_y| \ge e\big) \le \alpha,$$

or equivalently,

$$\mathsf{P}\big(|\bar{y}-\mu_y| < e\big) \le 1-\alpha,$$

for a given $\alpha \in (0,1)$ and a pre-specified error margin e. What is the required n?

We assume that

$$\frac{\bar{y} - \mu_y}{\sqrt{\mathsf{V}(\bar{y})}} \xrightarrow{d} \mathcal{N}(0,1)$$

can be used as an approximation to derive the required sample size.

We compare

$$\mathsf{P}\left(\frac{|\bar{y} - \mu_y|}{\sqrt{\mathsf{V}(\bar{y})}} < \frac{e}{\mathsf{V}(\bar{y})}\right) \geq 1 - \alpha$$

with

$$\mathsf{P}\big(|Z| < Z_{\alpha/2}\big) = 1 - \alpha,$$

where $Z \sim \mathcal{N}(0,1)$ and $Z_{\alpha/2}$ is the upper $\alpha/2$ quantile of $\mathcal{N}(0,1)$.

$$\begin{aligned} \frac{e}{\sqrt{\mathsf{V}(\bar{y})}} = Z_{\alpha/2} \implies \mathsf{V}(\bar{y}) &= \frac{e^2}{Z_{\alpha/2}^2} \\ \left(\frac{1}{n} - \frac{1}{N}\right)\sigma_y^2 &= \frac{e^2}{Z_{\alpha/2}^2}. \end{aligned}$$

Doing some algebra,

$$\begin{aligned} n &= \frac{Z_{\alpha/2}^2 \sigma_y^2 / e^2}{1 + (Z_{\alpha/2}^2 \sigma_y^2 / e^2)/N} = \frac{n_0}{1 + n_0/N} < n_0 \\ n_0 &= Z_{\alpha/2}^2 \sigma_y^2 / e^2, \end{aligned}$$

where $n \approx n_0$ for large N ($N = +\infty$).

(2) Accuracy specified by the relative tolerable error

Suppose that $\mu_y \neq 0$. We want the estimator $\bar{y}$ satisfies

$$\mathsf{P}\left(\frac{|\bar{y} - \mu_y|}{|\mu_y|} \geq e\right) \leq \alpha.$$

What is the required n?

Why is sometimes relative tolerable error preferred?

The absolute tolerable error e specified in $|\bar{y} - \mu_y| < e$ is scale-dependent. The choice of e in the relative tolerable error is scale-free, and can easily be decided as, for instance, 0.01–0.03 (that is, 1%–3%).

The accuracy requirement can be re-written as

$$\mathsf{P}\big(|\bar{y} - \mu_y| \geq e^*\big) \leq \alpha,$$

where $e^* = e|\mu_y|$. The required sample size n is given by

$$n = \frac{n_0}{1 + n_0/N}.$$

$$\begin{aligned} n_0 &= Z_{\alpha/2}^2 \sigma_y^2 / (e^*)^2 \\ &= Z_{\alpha/2}^2 \left(\frac{\sigma_y^2}{\mu_y^2}\right) / e^2 \\ &= Z_{\alpha/2}^2 \big[\mathsf{CV}(y)\big]^2 / e^2, \end{aligned}$$

where $\mathsf{CV}(y) = \frac{\sigma_y}{\mu_y}$.

A useful result: if $y_i = ax_i$ for all i, then $\mathsf{CV}(y) = \mathsf{CV}(x)$.

Notes on sample size calculations:

- The question of sample size calculation or sample size determination is part of the survey planning; the actual survey sample data are not available at this stage.
- Formulas for sample size calculations typically involve unknown population quantities such as μ_y and σ_y^2.
- How to obtain the required population information to calculate n?
    - Existing data sources:
        - Other similar surveys.
        - Census data.
    - Pilot surveys
        - Do a small survey first ($n = 50$?)
- The population information for sample size calculations does not need to be very accurate, because the calculated n is used for survey planning, which needs to be further adjusted by cost and other factors.

Example 2.1. Sample size calculation for estimating a population proportion

Suppose that the goal is to estimate the population proportion $P = M/N$ using a survey sample to be selected by SRSWOR. Using the sample proportion $p = m/n$ to estimate P, a common absolute tolerable error is 3% and the α is set to 0.05. In other words, the estimation accuracy is specified as

$$\mathsf{P}\big(|p - P| \le 0.03\big) \ge 0.95.$$

Noting that $0.95 = 19/20$, the probability statement is often quoted in media reports as "*The result is accurate within three percentage points, 19 times out of 20*."

What is the required sample size n?

$$n = \frac{n_0}{1 + n_0/N} < n_0.$$

n_0 is a conservative choice for any N.

$$\sigma_y^2 \approx P(1-P) \le \frac{1}{4}.$$

$$\begin{aligned} n_0 &= Z_{\alpha/2}^2 \sigma_y^2 / e^2 \\ &= 1.96^2 \sigma_y^2 / 0.03^2 \\ &\le 1.96^2 \times \frac{1}{4} / 0.03^2 \\ &\approx 1067. \end{aligned}$$

Chapter 3

Stratified Sampling and Cluster Sampling

Lecture 6
24th January

3.1 Stratified Simple Random Sampling

The survey population is divided into H non-overlapping strata:

$$U = U_1 \cup \cdots \cup U_H,$$

with corresponding break-down of population size as

$$N = \sum_{h=1}^{H} N_h,$$

where N_h is the size of stratum h.

For any *stratified sampling* designs, there are two basic features:

- A sample S_h of size n_h is taken from stratum h using a chosen sampling design, and this is done for every stratum.
- The H stratum samples S_h, $h = 1, 2, \dots, H$ are selected independent of each other.

The stratum sample sizes $(n_1, n_2, \dots, n_H)$ are pre-determined at the design stage. The total sample size is

$$n = \sum_{h=1}^{H} n_h.$$

Stratified Simple Random Sampling:

The stratum sample S_h is selected by SRSWOR, for every stratum $h = 1, 2, \dots, H$.

The required sampling frames:

Complete list of N_h units in stratum h, for every stratum $h = 1, 2, \dots, H$.

Notes:

- The population size N and the stratum sizes N_h are all known under stratified sampling (as part of the frame information).

- Even if a complete list of all N units is available, it does not imply that stratified sampling frames are automatically available.

The stratum weights:

$$W_h = \frac{N_h}{N},\ h = 1, 2, \ldots, H.$$

$$\sum_{h=1}^{H} N_h = N, \qquad \sum_{h=1}^{H} W_h = 1.$$

The variables:

$(y_{hi}, \boldsymbol{x}_{hi})$: the value of $(y, \boldsymbol{x})$ for unit i in stratum h, $i = 1, 2, \ldots, N_h$, $h = 1, 2, \ldots, H$.

The population (i.e., the census) "data file":

$$\big\{(y_{hi}, \boldsymbol{x}_{hi}) : i = 1, 2, \ldots, N_h,\ h = 1, 2, \ldots, H\big\}.$$

The sample data set:

$$\big\{(y_{hi}, \boldsymbol{x}_{hi}) : i \in S_h,\ h = 1, 2, \ldots, H\big\}.$$

3.1.1 Population parameters

The stratum population mean and population total:

$$\mu_{yh} = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}, \quad \text{and} \quad T_{yh} = \sum_{i=1}^{N_h} y_{hi}.$$

$$T_{yh} = N_h \mu_{yh}, \qquad \text{and} \qquad \mu_{yh} = \frac{T_{yh}}{N_h}.$$

The overall population mean and population total:

$$\mu_y = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi}, \quad \text{and} \quad T_y = \sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi}.$$

The relations between μ_y, T_y and μ_{yh}, T_{yh}:

$$T_y = \sum_{h=1}^{H} T_{yh} = \sum_{h=1}^{H} N_h \mu_{yh}.$$

$$\mu_y = \sum_{h=1}^{H} W_h \mu_{yh}.$$

The stratum population variances:

$$\sigma_{yh}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h} (y_{hi} - \mu_{yh})^2,\ h = 1, 2, \ldots, H.$$

The overall population variance:

$$\sigma_y^2 = \frac{1}{N-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h} (y_{hi} - \mu_y)^2.$$

The relation between σ_y^2 and σ_{yh}^2:

$$\sigma_y^2 \approx \sum_{h=1}^{H} W_h \sigma_{yh}^2 + \sum_{h=1}^{H} W_h (\mu_{yh} - \mu_y)^2.$$

Total variation = Variation within stratum + Variation between strata.

$$\begin{aligned}
(N-1)\sigma_y^2 &= \sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\mu_y)^2 \\
&= \sum_{h=1}^{H}\sum_{i=1}^{N_h}\big((y_{hi}-\mu_{yh})+(\mu_{yh}-\mu_y)\big)^2 \\
&= \sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\mu_{yh})^2+\sum_{h=1}^{H}\sum_{i=1}^{N_h}(\mu_{yh}-\mu_y)^2+2\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\mu_{yh})(\mu_{yh}-\mu_y) \\
&= \sum_{h=1}^{H}(N_h-1)\sigma_{yh}^2+\sum_{h=1}^{H}N_h(\mu_{yh}-\mu_y)^2 \\
\sigma_y^2 &= \sum_{h=1}^{H}\frac{N_h-1}{N-1}\sigma_{yh}^2+\sum_{h=1}^{H}\frac{N_h}{N-1}(\mu_{yh}-\mu_y)^2,
\end{aligned}$$

where

$$W_h=\frac{N_h}{N}, \qquad \frac{N_h-1}{N-1}\approx W_h, \qquad \frac{N_h}{N-1}\approx W_h.$$

3.1.2 Sample data and summary statistics

Let's focus on the study variable y. The sample data under stratified sampling are given by

$$\{y_{hi}, i\in S_h,\ h=1,2,\ldots,H\}.$$

The **stratum sample mean** and the **stratum sample variance** are defined as

$$\bar{y}_h=\frac{1}{n_h}\sum_{i\in S_h}y_{hi}, \qquad s_{yh}^2=\frac{1}{n_h-1}\sum_{i\in S_h}(y_{hi}-\bar{y}_h)^2,$$

where n_h is the stratum sample size.

The overall sample mean

$$\bar{y}=\frac{1}{n}\sum_{h=1}^{H}\sum_{i\in S_h}y_{hi}$$

is not a useful statistic (generally a biased estimator for μ_y).

3.1.3 Estimation of the overall population mean μ_y

In general, the overall population mean $\mu_y=\sum_{h=1}^{H}W_h\mu_{yh}$ can be estimated by

$$\hat{\mu}_y=\sum_{h=1}^{H}W_h\hat{\mu}_{yh},$$

where $\hat{\mu}_{yh}$ is an estimator of μ_{yh} using the data from the h^{th} stratum.

Three general properties of $\hat{\mu}_y$ under any stratified sampling designs:

1. $\mathsf{E}[\hat{\mu}_y]=\sum_{h=1}^{H}W_h\,\mathsf{E}[\hat{\mu}_{yh}]$.
2. $\mathsf{V}(\hat{\mu}_y)=\sum_{h=1}^{H}W_h^2\,\mathsf{V}(\hat{\mu}_{yh})$.
3. $\mathsf{v}(\hat{\mu}_y)=\sum_{h=1}^{H}W_h^2\,\mathsf{v}(\hat{\mu}_{yh})$.

Estimation of μ_y under stratified simple random sampling:

$$\bar{y}_{\text{st}} = \sum_{h=1}^{H} W_h \bar{y}_h.$$

This is called the stratified sample mean estimator.

Under **stratified simple random sampling**,

- The stratum weights W_h, $h = 1, \ldots, H$ are known constants.
- $\bar{y}_h$, $h = 1, \ldots, H$ are independent.
- $\mathsf{E}[\bar{y}_h] = \mu_{yh}$.
- $\mathsf{E}[s_{yh}^2] = \sigma_{yh}^2$.
- $\mathsf{V}(\bar{y}_h) = \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_{yh}^2}{n_h}$.
- $\mathsf{v}(\bar{y}_h) = \left(1 - \frac{n_h}{N_h}\right) \frac{s_{yh}^2}{n_h}$.

Three main properties of $\bar{y}_{\text{st}}$ under stratified simple random sampling:

(a) $\mathsf{E}[\bar{y}_{\text{st}}] = \sum_{h=1}^{H} W_h \, \mathsf{E}[\bar{y}_h] = \sum_{h=1}^{H} W_h \mu_{yh} = \mu_y$.

(b) $\mathsf{V}(\bar{y}_{\text{st}}) = \sum_{h=1}^{H} W_h^2 \, \mathsf{V}(\bar{y}_h) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_{yh}^2}{n_h}$.

(c) $\mathsf{v}(\bar{y}_{\text{st}}) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{yh}^2}{n_h}$.

Homework: Show that the overall sample mean

$$\bar{y} = \frac{1}{n} \sum_{h=1}^{H} \sum_{i \in S_h} y_{hi}$$

is not an unbiased estimator of μ_y under stratified simple random sampling unless

$$\frac{n_h}{n} = W_h, \; h = 1, \ldots, H.$$

(This is called the so-called proportional sample size allocation)

$$\bar{y} = \frac{1}{n} \sum_{h=1}^{H} n_h \bar{y}_h.$$

3.1.4 Justifications for using stratified sampling

- *Administrative convenience*. A survey at the national level can be organized more conveniently if each province surveys the allocated portion of the sample independently. In this case the provinces would be a natural choice for stratification.
- *Estimation of subpopulation parameters*. Large surveys often have multiple objectives. In addition to estimates for the entire population, estimates for certain subpopulations could also be required.
- *Efficiency considerations*. With suitable stratification and reasonable sample size allocation, stratified sampling leads to more efficient statistical inference.
- *More balanced or controlled samples*. Stratified sampling can protect from possible disproportionate samples under probability sampling among subpopulations

LECTURE 7
26th January

3.2 Sample Size Allocation Under Stratified Sampling

Sample size allocations need to be addressed at the survey design stage. There are practical constraints on sample size allocations.

We consider three theoretical questions on sample size allocations:

- For a given overall sample size n, how to find the "optimal allocation" $(n_1, n_2, \ldots, n_H)$?
- For a total cost C and cost per unit, how to find the "optimal allocation?"
- For a pre-specified requirement on variance of the estimators, how to find the "optimal allocation?"

Sample size allocations can be complicated by the use of complex survey sampling methods within each of the strata and more advanced inferential problems.

We focus on stratified simple random sampling and the estimation of the population mean μ_y.

3.2.1 Proportional allocation

The overall sample size n has already been decided. The question is how to choose n_h such that $\sum_{h=1}^H n_h = n$.

The **proportional allocation** method chooses $n_h \propto N_h$ under the constraint $\sum_{h=1}^H n_h = n$.

$$n_h = cN_h,\ h = 1, \ldots, H.$$

$$n = \sum_{h=1}^H cN_h = cN.$$

$$c = \frac{n}{N}, \qquad n_h = \frac{n}{N} N_h.$$

The allocation methods leads to

$$n_h = \frac{n}{N} N_h = nW_h,\ h = 1, \ldots, H.$$

Under stratified simple random sampling with proportional allocation:

- The point estimator $\bar{y}_{\text{st}}$ remains unbiased for μ_y.
- The theoretical variance formula $\mathsf{V}(\bar{y}_{\text{st}})$ reduces to

$$\mathsf{V}_{\text{prop}}(\bar{y}_{\text{st}}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{h=1}^H W_h \sigma_{yh}^2.$$

$$\mathsf{V}(\bar{y}_{\text{st}}) = \sum_{h=1}^H W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_{yh}^2}{n_h}.$$

$$n_h = nW_h = n\frac{N_h}{N} \implies \frac{n_h}{N_h} = \frac{n}{N}.$$

$$W_h \cdot \frac{1}{n_h} = \frac{1}{n}.$$

$$W_h \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} = \left(1 - \frac{n}{N}\right) \frac{1}{n}.$$

A comparison between $\bar{y}$ under SRSWOR and $\bar{y}_{\text{st}}$ under stratified simple random sampling with proportional allocation, with the same overall sample size n:

- Point estimators: Both $\bar{y}$ and $\bar{y}_{\text{st}}$ are unbiased for μ_y under the respective sampling design.

- The two variances satisfy

$$\mathsf{V}(\bar{y}) - \mathsf{V}_{\text{prop}}(\bar{y}_{\text{st}}) \approx \left(1 - \frac{n}{N}\right)\frac{1}{n}\sum_{h=1}^{H} W_h(\mu_{yh} - \mu_y)^2.$$

$$\mathsf{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\sigma_y^2.$$

$$\mathsf{V}_{\text{prop}}(\bar{y}_{\text{st}}) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\sum_{h=1}^{H} W_h\sigma_{yh}^2.$$

$$\sigma_y^2 \approx \sum_{h=1}^{H} W_h\sigma_{yh}^2 + \sum_{h=1}^{H} W_h(\mu_{yh} - \mu_y)^2.$$

Two important implications:

- The stratified simple random sampling design under proportional sample size allocation always provides more efficient estimate of the population mean than SRSWOR.
- The gain of efficiency under stratified sampling is larger when units within each stratum are more homogeneous, or equivalently, the units from different strata are more heterogeneous so that the between strata variation is large.

Example. $N = 10$; $\{y_1, \ldots, y_{10}\} = \{0, 1, 0, 0, 1, 1, 1, 0, 0, 0\}$.

$$\mu_y = \frac{4}{10}, \qquad \sigma_y^2 = \frac{N}{N-1}P(1-P) = \frac{10}{9}\frac{4}{10}\frac{6}{10}.$$

Take a sample with $n = 4$:

(a) SRSWOR: $\bar{y}$, $\mathsf{E}[\bar{y}] = \mu_y$, $\mathsf{V}(\bar{y}) = \cdots$.

(b) Stratified sampling: $N_1 = 6$, $N_2 = 4$.

$$\underbrace{\{0, 0, 0, 0, 0, 0\}}_{n_1=2}, \qquad \underbrace{\{1, 1, 1, 1\}}_{n_2=2}.$$

$$\bar{y}_{\text{st}} = W_1\bar{y}_1 + W_2\bar{y}_2 = \frac{6}{10} \times 0 + \frac{4}{10} \times 1 = \frac{4}{10} = \mu_y.$$

3.2.2 Neyman allocation

The overall sample size n is fixed. Find the optimal allocation $(n_1, \ldots, n_H)$ such that $\mathsf{V}(\bar{y}_{\text{st}})$ is minimized subject to the constraint $\sum_{h=1}^{H} n_h = n$.

This is called the *Neyman allocation* (Neyman, 1934). The solution is given by

$$n_h \propto W_h\sigma_{yh}, \ h = 1, 2, \ldots, H.$$

The constraint $\sum_{h=1}^{H} n_h = n$ leads to the allocation formula

$$n_h = n\frac{W_h\sigma_{yh}}{\sum_{k=1}^{H} W_k\sigma_{yk}} = n\frac{N_h\sigma_{yh}}{\sum_{k=1}^{H} N_k\sigma_{yk}}, \ h = 1, 2, \ldots, H.$$

$$n_h = cW_h\sigma_{yh}, \ h = 1, \ldots, H.$$

$$n = \sum_{h=1}^{H} n_h = c\sum_{h=1}^{H} W_h\sigma_{yh}.$$

$$c = \frac{n}{\sum_{h=1}^{H} W_h \sigma_{yh}}.$$

$$\begin{aligned}\mathsf{V}(\bar{y}_{\mathrm{st}}) &= \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_{yh}^2}{n_h} \\ &= \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \sigma_{yh}^2.\end{aligned}$$

$$\begin{aligned}L(n_1, \ldots, n_H) &= \mathsf{V}(\bar{y}_{\mathrm{st}}) + \lambda\left(\sum_{h=1}^{H} n_h - n\right). \\ 0 = \frac{\partial L}{\partial n_h} &= -\frac{1}{n_h^2} W_h^2 \sigma_{yh}^2 + \lambda. \\ n_h^2 &= \frac{1}{\lambda} W_h^2 \sigma_{yh}^2,\end{aligned}$$

that is, $n_h \propto W_h \sigma_{yh}$.

The theoretical variance $\mathsf{V}(\bar{y}_{\mathrm{st}})$ under Neyman allocation reduces to

$$\mathsf{V}_{\mathrm{neym}}(\bar{y}_{\mathrm{st}}) = \frac{1}{n}\left(\sum_{h=1}^{H} W_h \sigma_{yh}\right)^2 - \frac{1}{N} \sum_{h=1}^{H} W_h \sigma_{yh}^2.$$

(Details can be skipped)

Two major implications of Neyman allocation:

$$n_h \propto W_h \sigma_{yh},\ h = 1, 2, \ldots, H.$$

- Under Neyman allocation, population strata with bigger size N_h or bigger variation (i.e., bigger σ_{yh}^2) or both should be assigned to a bigger sample size n_h.
- If all strata have similar variation, i.e., similar values of σ_{yh}^2, Neyman allocation reduces to $n_h \propto W_h$, which is proportional allocation.

3.2.3 Optimal allocation with pre-specified cost or variance

The total direct cost for the overall sample is C_1. Cost for sampling one unit in stratum h is c_h. The cost constraint for allocation $(n_1, \ldots, n_H)$:

$$C_1 = \sum_{h=1}^{H} c_h n_h.$$

The variance formula $\mathsf{V}(\bar{y}_{\mathrm{st}})$ can be re-written as

$$\mathsf{V}(\bar{y}_{\mathrm{st}}) = \sum_{h=1}^{H} W_h^2 \frac{\sigma_{yh}^2}{n_h} - \sum_{h=1}^{H} W_h^2 \frac{\sigma_{yh}^2}{N_h}.$$

The variance constraint for allocation $(n_1, \ldots, n_H)$:

$$V_1 = \sum_{h=1}^{H} W_h^2 \frac{\sigma_{yh}^2}{n_h}.$$

Under both allocation constraints, the overall sample size n depends on C_1 and V_1.

Two optimal allocation methods: Find $(n_1, \ldots, n_H)$ to

- Minimize V_1 with a pre-specified C_1.
- Minimize C_1 with a pre-specified V_1.

The solution to (either) the optimal allocation:

$$n_h \propto W_h\sigma_{yh}/\sqrt{c_h},\ h = 1, 2, \ldots, H.$$

The formulas for calculating the n_h are given by

$$n_h = n\frac{W_h\sigma_{yh}/\sqrt{c_h}}{\sum_{k=1}^{H} W_k\sigma_{yk}/\sqrt{c_k}},\ h = 1, 2, \ldots, H,$$

where the overall sample size n is determined by the pre-specified C_1 or V_1.

The Cauchy-Schwarz inequality:

$$\big(\mathsf{E}[XY]\big)^2 \le \mathsf{E}[X^2]\,\mathsf{E}[Y^2].$$

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \le \sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2.$$

The equality holds iff $y_i = ax_i$ for all i.

Consider

$$\begin{aligned} V_1C_1 &= \left(\sum_{h=1}^{H} W_h^2\sigma_{yh}^2\frac{1}{n_h}\right)\cdot\left(\sum_{h=1}^{H} c_h n_h\right) \\ &\ge \left(\sum_{h=1}^{H} W_h\sigma_{yh}\frac{1}{\sqrt{n_h}}\cdot\sqrt{c_h}\sqrt{n_h}\right)^2, \end{aligned}$$

where we note that the RHS does not involve n_h.

The minimum of V_1C_1 is achieved when

$$W_h\sigma_{yh}\frac{1}{\sqrt{n_h}} \propto \sqrt{c_h}\sqrt{n_h} \implies n_h \propto W_h\sigma_{yh}/\sqrt{c_h}.$$

Implications of the two optimal allocation methods:

$$n_h \propto W_h\sigma_{yh}/\sqrt{c_h},\ h = 1, 2, \ldots, H.$$

1. With unequal costs for different strata, the more expensive stratum should be assigned a smaller sample size.
2. With equal cost for all strata, the two versions of optimal allocation both reduce to Neyman allocation, and hence the stratum sample size n_h is decided by the stratum population size N_h and the stratum variance σ_{yh}^2.
3. With equal or nearly equal cost and no information on stratum variations, proportional allocation would be the natural choice for sample size allocations.

3.3 Post-stratification

(Problem 3.7 in the textbook)

- Stratified sampling cannot be implemented if the sampling frames are not available, such as stratification by the gender and age groups for a large human population.

- Stratum membership can be determined relatively easily for all units in the sample once the sample is selected.
- Post-stratification: Divide a "non-stratified" sample into subsamples by the stratum membership, and construct a stratified estimator from the non-stratified sample data set, assuming the stratum weights W_h, $h = 1, \ldots, H$ are known.

Let $\{y_i, i \in S\}$ be the survey data set and S is a sample of size n selected by SRSWOR. The sample S can be post-stratified as

$$S = S_1 \cup \cdots \cup S_H,$$

with corresponding breakdown of n as $n = n_1 + \cdots + n_H$.

The post-stratified estimator of μ_y is computed as

$$\bar{y}_{\text{post}} = \sum_{h=1}^{H} W_h \bar{y}_h.$$

The key differences between $\bar{y}_{\text{post}}$ and $\bar{y}_{\text{st}}$:

- Under stratified sampling, the stratum sample sizes n_h are decided at the survey design stage and are fixed.
- Under post-stratification, the stratum sample sizes n_h are random numbers. The technical arguments for $\bar{y}_{\text{st}}$ cannot be used directly for $\bar{y}_{\text{post}}$.

Homework for STAT 854: Argue that the post-stratified estimator $\bar{y}_{\text{post}}$ is usually more efficient than $\bar{y}$ under SRSWOR. (Hint: Need to go through Problem 3.7)

LECTURE 8
31st January

Basic Concepts of Cluster Sampling:

- The population consists of K clusters (groups).
- *Single-stage cluster sampling*: A subset of the clusters is selected, and all units in the selected cluster are observed for the final sample.
- *Two-stage cluster sampling*: A subset of the clusters is selected, and within each selected cluster, a subset of units is selected for the final sample.
- Sampling frames for single-stage and two-stage cluster sampling:
  (1) First stage sampling frame: A complete list of clusters in the population
  (2) Second stage sampling frames: A complete list of units for each selected cluster
- More complex sampling designs: Stratified *multi-stage cluster sampling* with unequal selection probabilities at each stage.

3.4 Single-stage Cluster Sampling

3.4.1 Notation

- K: The total number of clusters in the population.
- M_i: The total number of units in cluster i.
- y_{ij}: The value of y for unit j in cluster i.
- $N = \sum_{i=1}^{K} M_i$: The overall population size.

The mean and the total for the i^{th} cluster are given by

$$\mu_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij}, \qquad T_i = \sum_{j=1}^{M_i} y_{ij} = M_i\mu_i,\ i = 1, 2, \ldots, K.$$

The population total is given by

$$T_y = \sum_{i=1}^{K}\sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{K} T_i = \sum_{i=1}^{K} M_i\mu_i,$$

and the population mean is given by $\mu_y = T_y/N$.

3.4.2 Single-stage cluster sampling with clusters selected by SRSWOR

The sampling procedure:

1. Select k clusters from the list of K clusters using SRSWOR, with a pre-specified k. Let S_c be the set of labels for the k selected clusters.
2. For $i \in S_c$, select all M_i units for the final sample.

The total number of units in the final sample (overall sample size):

$$n = \sum_{i \in S_c} M_i.$$

The sample data on the y-variable:

$$\{y_{ij} : j = 1, 2, \ldots, M_i,\ i \in S_c\}.$$

The cluster total $T_i = \sum_{j=1}^{M_i} y_{ij}$ is known for $i \in S_c$. The “condensed” sample data set:

$$\{T_i,\ i \in S_c\}.$$

Other information available from the sampling frames and the design:

- The total number of clusters, K.
- The number of clusters selected, k.
- The cluster size M_i for $i \in S_c$ (selected clusters). M_i may not be known if $i \notin S_c$.

Other notes:

- The overall sample size $n = \sum_{i \in S_c} M_i$ is typically a random number and is not controlled at the design stage except for the special case where the cluster sizes $M_i = M$ are all equal. In this case, $n = kM$.
- The overall population size $N = \sum_{i=1}^{K} M_i$ is often **unknown**.
- Estimation of $T_i = \sum_{i=1}^{K} T_i$ does not lead to estimation of $\mu_y = T_y/N$ and vice versa.
- Need to have an estimator for N.

3.4.3 Estimation of the population total T_y

Re-write the population total as

$$T_y = \sum_{i=1}^{K} T_i = K\left(\frac{1}{K}\sum_{i=1}^{K} T_i\right) = K\mu_T.$$

The question: Why do we introduce

$$\mu_T = \frac{1}{K}\sum_{i=1}^{K} T_i? \qquad \mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

The answer: The μ_T is not a parameter of interest, but it is a "population mean" and can be estimated by the corresponding "sample mean" under SRSWOR,

$$\hat{\mu}_T = \frac{1}{k}\sum_{i\in S_c} T_i. \qquad \bar{y} = \frac{1}{n}\sum_{i\in S} y_i.$$

This leads to $\hat{T}_y = K\hat{\mu}_T$, where K is known.

Main results on estimating T_y: Under single-stage cluster sampling with clusters selected by SRSWOR,

(a) An unbiased estimator for the population total T_y is given by

$$\hat{T}_y = K\Big(\frac{1}{k}\sum_{i\in S_c} T_i\Big) = K\hat{\mu}_T,$$

where $\hat{\mu}_T = k^{-1}\sum_{i\in S_c} T_i$ is the sample mean of cluster totals.

(b) The design-based variance of $\hat{T}_y$ is given by

$$\mathsf{V}(\hat{T}_y) = K^2\Big(1 - \frac{k}{K}\Big)\frac{\sigma_T^2}{k},$$

where $\sigma_T^2 = (K-1)^{-1}\sum_{i=1}^{K}(T_i - \mu_T)^2$, and $\mu_T = K^{-1}\sum_{i=1}^{K} T_i$ is the population mean of cluster totals.

(c) An unbiased variance estimator for $\hat{T}_y$ is given by

$$\mathsf{v}(\hat{T}_y) = K^2\Big(1 - \frac{k}{K}\Big)\frac{s_T^2}{k},$$

where $s_T^2 = (k-1)^{-1}\sum_{i\in S_c}(T_i - \hat{\mu}_T)^2$ and $\hat{\mu}_T = k^{-1}\sum_{i\in S_c} T_i$.

3.4.4 Estimation of the population mean μ_y

If N is known, we can simply use $\hat{\mu}_y = \hat{T}_y/N$.

If $N = \sum_{i=1}^{K} M_i$ is unknown, we re-write the population mean as

$$\mu_y = \frac{1}{N}\sum_{i=1}^{K} T_i = \frac{\sum_{i=1}^{K} T_i}{\sum_{i=1}^{K} M_i} = \frac{K^{-1}\sum_{i=1}^{K} T_i}{K^{-1}\sum_{i=1}^{K} M_i} = \frac{\mu_T}{\mu_M},$$

where

$$\mu_M = \frac{1}{K}\sum_{i=1}^{K} M_i$$

is the "population mean" for the variable M_i (average cluster size), and can be estimated by the corresponding "sample mean"

$$\hat{\mu}_M = \frac{1}{k}\sum_{i\in S_c} M_i.$$

The population mean μ_y can be estimated by

$$\hat{\mu}_y = \frac{\hat{\mu}_T}{\hat{\mu}_M} = \frac{k^{-1}\sum_{i\in S_c} T_i}{k^{-1}\sum_{i\in S_c} M_i} = \frac{\sum_{i\in S_c} T_i}{\sum_{i\in S_c} M_i} = \frac{1}{n}\sum_{i\in S_c}\sum_{j=1}^{M_i} y_{ij},$$

where $n = \sum_{i \in S_c} M_i$ is the overall sample size.

Notes:

- The overall sample size n is usually a random number.
- The $\hat{\mu}_y$ looks like a sample mean, but its theoretical properties need to be derived using a "ratio estimator."
- Ratio estimators will be discussed in Chapter 5.

3.4.5 A comparison between SRSWOR and Single-stage cluster sampling

(This is technically a challenge topic under general scenarios with unequal M_i)

Consider a simple scenario where

- All clusters have the same size: $M_i = M$ $(M \geq 2)$.
- The overall population size is $N = KM$ (and is known).
- The overall sample size is $n = kM$ (and is a fixed number).
- The sampling fraction $n/N = (kM)/(KM) = k/K$.
- The estimators $\hat{\mu}_M$ and $\hat{\mu}_y$ reduce to

$$\hat{\mu}_M = k^{-1} \sum_{i \in S_c} M_i = M, \qquad \hat{\mu}_y = \frac{\hat{\mu}_T}{\hat{\mu}_M} = M^{-1} \hat{\mu}_T,$$

  and $\hat{\mu}_T = \frac{1}{k} \sum_{i \in S_c} T_i$ is a "sample mean."
- The $\hat{\mu}_y$ is an unbiased estimator of μ_y.

Under single-stage cluster sampling with clusters selected by SRSWOR,

$$\mathsf{V}(\hat{\mu}_y) = \frac{1}{M^2}\left(1 - \frac{k}{K}\right)\frac{\sigma_T^2}{k} = \left(1 - \frac{n}{N}\right)\frac{M^{-1}\sigma_T^2}{n}.$$

It can be shown (Problem 3.8 for STAT 854)

$$M^{-1}\sigma_T^2 \approx \sigma_y^2 \big(1 + (M-1)\rho\big),$$

where ρ is the *intra-cluster correlation coefficient* and is defined as follows: Randomly select a cluster, and then randomly select two units from the cluster without replacement; let Z_1 and Z_2 be the values of y for the two selected units,

$$\rho = \frac{\mathsf{Cov}(Z_1, Z_2)}{\sqrt{\mathsf{V}(Z_1)\,\mathsf{V}(Z_2)}}.$$

We have

$$\mathsf{V}(\hat{\mu}_y) \approx \left(1 - \frac{n}{N}\right)\frac{\sigma_y^2}{n}\big(1 + (M-1)\rho\big).$$

Key results for the comparison of two sampling strategies:

- Under the simple scenario with single stage cluster sampling,

$$\mathsf{V}(\hat{\mu}_y) \approx \left(1 - \frac{n}{N}\right)\frac{\sigma_y^2}{n}\big(1 + (M-1)\rho\big).$$

- If we take a sample of the same overall size n by SRSWOR and use the sample mean $\bar{y}$ to estimate μ_y, we have

$$\mathsf{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_y^2}{n}.$$

(i) It is very common in survey practice that units within the same cluster are positively correlated, i.e., $\rho > 0$ and consequently single-stage cluster sampling is less efficient than SRSWOR.

(ii) For situations where $\rho < 0$, cluster sampling can be more efficient.

(iii) When $\rho = 0$, the clusters behave like random groups of units from the population. Under such scenarios single-stage cluster sampling will result in a final sample which is similar to the one selected by non-cluster sampling methods.

- **Homework**: Find examples of the three scenarios listed above.

LECTURE 9
2nd February

3.5 Two-stage Cluster Sampling

Primary sampling unit (PSU): clusters.
(The first-stage sample selects k clusters from the population of K clusters)

Secondary sampling unit (SSU): units within clusters.
(The second-stage sample selects m_i units from the list of M_i units if cluster i is selected in the first stage)

3.5.1 Two-stage cluster sampling with SRSWOR at both stages

The sampling procedures:

1. Select k clusters from the list of K clusters using SRSWOR, with a pre-specified k. Let S_c be the set of labels for the k selected clusters.
2. For $i \in S_c$ and a pre-specified m_i, select a second-stage sample S_i of m_i units from the list of M_i units in cluster i using SRSWOR; the processes are carried out independently for different clusters.

The overall sample size is

$$n = \sum_{i \in S_c} m_i.$$

The choice of m_i (as part of the survey design):

- A constant $m_i = m$ is used across all clusters; $n = mk$ is fixed.
- A fixed second-stage sampling fraction, i.e., choose m_i such that $m_i/M_i = c$ for a pre-specified proportion c across all clusters; $n = c\sum_{i\in S_c} M_i$ is a random number (e.g., $c = 5\%, 10\%, \ldots$).

The sample data on the y-variable:

$$\{y_{ij} : j \in S_i, i \in S_c\}.$$

Other information available:

- The total number of clusters, K, and the number of clusters sampled, k.
- The cluster size M_i and the second-stage sample size m_i for $i \in S_c$.

The cluster mean and the cluster variance (cluster level population parameters):

$$\mu_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij}, \qquad \sigma_i^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \mu_i)^2.$$

Note:

- Under stage-stage cluster sampling, both μ_i and σ_i^2 can be computed from the sample data for cluster $i \in S_c$.

- Under two-stage cluster sampling, both μ_i and σ_i^2 are **unknown** even if cluster i is selected in the first stage.
- Under two-stage cluster sampling, $T_i = \sum_{j=1}^{M_i} y_{ij} = M_i\mu_i$ are unknown.

3.5.2 Estimation of the population total T_y

The second-stage cluster sample mean and sample variance:

$$\bar{y}_i = \frac{1}{m_i}\sum_{j\in S_i} y_{ij}, \qquad s_i^2 = \frac{1}{m_i-1}\sum_{j\in S_i}(y_{ij}-\bar{y}_i)^2.$$

The second-stage sample S_i of size m_i is selected by SRSWOR from the cluster of M_i units:

$$\mathsf{E}[\bar{y}_i] = \mu_i, \qquad \mathsf{V}(\bar{y}_i) = \left(1-\frac{m_i}{M_i}\right)\frac{\sigma_i^2}{m_i}.$$

The cluster total $T_i = M_i\mu_i$ can be estimated by

$$\hat{T}_i = M_i\bar{y}_i.$$

Note: M_i is known for $i \in S_c$. (Part of second stage frame info).

The “introduced bridge parameter” $\mu_T = K^{-1}\sum_{i=1}^K T_i$ can be estimated by

$$\tilde{\mu}_T = \frac{1}{k}\sum_{i\in S_c}\hat{T}_i = \frac{1}{k}\sum_{i\in S_c} M_i\bar{y}_i.$$

Comparison to single-stage cluster sampling:

$$\hat{\mu}_T = \frac{1}{k}\sum_{i\in S_c} T_i.$$

(Difference in notation: tilde vs hat).

The population total $T_y = \sum_{i=1}^K T_i = K\mu_T$ can be estimated by

$$\tilde{T}_y = K\tilde{\mu}_T.$$

K is available from the first-stage sampling frame information.

1. $\mathsf{E}[\tilde{T}_y] = K\,\mathsf{E}[\tilde{\mu}_T]$.
2. $\mathsf{V}(\tilde{T}_y) = K^2\,\mathsf{V}(\tilde{\mu}_T)$.
3. $\mathsf{v}(\tilde{T}_y) = K^2\,\mathsf{v}(\tilde{\mu}_T)$.

Main theoretical results on $\tilde{\mu}_T$: Under two stage-cluster sampling with SRSWOR at both stages,

(a) The estimator $\tilde{\mu}_T$ is unbiased for μ_T.

(b) The design-based variance of $\tilde{\mu}_T$ is given by

$$\mathsf{V}(\tilde{\mu}_T) = \left(1-\frac{k}{K}\right)\frac{\sigma_T^2}{k} + \frac{1}{k}\frac{1}{K}\sum_{i=1}^K M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{\sigma_i^2}{m_i},$$

where $\sigma_T^2 = (K-1)^{-1}\sum_{i=1}^K (T_i-\mu_T)^2$ and σ_i^2 is the cluster variance.

(c) An unbiased variance estimator for $\tilde{\mu}_T$ is given by

$$\mathsf{v}(\tilde{\mu}_T) = \left(1-\frac{k}{K}\right)\frac{\hat{\sigma}_T^2}{k} + \frac{1}{K}\frac{1}{k}\sum_{i\in S_c} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i},$$

where $\hat{\sigma}_T^2 = (k-1)^{-1}\sum_{i\in S_c}(\hat{T}_i-\tilde{\mu}_T)^2$ and s_i^2 is the cluster sample variance.

Two technical arguments for the proofs of (a), (b) and (c):

1. For any random variables X and Y, we have

$$\mathsf{E}[X] = \mathsf{E}\big[\mathsf{E}[X \mid Y]\big],$$

and

$$\mathsf{V}(X) = \mathsf{E}\big[\mathsf{V}(X \mid Y)\big] + \mathsf{V}\big(\mathsf{E}[X \mid Y]\big).$$

The proofs involve

- $\mathsf{E}_1[\,\cdot\,]$ and $\mathsf{V}_1(\,\cdot\,)$: the expectation and the variance with respect to the first stage sampling design.
- $\mathsf{E}_2[\,\cdot\,]$ and $\mathsf{V}_2(\,\cdot\,)$: the conditional expectation and the conditional variance with respect to the second stage sampling design given the first stage sample.

Proof of (a):

$$\tilde{\mu}_T = \frac{1}{k}\sum_{i\in S_c} M_i \bar{y}_i.$$

$$\begin{aligned}
\mathsf{E}[\tilde{\mu}_T] &= \mathsf{E}_1\big[\mathsf{E}_2[\tilde{\mu}_T]\big] \\
&= \mathsf{E}_1\left[\frac{1}{k}\sum_{i\in S_c} M_i\, \mathsf{E}_2[\bar{y}_i]\right] \\
&= \mathsf{E}_1\left[\frac{1}{k}\sum_{i\in S_c} M_i \mu_i\right] \\
&= \mathsf{E}_1\left[\frac{1}{k}\sum_{i\in S_c} T_i\right] \\
&= \frac{1}{K}\sum_{i=1}^{K} T_i \\
&= \mu_T.
\end{aligned}$$

Proof of (b):

$$\tilde{\mu}_T = \frac{1}{k}\sum_{i\in S_c} M_i \bar{y}_i.$$

$$\begin{aligned}
\mathsf{V}(\tilde{\mu}_T) &= \mathsf{V}_1\big(\mathsf{E}_2[\tilde{\mu}_T]\big) + \mathsf{E}_1\big[\mathsf{V}_2(\tilde{\mu}_T)\big] \\
&= \mathsf{V}_1\left(\frac{1}{k}\sum_{i\in S_c} T_i\right) + \mathsf{E}_1\left[\frac{1}{k^2}\sum_{i\in S_c} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{\sigma_i^2}{m_i}\right] \\
&= \left(1-\frac{k}{K}\right)\frac{\sigma_T^2}{k} + \frac{1}{k}\,\mathsf{E}_1\underbrace{\left[\frac{1}{k}\sum_{i\in S_c} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{\sigma_i^2}{m_i}\right]}_{\text{first stage sample mean}} \\
&= \left(1-\frac{k}{K}\right)\frac{\sigma_T^2}{k} + \frac{1}{k}\underbrace{\frac{1}{K}\sum_{i=1}^{K} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{\sigma_i^2}{m_i}}_{\text{first stage population mean}}.
\end{aligned}$$

(2) Re-write $\mathsf{V}(\tilde{\mu}_T)$ as

$$\mathsf{V}(\tilde{\mu}_T) = \left(\frac{1}{k}-\frac{1}{K}\right)\sigma_T^2 + \frac{1}{k}W,$$

where

$$\sigma_T^2 = \frac{1}{K-1}\sum_{i=1}^{K}(T_i - \mu_T)^2, \qquad W = \frac{1}{K}\sum_{i=1}^{K} M_i^2\left(1 - \frac{m_i}{M_i}\right)\frac{\sigma_i^2}{m_i}.$$

The "plug-in" estimator

$$\hat{\sigma}_T^2 = \frac{1}{k-1}\sum_{i\in S_c}(\hat{T}_i - \tilde{\mu}_T)^2$$

is not unbiased for σ_T^2, and instead satisfies (**homework for STAT 854**, hints from Problem 3.11 in the textbook)

$$\mathsf{E}[\hat{\sigma}_T^2] = \sigma_T^2 + W.$$

Proof of (c):

$$\mathsf{V}(\tilde{\mu}_T) = \left(\frac{1}{k} - \frac{1}{K}\right)\sigma_T^2 + \frac{1}{k}W,$$

$$W = \frac{1}{K}\sum_{i=1}^{K} M_i^2\left(1 - \frac{m_i}{M_i}\right)\frac{\sigma_i^2}{m_i}.$$

(i) Homework: Show that $\mathsf{E}[\hat{W}] = W$ (using the same argument from (a)), where

$$\hat{W} = \frac{1}{k}\sum_{i\in S_c} M_i^2\left(1 - \frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}.$$

(ii) $\mathsf{E}[\hat{\sigma}_T^2] = \sigma_T^2 + W$.

(iii) Homework: Show that $\mathsf{E}\big[\mathsf{v}(\tilde{\mu}_T)\big] = \mathsf{V}(\tilde{\mu}_T)$, where

$$\mathsf{v}(\tilde{\mu}_T) = \left(\frac{1}{k} - \frac{1}{K}\right)\hat{\sigma}_T^2 + \frac{1}{K}\hat{W}.$$

Chapter 4

General Theory and Methods of Unequal Probability Sampling

Lecture 10
7th February

4.1 Sample Inclusion Probabilities

The first order and the second order inclusion probabilities:

$$\pi_i = \mathsf{P}(i \in S), \qquad \pi_{ij} = \mathsf{P}(i, j \in S).$$

- Inclusion probabilities are defined for all units in the population.
- Inclusion probabilities are usually only computed for units in the sample (to construct point and variance estimators).
- Inclusion probabilities are the fundamental tool for general theory of unequal probability sampling.
- A useful special case: $\pi_{ii} = \pi_i$:

$$\pi_{ii} = \mathsf{P}(i \in S, i \in S) = \mathsf{P}(i \in S) = \pi_i$$

Inclusion probabilities for some simple sampling designs

(1) SRSWOR (N, n):

$$\pi_i = \mathsf{P}(i \in S) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

$$\pi_{ij} = \mathsf{P}(i \in S, j \in S) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)},\ i \neq j.$$

(2) Stratified SRSWOR:

- $U = U_1 \cup \cdots \cup U_H$;
- $N = N_1 + \cdots + N_H$;
- $n = n_1 + \cdots + n_H$.

$$\pi_i = \frac{n_h}{N_h},\ i \in U_h.$$

$$\pi_{ij} = \begin{cases} \dfrac{n_h(n_h-1)}{N_h(N_h-1)}, & i,j \in U_h \\ \dfrac{n_h}{N_h} \cdot \dfrac{n_{h'}}{n_{h'}}, & i \in U_h,\ j \in U_{h'}. \end{cases}$$

(3) Single-stage cluster sampling with clusters selected by SRSWOR (S_c, K, k):

$$S\colon n = \sum_{i \in S_c} M_i.$$

$$\pi_i = \mathsf{P}(i \in S) = \frac{k}{K}.$$

$$\pi_{ij} = \begin{cases} \dfrac{k}{K}, & i,j \text{ in the same cluster,} \\ \dfrac{k(k-1)}{K(K-1)}, & i,j \text{ in two different clusters.} \end{cases}$$

(4) Two-stage cluster sampling with SRSWOR at both stages (S_c, K, k; S_i, M_i, m_i):

$$\begin{aligned} \pi_i &= \mathsf{P}(i \in S) \\ &= \mathsf{P}(i \in S_\ell, \ell \in S_c) \\ &= \mathsf{P}(\ell \in S_c)\,\mathsf{P}(i \in S_\ell \mid \ell \in S_c) \\ &= \frac{k}{K} \cdot \frac{m_\ell}{M_\ell}. \end{aligned}$$

$$\pi_{ij} = \begin{cases} \dfrac{k}{K} \cdot \dfrac{m_\ell(m_\ell-1)}{M_\ell(M_\ell-1)}, & i,j \text{ in cluster } \ell, \\ \dfrac{k(k-1)}{K(K-1)} \cdot \dfrac{m_\ell}{M_\ell} \cdot \dfrac{m_{\ell'}}{M_{\ell'}}, & i \text{ in cluster } \ell;\ j \text{ in cluster } \ell'. \end{cases}$$

4.1.1 Equalities related to inclusion probabilities

The sample indicator variables are the basic tool:

$$A_i = 1,\ i \in S, \qquad A_i = 0,\ i \notin S.$$

(1) For any sampling design,

$$\mathsf{E}[A_i] = \mathsf{P}(i \in S) = \pi_i, \qquad \mathsf{V}(A_i) = \pi_i(1-\pi_i).$$

(2) For any sampling design, with $i \neq j$:

$$\mathsf{Cov}(A_i, A_j) = \mathsf{E}[A_iA_j] - \mathsf{E}[A_i]\,\mathsf{E}[A_j] = \pi_{ij} - \pi_i\pi_j.$$

$$\mathsf{E}[A_iA_j] = \mathsf{P}(A_i = 1, A_j = 1) = \mathsf{P}(i \in S, j \in S) = \pi_{ij}.$$

If $i = j$, then

$$\mathsf{Cov}(A_i, A_i) = \mathsf{V}(A_i) = \pi_{ii} - \pi_i\pi_i = \pi_i(1-\pi_i).$$

(3) For any sampling design,

$$\sum_{i=1}^{N} A_i = n,$$

where n is the overall sample size (could be a random number under certain designs).

This leads to

$$\sum_{i=1}^{N} \pi_i = \mathsf{E}[n].$$

If the design has a fixed sample size, we have

$$\sum_{i=1}^{N} \pi_i = n.$$

(4) For any sampling design,

$$\sum_{j=1}^{N} A_i A_j = nA_i, \qquad \sum_{i=1}^{N}\sum_{j=1}^{N} A_i A_j^2 = n^2,$$

which leads to

$$\sum_{j=1}^{N} \pi_{ij} = \mathsf{E}[nA_i], \qquad \sum_{i=1}^{N}\sum_{j=1}^{N} \pi_{ij} = \mathsf{E}[n^2].$$

When n is random, $\mathsf{E}[nA_i] \neq \mathsf{E}[n]\,\mathsf{E}[A_i]$.

For sampling designs with a fixed sample size n, we have

$$\sum_{j=1}^{N} \pi_{ij} = n\pi_i, \qquad \sum_{i=1}^{N}\sum_{j=1}^{N} \pi_{ij} = n^2, \qquad \sum_{i\neq j}^{N}\sum_{j=1}^{N} \pi_{ij} = n(n-1).$$

Those equalities are useful to check computational errors in applications or simulation studies.

4.2 The Horvitz-Thompson Estimator

4.2.1 The general setting and the estimator

- The parameter of interest: $T_y = \sum_{i=1}^{N} y_i$.
- A general sampling design with $\pi_i > 0$ and $\pi_{ij} > 0$.
- The survey data: $\{y_i, i \in S\}$, or $\big\{(i, y_i), i \in S\big\}$ (with labels/ID).
- Information available from the survey design:

$$\{\pi_i, i \in S\}, \qquad \{\pi_{ij}, i, j \in S\}.$$

The Horvitz-Thompson (HT, 1952 JASA) estimator of T_y:

$$\hat{T}_{y\text{HT}} = \sum_{i\in S} \frac{y_i}{\pi_i} = \sum_{i\in S} d_i y_i,$$

where $d_i = 1/\pi_i$ are called the *basic design weights*.

Narain (1953) published a paper in an Indian journal with the same proposed estimator.

Notes on the Horvitz-Thompson estimator:

- The HT estimator is the most important fundamental piece of modern design-based sampling theory.
- The HT estimator was adopted (much later) by researchers on missing data problems as the *inverse probability weighted* (IPW) estimator.

- The value of the basic design weight $d_i = 1/\pi_i$ can be interpreted as: "*the number of units in the survey population which are represented by unit i selected for the survey sample*."

  In SRSWOR: $N = 100$, $n = 5$, we have

$$\pi_i = \frac{5}{100} = \frac{1}{20}, \qquad d_i = \frac{1}{\pi_i} = 20$$

4.2.2 Properties of the Horvitz-Thompson estimator

(1) The HT estimator is design unbiased for T_y.

$$\mathsf{E}[\hat{T}_{y\text{HT}}] = T_y \qquad \text{(Point Estimator)}.$$

(2) The theoretical variance of $\hat{T}_{y\text{HT}}$ is given by

$$\mathsf{V}(\hat{T}_{y\text{HT}}) = \sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} \qquad \text{(Theoretical Variance)}.$$

(3) An unbiased variance estimator for $\hat{T}_{y\text{HT}}$ is given by

$$\mathsf{v}(\hat{T}_{y\text{HT}}) = \sum_{i\in S}\sum_{j\in S}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} \qquad \text{(Variance estimator)}.$$

Also, $\mathsf{E}\big[\mathsf{v}(\hat{T}_{y\text{HT}})\big] = \mathsf{V}(\hat{T}_{y\text{HT}})$.

Sketch of Proofs: (1) and (2): Use indicators A_i.

$$\hat{T}_{y\text{HT}} = \sum_{i\in S}\frac{y_i}{\pi_i} = \sum_{i=1}^{N} A_i\frac{y_i}{\pi_i}.$$

$$\mathsf{E}[\hat{T}_{y\text{HT}}] = \sum_{i=1}^{N}\underbrace{\mathsf{E}[A_i]}_{\pi_i}\frac{y_i}{\pi_i} = \sum_{i=1}^{N} y_i = T_y.$$

$$\mathsf{V}(\hat{T}_{y\text{HT}}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\underbrace{\mathsf{Cov}(A_i, A_j)}_{\pi_{ij}-\pi_i\pi_j}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}.$$

(3): A more general question: How to estimate a quadratic quantity

$$Q = \sum_{i=1}^{N}\sum_{j=1}^{N} c(y_i, y_j)?$$

The answer:

$$\hat{Q} = \sum_{i\in S}\sum_{j\in S}\frac{c(y_i, y_j)}{\pi_{ij}}.$$

Homework: Show that $\mathsf{E}[\hat{Q}] = Q$.

$$\hat{Q} = \sum_{i=1}^{N}\sum_{j=1}^{N} A_iA_j\frac{c(y_i, y_j)}{\pi_{ij}}.$$

4.2.3 Estimation of the population mean μ_y and the Hájek estimator

(1) When the population size N is known, the Horvitz-Thompson estimator for the population mean μ_y is given by

$$\hat{\mu}_{y\text{HT}} = \frac{1}{N}\sum_{i\in S}\frac{y_i}{\pi_i} = \frac{1}{N}\hat{T}_{y\text{HT}}.$$

It is a design-unbiased estimator for μ_y with theoretical variance and variance estimator given respectively by

$$\mathsf{V}(\hat{\mu}_{y\text{HT}}) = \frac{1}{N^2}\mathsf{V}(\hat{T}_{y\text{HT}}), \qquad \mathsf{v}(\hat{\mu}_{y\text{HT}}) = \frac{1}{N^2}\mathsf{v}(\hat{T}_{y\text{HT}}).$$

(2) When the population size N is unknown, which is often the case for two-stage or multi-stage cluster sampling, an exactly design-unbiased estimator of μ_y might not be available.

A design-unbiased estimator for N:

$$\hat{N} = \sum_{i\in S}\frac{1}{\pi_i} = \sum_{i\in S} d_i.$$

(i) $\hat{N} = \sum_{i=1}^{N} A_i \frac{1}{\pi_i} \implies \mathsf{E}[\hat{N}] = N$.

(ii) $\hat{N} = \hat{T}_{y\text{HT}}$ when $y_i = 1$ for all i: $T_y = N$.

The population mean μ_y can be estimated by the Hájek estimator

$$\hat{\mu}_{yH} = \frac{1}{\hat{N}}\sum_{i\in S}\frac{y_i}{\pi_i} = \frac{1}{\hat{N}}\sum_{i\in S} d_i y_i = \frac{\sum_{i\in S} d_i y_i}{\sum_{i\in S} d_i}.$$

(Properties to be discussed in Chapter 5).

Lecture 11
9th February

4.2.4 The Yates-Grundy-Sen Variance Formula for the HT Estimator

For sampling designs with fixed sample size n, there are useful alternative expressions for $\mathsf{V}(\hat{T}_{y\text{HT}})$ and $\mathsf{v}(\hat{T}_{y\text{HT}})$.

(1) The theoretical variance

$$\mathsf{V}(\hat{T}_{y\text{HT}}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

(2) The variance estimator

$$\mathsf{v}(\hat{T}_{y\text{HT}}) = \frac{1}{2}\sum_{i\in S}\sum_{j\in S}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

Proof. Not required for tests; results useful.

4.3 PPS Sampling and the HT Estimator: An Optimal Strategy

4.3.1 A hypothetical scenario

Suppose that $y_i > 0$ for all i. All values $\{y_1, y_2, \ldots, y_N\}$ are known. We select a sample size S of fixed size n with $\pi_i \propto y_i$. We must have

$$\sum_{i=1}^{N}\pi_i = n, \qquad \pi_i = cy_i \implies c\sum_{i=1}^{N} y_i = n \implies c = \frac{n}{T_y}.$$

This leads to

$$\pi_i = n\frac{y_i}{T_y},\ i = 1, 2, \ldots, N.$$

The HT estimator of $T_y = \sum_{i=1}^N y_i$ is given by

$$\hat{T}_{y\mathrm{HT}} = \sum_{i\in S}\frac{y_i}{\pi_i} = T_y\sum_{i\in S}\frac{y_i}{ny_i} = T_y.$$

The HT estimator equals exactly the true value T_y (no error in estimation).

4.3.2 A practical scenario

There exists a variable z which is correlated to y and provides a measure for the “size” of the sampling units. Some examples:

- Expenditure survey: y — expenses; z — previous income;
- Agriculture survey: y — yield of a farm product; z — acreage of the farm;
- Business survey: y — total sales; z — number of workers;
- Multi-stage cluster sampling: z — cluster size (M_i).

We assume that

- The values $z_1, z_2, \ldots, z_N$ are available at the survey design stage;
- The value z_i provides a measure of the “size” for unit i, and $z_i > 0$ for all i;
- The size variable z and the study variable y are **positively correlated**.

The PPS (*the inclusion probability proportional to size*) sampling design, i.e., $\pi_i \propto z_i$:

$$\pi_i = n\frac{z_i}{T_z},\ i = 1, 2, \ldots, N.$$

For most of the discussions going forward, we assume that the size variable z is re-scaled such that

$$\sum_{i=1}^N z_i = 1.$$

Equal inclusion probabilities:

$$z_1 = z_2 = \cdots = z_N = \frac{1}{N}, \quad \pi_i = nz_i = \frac{n}{N} \qquad \text{(SRSWOR)}.$$

We have

$$\pi_i = nz_i,\ i = 1, 2, \ldots, N.$$

The re-scaled size variable must satisfy $z_i \le 1/n$ with the given n.

4.3.3 An optimal strategy

We assume that $z_i > 0$ for all i, and y_i and z_i are highly correlated.

We show that the PPS sampling design, combined with the Horvitz-Thompson estimator for the population total, is an “optimal strategy” in terms of the “anticipated variance.”

(1) The concept of *superpopulation* models

The finite population values $\{(y_i, z_i), i = 1, 2, \ldots, N\}$ are treated as fixed under the design-based framework.

Under the *superpopulation model* concept, $\{(y_i, z_i), i = 1, 2, \ldots, N\}$ are viewed as a random sample from a statistical model, denoted as ξ.

$$\text{Super population } \xi \to \text{Finite population } U \to \text{Survey sample } S.$$

Example ($\star$). Suppose that the finite population values $\{(y_i, z_i), i = 1, 2, \ldots, N\}$ follow a simple linear regression model (ξ),

$$y_i = \beta z_i + z_i \varepsilon_i, \; i = 1, 2, \ldots, N,$$

where the error terms are independent and satisfy

$$\mathsf{E}_\xi[\varepsilon_i] = 0, \qquad \mathsf{V}_\xi(\varepsilon_i) = \tau^2,$$

with $\mathsf{E}_\xi[\,\cdot\,]$ and $\mathsf{V}_\xi(\,\cdot\,)$ denoting expectation and variance under the model, ξ. We have

$$\mathsf{E}_\xi[y_i \mid z_i] = \beta z_i, \qquad \mathsf{V}_\xi(y_i \mid z_i) = z_i^2 \tau^2.$$

A semi-parametric model specified through the first two conditional moments.

(2) An optimal estimator of T_y

A well-known (negative) result (Godambe, 1955): The minimum variance linear unbiased estimator does not exist among the general Godambe-class of linear estimator under the design-based framework.

A useful concept for optimality: The anticipated variance under a superpopulation mode (ξ).

$$\mathsf{E}_\xi\big[\mathsf{V}_\mathrm{p}(\hat{T}_{y\mathrm{HT}})\big],$$

where $\mathsf{V}_\mathrm{p}(\hat{T}_{y\mathrm{HT}})$ is the design-based variance (p: probability sampling design).

- Under the probability sampling design, p: (y_i, z_i) are fixed, but S is random.
- Under the super population model, ξ: y_i is random given z_i and the sampling selection becomes irrelevant (the survey design is non-informative: the superpopulation model holds for the sample).

An important (positive) result (Godambe, 1955): The anticipated variance of the HT estimator under model ($\star$) is minimized under the PPS sampling design with $\pi_i \propto z_i$ and fixed sample size n.

n is fixed; The Yates-Grundy-Sen variance formula for the HT estimator:

$$\mathsf{V}_\mathrm{p}(\hat{T}_{y\mathrm{HT}}) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

$$\mathsf{E}_\xi\big[\mathsf{V}_\mathrm{p}(\hat{T}_{y\mathrm{HT}})\big] = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij})\, \mathsf{E}_\xi\left[\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2\right].$$

Need to find

$$\mathsf{E}_\xi\left[\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2\right], \; i \neq j.$$

Major steps for the proof:

- Under the model ($\star$) and for $i \neq j$, we have

$$\mathsf{E}_\xi\left[\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2\right] = \beta^2 \left(\frac{z_i}{\pi_i} - \frac{z_j}{\pi_j}\right)^2 + \tau^2 \left(\frac{z_i^2}{\pi_i^2} + \frac{z_j^2}{\pi_j^2}\right).$$

$$\mathsf{E}_\xi[Y^2] = \big(\mathsf{E}_\xi[Y]\big)^2 + \mathsf{V}_\xi(Y).$$

From earlier, we know that:

$$\mathsf{E}_\xi[y_i \mid z_i] = \beta z_i, \qquad \mathsf{V}_\xi(y_i \mid z_i) = z_i^2 \tau^2.$$

Therefore,

$$\mathsf{E}_\xi\left[\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right] = \beta\left(\frac{z_i}{\pi_i} - \frac{z_j}{\pi_j}\right).$$

$$\mathsf{V}_\xi\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right) = \tau^2\left(\frac{z_i^2}{\pi_i^2} + \frac{z_j^2}{\pi_j^2}\right).$$

- Identities under any fixed sample size design:

$$\sum_{i=1}^{N}\sum_{j\neq i, j=1}^{N}(\pi_i\pi_j - \pi_{ij})\frac{z_i^2}{\pi_i^2} = \sum_{i=1}^{N}\pi_i(1-\pi_i)\frac{z_i^2}{\pi_i^2},$$

$$\sum_{i=1}^{N}\sum_{j\neq i, j=1}^{N}(\pi_i\pi_j - \pi_{ij})\frac{z_j^2}{\pi_j^2} = \sum_{i=1}^{N}\pi_i(1-\pi_i)\frac{z_i^2}{\pi_i^2}$$

$$\text{First equation you get by: } \sum_{i=1}^{N}\left(\sum_{j\neq i, j=1}^{N}(\pi_i\pi_j - \pi_{ij})\right)\frac{z_i^2}{\pi_i^2}.$$

$$\sum_{i\neq j, j=1}^{N}(\pi_i\pi_i - \pi_{ij}) = \pi_i(n - \pi_i) - (n\pi_i - \pi_{ii}) = \pi_i(1-\pi_i).$$

$$\pi_i(1-\pi_i)\frac{z_i^2}{\pi_i^2} = \frac{z_i^2}{\pi_i^2} - z_i^2.$$

- Use the Yates-Grundy-Sen variance formula to obtain

$$\mathsf{E}_\xi\left[\mathsf{V}_{\mathrm{p}}(\hat{T}_{y\mathrm{HT}})\right] = \frac{\beta^2}{2}D_1 + \tau^2 D_2 - \tau^2 D_3,$$

where

$$D_1 = \sum_{i=1}^{N}\sum_{j\neq i, j=1}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{z_i}{\pi_i} - \frac{z_j}{\pi_j}\right)^2,$$

$$D_2 = \sum_{i=1}^{N}\frac{z_i^2}{\pi_i},$$

$$D_3 = \sum_{i=1}^{N} z_i^2.$$

- Under the constraint $\sum_{i=1}^{N}\pi_i = n$, D_2 is minimized when $\pi_i \propto z_i$.

$$\mathcal{L}(\pi_1, \ldots, \pi_N) = \sum_{i=1}^{N}\frac{z_i^2}{\pi_i} + \lambda\left(\sum_{i=1}^{N}\pi_i - n\right).$$

$$\frac{\partial \mathcal{L}}{\partial \pi_i} = -\frac{z_i^2}{\pi_i^2} + \lambda = 0.$$

$$\pi_i^2 = \frac{1}{\lambda}z_i^2 \implies \pi_i \propto z_i.$$

- The anticipated variance $\mathsf{E}_\xi\left[\mathsf{V}_{\mathrm{p}}(\hat{T}_{y\mathrm{HT}})\right]$ is minimized when $\pi_i \propto z_i$.

  - D_2 is minimized when $\pi_i \propto z_i$.
  - D_3 does not depend on π_i.
  - D_1:

$$D_1 = \mathsf{V}_{\mathrm{p}}(\hat{T}_{z\mathrm{HT}}) \geq 0.$$

$$D_1 = 0, \text{ if } \pi_i \propto z_i.$$

- The PPS sampling design combined with the HT estimator is an **optimal strategy**. An important aspect of the optimal strategy is the assumption that the response variable y and the size variable z is positively correlated.

4.4 PPS Sampling Procedures

Lecture 12

14th February

Lecture 13

16th February