Sampling Theory and Practice

STAT 454[1]

Winter 2022 (1221)[2]

Cameron Roopnarine[3] Changbao Wu[4]

18th January 2022

[1] STAT 454≡ STAT 854

[2] Online Course until January 27th, 2022

[3] LaTeXer

[4] Instructor

Contents

Chapter 1

Review of Basic Concepts in Survey Sampling

WEEK 1
5th to 7th January

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

- *Target population*: Who is the group to be studied?
- *Sample data* (variables to be measured): What information should we collect?
- *Sampling frame(s)*: From what can we select individuals to be surveyed?
- *Sampling methods/procedures*: How do we select individuals to be surveyed?
- *Method of data collection*: What method(s) can we use to collect data?
  - Examples: face-to-face, telephone, mail, questionnaire.
- *Statistical analysis*: How do we use the data to draw conclusions?

Survey Populations

- **Target population**: The set of all units covered by the main objective of the study.

  Target Population of Example 1.1

  All students who received a formal degree from Waterloo between 2016 and 2019.

- **Frame population**: The set of all units covered by the sampling frame(s).

  Sampling Frame of Example 1.1

  The list of personal email addresses of students who graduated between 2016 and 2019.

- **Sampled/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

A general **population** is $U = \{1, 2, \ldots, N\}$, 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 for $h = 1, 2, \ldots, H$. 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.

- **Stratified sampling**: Randomly select 20 students from each high school.
- **Two-stage cluster sampling**: Randomly select 5 high schools from the list of 15 schools, and then randomly select 60 students from each of the 5 selected schools.
- **Stratified two-stage cluster sampling**: 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.

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.

- *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.
- *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.
- *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.

WEEK 2
10th to 14th January

Survey Samples

A **survey sample** S, is a subset of the population $U = \{1, 2, \ldots, N\}$.

The sample size $n = |S|$ is the number of units in the sample (a set of n "unordered" units):

$$S = \{i_1, i_2, \ldots, i_n\}.$$

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

Survey Sample

If $N = 10$ and $n = 3$, then $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, and some possible samples are:

- $S = \{7, 4, 9\} = \{i_1, i_2, i_3\}$, or
- $S = \{1, 2, 3\}$.

Non-probability Samples versus Probability Samples

Non-probability samples are selected by subjective or any convenient methods.

We will list some examples of non-probability sampling.

- **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.

Remarks:

- The most serious issue with non-probability survey samples is that the sample is *biased*. A sample is **biased** if it has 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

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

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

Example 1.4. If $N = 3$ and $U = \{1, 2, 3\}$, then we have seven possible candidate samples.

- $n = 1$: $S_1 = \{1\}$, $S_2 = \{2\}$, $S_3 = \{3\}$.
- $n = 2$: $S_4 = \{1, 2\}$, $S_5 = \{1, 3\}$, $S_6 = \{2, 3\}$.
- $n = 3$: $S_7 = \{1, 2, 3\}$ (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 $\sum_{\{S:S\in\Omega\}} \mathcal{P}(S) = 1$.
To select a sample under design 2, we first generate a random number R from the uniform distribution over $[0, 1]$. If $0 \leq R \leq 1/3$, S_4 is selected; if $1/3 < R \leq 2/3$, S_5 is selected; if $2/3 < R \leq 1$, S_6 is selected.

A sampling design $\mathcal{P}$ has a fixed sample size n if $\mathcal{P}(S) = 0$ for any S such that $|S| \neq n$. That is, the probability measure is defined over the set

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

Discrete Random Number Generator for Sample Selection (Problem 1.4)

Let $X \sim f(x)$ such that $p_i = f(x_i) = \mathbb{P}(X = x_i), i = 1, 2, \ldots$.

- *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 $b_{j-1} < R \leq b_j$.

Show that $X^\star$ has the same distribution as X.

Solution: Not completed.

Survey Variables

- 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 Parameters

- **Population totals**:
$$T_y = \sum_{i=1}^{N} y_i \quad \text{and} \quad T_{\boldsymbol{x}} = \sum_{i=1}^{N} \boldsymbol{x}_i.$$
- **Population means**:
$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i \quad \text{and} \quad \mu_{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i.$$
- **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 is when y is an indicator variable:

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

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

We can define our population parameters in terms of the new indicator variable.

- 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:

$$\begin{aligned}\sigma_y^2 &= \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - N\mu_y^2\right) \\ &= \frac{1}{N-1}(M - NP^2) \\ &= \frac{N}{N-1}P(1-P) \\ &\approx P(1-P) \qquad \text{if } N \text{ is large.}\end{aligned}$$

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)

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 is

$$\binom{N}{n} = \frac{N(N-1)\cdots(N-n+1)}{n!}.$$

The **probability measure** for the sampling design is

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

Remark: $\mathcal{P}(S)$ is a theoretical tool. That is, $\mathcal{P}(S)$ cannot be used to select a sample in practice. For example, if $N = 1000$ and $n = 3$, then

$$\binom{N}{n} = \frac{1000 \times 999 \times 998}{6} = 166\,167\,000.$$

Sampling Scheme/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.

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.

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

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

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

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

Proof: Not given.

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

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

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 $\mathbb{E}[X_i] = \mu$, $\mathbb{V}(X_i) = \sigma^2$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

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

However, in SRSWOR, note that

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

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

Theorem 2.2. 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 $\mu_y = N^{-1}\sum_{i=1}^N y_i$, that is,
$$\mathbb{E}[\bar{y}] = \mu_y.$$

(b) The design-based variance of $\bar{y}$ under SRSWOR is given by
$$\mathbb{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_y^2}{n},$$
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**.

(c) An unbiased variance estimator for $\bar{y}$ is given by
$$v(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n},$$
which satisfies $\mathbb{E}[v(\bar{y})] = \mathbb{V}(\bar{y})$, where
$$\mathbb{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_y^2}{n}.$$
Furthermore, the sample variance s_y^2 is an unbiased estimator for the population variance σ_y^2 under SRSWOR, i.e., $\boxed{\mathbb{E}[s_y^2] = \sigma_y^2}$.

Proof.

(a)

- *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.
  $$\begin{aligned}
  \mathbb{E}[\bar{y}] &= \sum(\text{value})(\text{prob}) \\
  &= \sum_{S} \bar{y}(S)\mathcal{P}(S) \\
  &= \sum_{S:|S|=n} \frac{1}{n}\sum_{i\in S} y_i \frac{1}{\binom{N}{n}} \\
  &= \frac{1}{n}\frac{1}{\binom{N}{n}} \sum_{\{S:|S|=n\}} \sum_{i\in S} y_i \\
  &= \frac{1}{n}\frac{1}{\binom{N}{n}} \sum_{i=1}^{N} t_i y_i \\
  &= \frac{1}{N}\sum_{i=1}^{N} y_i \\
  &= \mu_y,
  \end{aligned}$$
  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,
  $$\mathbb{E}[\bar{y}] = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} Z_k\right] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[Z_k].$$
  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,
  $$\mathbb{E}[Z_k] = \sum_{i=1}^{N} y_i \frac{1}{N} = \mu_y.$$

- *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.

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

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

- To find $\mathsf{Cov}(A_i, A_j)$, we first need $\mathbb{E}[A_i A_j]$:

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

Now, we find $\mathsf{Cov}(A_i, A_j)$:

$$\begin{aligned}
\mathsf{Cov}(A_i, A_j) &= \mathbb{E}[A_i A_j] - \mathbb{E}[A_i] \, \mathbb{E}[A_j] \\
&= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = \frac{n(n-N)}{N^2(N-1)}.
\end{aligned}$$

Therefore,

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

(c) Homework.

Summary

- 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: $\mathbb{E}[\bar{y}] = \mu_y$.
- $\mathbb{V}(\bar{y}) = \left(1-\frac{n}{N}\right)\frac{\sigma_y^2}{n}$ is the theoretical variance of $\bar{y}$ and is a fixed, but unknown quantity depending on the population variance σ_y^2.
- $v(\bar{y}) = \left(1-\frac{n}{N}\right)\frac{s_y^2}{n}$ 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
sam=sample(N,n,replace=T)
sam
N=100
n=4
sam=sample(N,n)
sam
sam=sample(N,n,replace=T)
sam
```