Course on Survey Sampling

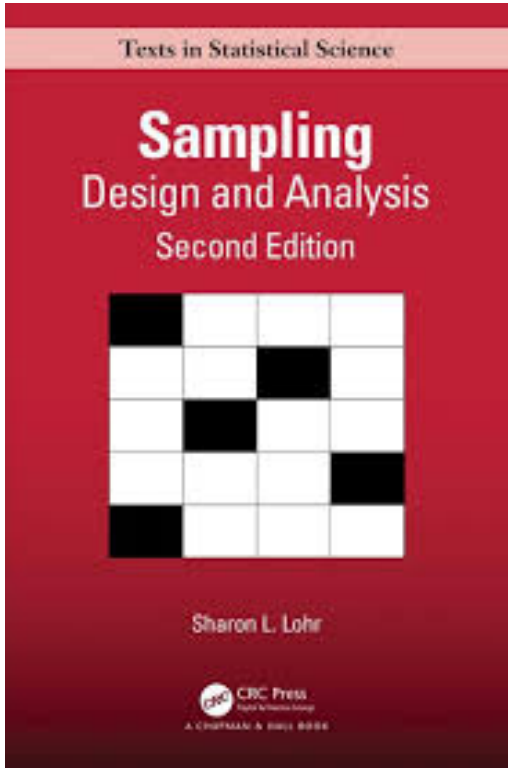
Melike Oğuz–Alper[1] & **Tatsiana Pekarskaya**[1]

[1]Statistics Norway, Methods Section

Short-term mission from SSB, Norway, to SSSU, Ukraine,
19-28 October 2020

Reference textbook

- **Lohr, Sharon L. (2019)**, *Sampling: design and analysis*, 2nd edition, Taylor& Francis Group, CRC Press.



Outline

Session 1: Introduction

Anders Nicolai Kiær

- **Aim**: To estimate totals, means, proportions, etc. of some *variables* in the population of interest based on a *sample* from it
- **Sampling**: Observations only on a subset of the population
  - **Reduced cost**, **faster** production, **greater** scope
- Anders Nicolai Kiær pioneered the use of sampling
  - Founded **official statistics** in Norway
  - **First director** of Statistics Norway established in 1876
  - A large paper: "The **representative method** of statistical surveys" (1897)

Anders Nicolai Kiær (1838-1919)

Survey sampling in Norway

- *Survey of personal income and poverty*
  - Two-stage cluster sampling with *Probabilities Proportional to Size (PPS)*
  - Purposive sampling of a set of administrative districts
  - Men selected from the 1981 census sheets by 5-year age intervals starting from 17 and by the first letter of their names (A,B,L,M and N) ($n = 11\,427$)
  - Double weights to the rural sample
- *Survey of living conditions*
  - Sample of houses constructed ensuring the same distribution of industry groups as in the census at district level
  - Half of the large streets and 1/10 of the houses
  - Quarter of the other streets and 1/5 of the houses

Representativeness

- Many critics in ISI 1895: 'No calculations can be made without observing the **whole**'
- Later studies increased doubts about the **representativeness** of his method
- ISI Commission 1925: 'Results should be generalised if the sample is **sufficiently representative** of the totality'
  - *Random selection*: equal probabilities of selection
  - *Purposive selection*: selected to nearly the same characteristics as the totality
- Neyman (1934): '**random samples** are preferable'
  - Probability samples largely accepted by National Statistical Offices

1936 Literacy digest voting survey

- 1936 Literacy digest magazine voting survey (Presidential election): $n = 10\,000\,000$ persons from telephone directories and automobile registration lists
  - Response rate: 24%
  - Predicted Landon to win with 54%, but Roosevelt won with 60%

Remark

The respondent set, consisting mainly of car and telephone owners and magazine's subscribers, is highly *unrepresentative* of the voting population

Women and love

- Shere Hite's book *Women and Love: A Cultural Revolution in Progress* (1987)
  - 84% of women are "not satisfied emotionally with their relationships" (p. 804)
  - 70% of all women "married five or more years are having sex outside of their marriages" (p. 856)
  - 95% of women "report forms of emotional and psychological harassment from men with whom they are in love relationships" (p. 810)
  - 84% of women report forms of condescension from the men in their love relationships (p. 809)
- 4.5% of 100 000 questionnaires mailed returned

Remark

The respondent women is *not representative* of women in Unites States

Survey sampling

Survey sampling

As the basic scientific method,

- How to select the sample?: *sampling*
- How to measure the quantities of interest?: *survey*

Terminology

- **Observation (elementary) unit** Object on which measurements are taken

  e.g. person, household, business, farm, animal, etc.

- **Target population** A set of observation units we want to study

  e.g. all adults 15+ residing in Ukraine, all persons eligible for voting, all farmers in a given region, etc.

- **Sample** A subset of a population

- **Sampled (study) population** The population from which the sample was actually taken

- **Sampling unit** A unit that can be sampled

  e.g. household, address block, residential telephone number, school, farm, etc.

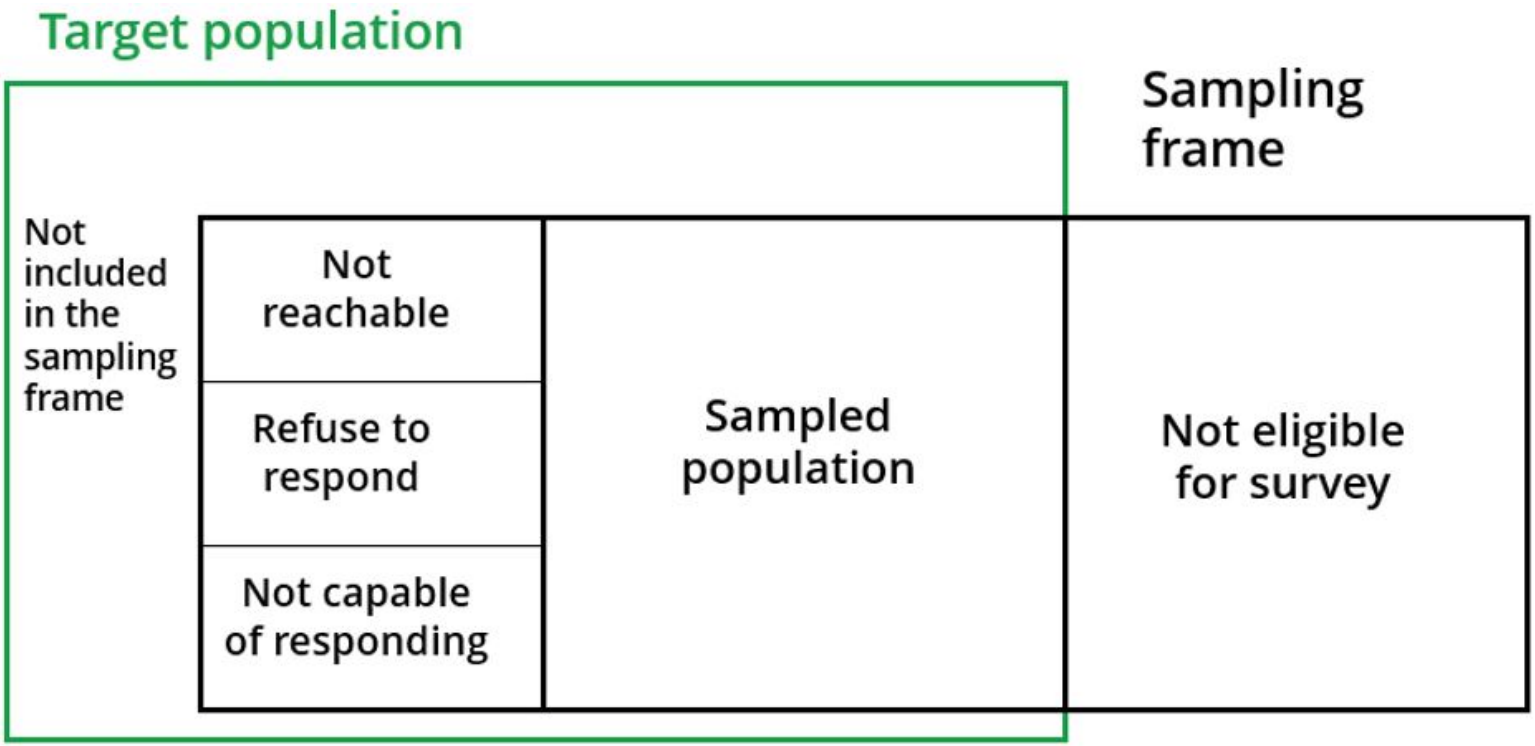
Terminology (cont.)

- **Sampling frame** A list, map, or any device providing access to the elements in the population. The sample is taken from the sampling frame
  - telephone surveys: list of residential telephone numbers,
  - person or household surveys: list of postcode addresses,
  - agricultural surveys: list of all farms or map of areas containing farms, etc.
  - business surveys: list of enterprises,
  - list of schools, medical centres, etc.

Ideal survey

The sampled population will be identical to the target population

Terminology (cont.)



Source: Lohr (2010, p.4)

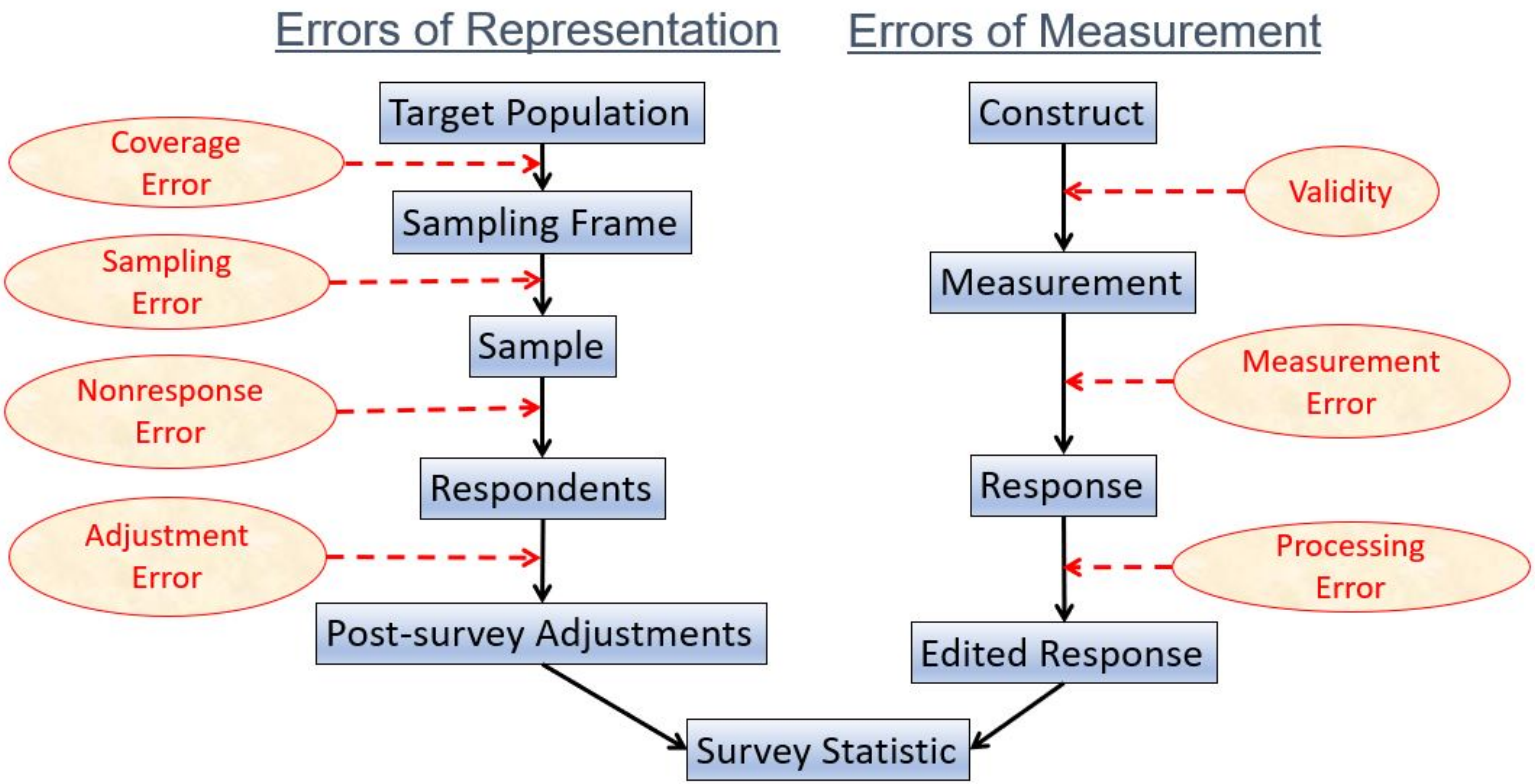
Terminology (cont.)

- **Target sample** sample we want to select
- **Achieved (effective) sample** sample we actually observed; that is, sample after removal of *ineligible* units and *nonresponse*

Example

Norwegian European Health Interview Survey (EHIS) 2015: target sample of 14 000 individuals; effective sample consisted of $8\,164 = 14\,000 - 252\,(\textit{ineligible}) - 5\,584\,(\textit{nonresponse})$ individuals

Total Survey Error (TSE) framework



Source: Groves et al. (2009, p.48)

Errors of representation

- Errors of representation may result in **selection bias**

> Selection bias occurs when some population units are sampled at a different rate then intended by the investigator

1. **Coverage error** (Sampled (study) population $\neq$ target population): effect of difference between target and sampled (study) populations
   - **under-coverage**: failing to include all of the target population in the study population
   - **over-coverage**: including units in the study population that are not in the target population (*not-eligible* (*out-of-scope*) units)
     - Does not lead to selection bias: waste of resources, loss of efficiency

Errors of representation (cont.)

2. **Non-probability sampling**:
   - *Convenience sampling*: selecting units easiest to select or most likely to respond (*voluntary* sample)
   - *Purposive (judgement) sampling*: deliberately or purposively selecting a sample of units with certain characteristics: causes under-coverage
   - *Quota sampling*: non-probability sampling of units until reaching pre-specified *quotas* within strata (age, sex, occupation, ethnicity, etc.): often used in market research
3. **Nonresponse**: effect of difference between target sample and achieved sample: nonresponse probability is generally unknown

Remark

Selection bias can cause sample estimates to be invalid

Measurement error

Measurement error occurs when the observed values are not equal to the true values

- May not be possible to measure the truth via an instrument: hypothetical truth
- Not willing to tell the truth
- Misunderstanding of questions
- Memory issues: *telescoping errors*
- Interviewer effect, etc.

Measurement bias

Occurs when there exists a tendency between the observed and the true values in one direction

Questionnaire design

Careful *questionnaire design* may reduce the measurement error

- **Testing** questions in advance
- Keeping it **simple** and **clear**
- Motivational, cognitive, contextual aspects
- **Order** or routing if multiple questions
- *Split-questionnaire* design, **randomised response**

Sampling vs. non-sampling errors

Sampling error: Error due to not observing the entire population

- Often, different samples produce different estimates
- Depends on sampling design
- Can be assessed based on the selected sample

Non-sampling error = under-coverage + nonresponse error + measurement error

Remark

Non-sampling errors may cause serious bias in the estimation

Session 2: Probability sampling

What is a "good" sample?

- A *representative* sample?

  e.g. proportion of people with certain characteristics (gender, age, religion, ethnicity, etc.) in the sample $\approx$ proportion of people with the same characteristics in the population

Remark

Difficulties in defining a *generally representative sample* (Neyman 1934)

- Neyman (1934) defined what should be termed a *representative method of sampling*
  - **probability sampling**

Definition

Definition

In **probability sampling**, each unit in the population has a known positive probability of selection to be included in the sample which is selected with respect to a *randomisation mechanism*

An objective inference framework

- An **objective inference framework**
  - Sample estimates can be generalised to make inference for population quantities
  - Avoids selection bias at the sample selection stage
- Allows **unbiased** or **approximately unbiased** estimation of population quantities
- Allows estimation of **standard errors**
- Ensures specification of **sample size** for a desired level of **precision**

Notation

- U: the **finite population** or **universe** of N units
- $i \in \{1, \ldots, N\}$: index for elementary units in U
- s: a subset of U consisting of n units
- $\pi_i = \Pr(i \in s)$: **inclusion probability**: probability of unit i being included in the sample
  - The π_i are known, and $\pi_i > 0$ for all $i \in U$
- Let Y be the **variable of interest** taking values $y_1, \ldots, y_N$
- θ: the **parameter** of interest, which is a function of $y_1, \ldots, y_N$
  - Total, mean, proportion, ratio, cumulative distribution function, etc.

Parameter of interest

- **Total**: $t = \sum_{i \in U} y_i$

  e.g. total number of employees, total turnover

- **Mean**: $\bar{Y} = \sum_{i \in U} y_i / N$

  e.g. average income

- **Proportion**: $p = \sum_{i \in U} y_i / N$, where $y_i \in \{0, 1\}$

  e.g. unemployment/employment rate

- **Ratio of totals/means**: $R = t_y / t_x$ or $R = \bar{Y} / \bar{X}$
- **Distribution function**: $F(y) = \sum_{i \in U} \delta\{y_i \leq y\} / N$, where $\delta\{a\} = 1$ if a is true, and $\delta\{a\} = 0$ otherwise
  - used in statistics based on quantiles: median income, poverty rate, etc.

Sampling design

The pair $\{\Omega, p(s)\}$, where

- Ω is the *sample space*; that is, the set of all possible samples
- $p(s)$ is probability of selecting s.

$$\sum_{s \in \Omega} p(s) = 1, \quad \text{for any } s \in \Omega$$

Sampling design (cont.)

Example

Let $U = \{1, 2, 3, 4\}$ and $\Omega = \{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}$

- **Design 1**: $\Pr(s_d) = 1/6$ for all $s_d \in \Omega$, with $d = 1, \ldots, 6$;
- **Design 2**: $\Pr(s_1) = 1/3$, $\Pr(s_2) = 1/6$, $\Pr(s_6) = 1/2$, and $\Pr(s_d) = 0$, for $d = 3, 4, 5$

Inclusion probabilities

$$\pi_i = \Pr(i \in s) = \sum_{s \in \Omega | i \in s} p(s)$$

Example

(cont.)

- **Design 1**:

$$\pi_1 = \Pr(s_1) + \Pr(s_2) + \Pr(s_3) = 1/2,$$
$$\pi_2 = \Pr(s_1) + \Pr(s_4) + \Pr(s_5) = 1/2,$$
$$\pi_3 = \Pr(s_2) + \Pr(s_4) + \Pr(s_6) = 1/2,$$
$$\pi_4 = \Pr(s_3) + \Pr(s_5) + \Pr(s_6) = 1/2.$$

Inclusion probabilities (cont.)

Example

(cont.)

- **Design 2**:

$$\pi_1 = \Pr(s_1) + \Pr(s_2) + \Pr(s_3) = 1/2,$$
$$\pi_2 = \Pr(s_1) + \Pr(s_4) + \Pr(s_5) = 1/3,$$
$$\pi_3 = \Pr(s_2) + \Pr(s_4) + \Pr(s_6) = 2/3,$$
$$\pi_4 = \Pr(s_3) + \Pr(s_5) + \Pr(s_6) = 1/2.$$

Fixed sample size designs

For any fixed sample-size designs,

$$\sum_{i \in U} \pi_i = n, \text{ (sample size)}$$

Example

(cont.)

- **Design 1**: $\sum_{i \in U} \pi_i = 4 * 1/2 = 2$
- **Design 2**: $\sum_{i \in U} \pi_i = 1/2 + 1/3 + 2/3 + 1/2 = 2$

Sampling strategy

- Let $\widehat{\theta}(s)$ be a **sample estimator** of θ, which is a function of $y_1, \ldots, y_n$, for $i \in s$

  e.g. $\bar{y}_s = \sum_{i \in s} y_i / n$ (**sample mean**); $\hat{t}_s = \sum_{i \in s} d_i y_i$, where $d_i = \pi_i^{-1}$ are **design (sampling) weights**

- Values that $\widehat{\theta}(s)$ takes are called **sample estimates**

Sampling strategy = sampling design + estimation method

Aim:

Choose a sampling strategy providing an *accuracy* of the estimators as high as possible given the resources

Sampling distribution

Definition

The distribution of the values of a sample estimator takes over all possible samples taken from a population

Example

Let $U = \{1, 2, 3, 4\}$, with y_i values $y_1 = 1$, $y_2 = 2$, $y_3 = 3$, $y_4 = 4$, and $\Omega = \{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}$ with $\Pr(s_d) = 1/6$ for $d = 1, \ldots, 6$. Consider $\hat{t} = N\bar{y}$ as an estimator for the population total Y.

Sampling distribution (cont.)

Example

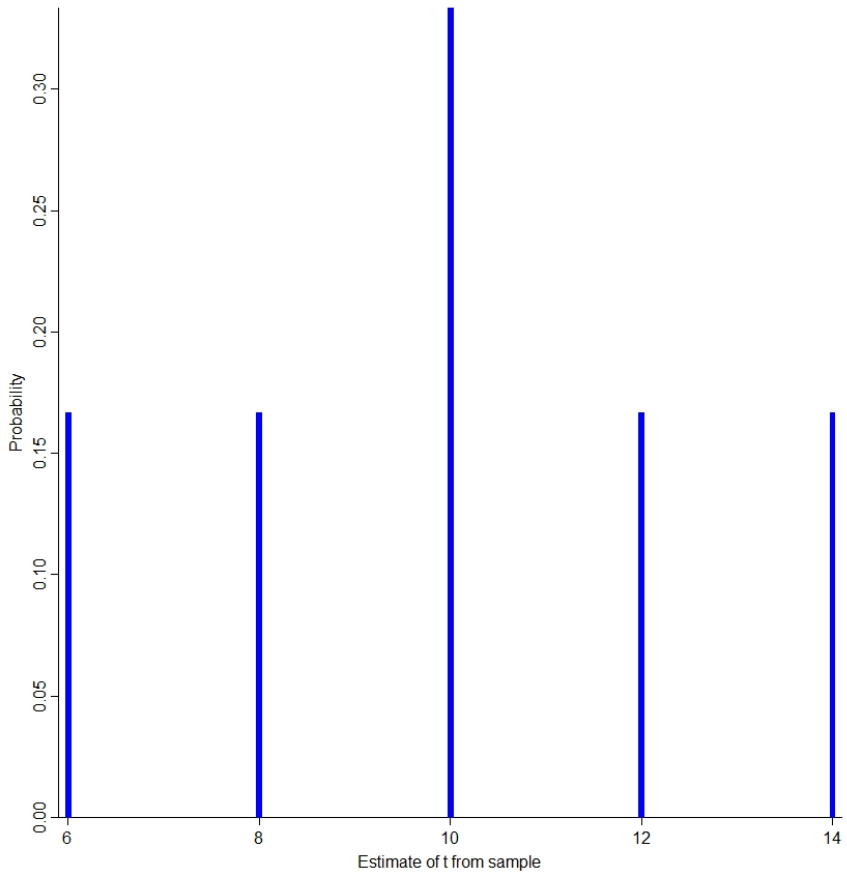
(cont.)

s	$\Pr(s)$	$\hat{t}_s$
$(1,2)$	$1/6$	6
$(1,3)$	$1/6$	8
$(1,4)$	$1/6$	10
$(2,3)$	$1/6$	10
$(2,4)$	$1/6$	12
$(3,4)$	$1/6$	14

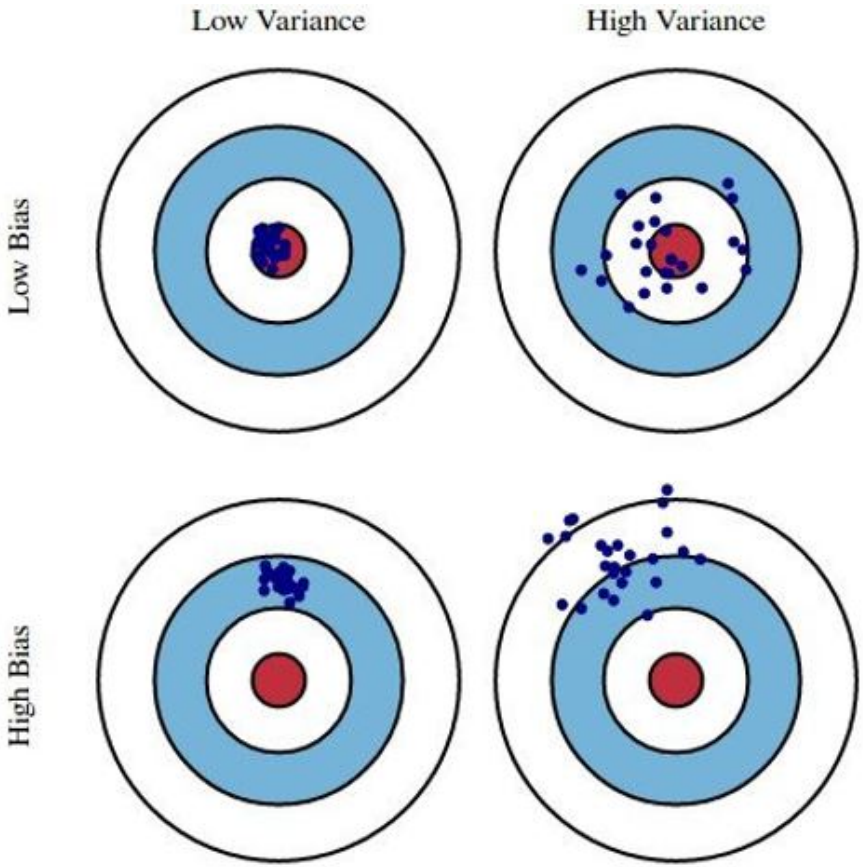
$$\Pr(\hat{t} = k) = \sum_{s|\hat{t}_s=k} \Pr(s)$$

Sampling distribution (cont.)

- $\Pr(\hat{t} = k) = 1/6$, for $k = 6, 8, 12, 14$, and $\Pr(\hat{t} = 10) = 2/6$.



Which one most "accurate"?



Bias of an estimator

Bias of $\widehat{\theta}$: $\text{Bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta$, where

$$E(\widehat{\theta}) = \sum_{s \in \Omega} \widehat{\theta}(s) \Pr(s) = \sum_{k} k \Pr(\widehat{\theta} = k)$$

- An estimator $\widehat{\theta}$ is **unbiased** for θ if $E(\widehat{\theta}) = \theta$

Example

(cont.) $$E(\hat{t}) = \frac{1}{6}(6 + 8 + 12 + 14) + \frac{2}{6}(10) = 10 = t$$

- $\hat{t}$ is **unbiased** for t

Variance of an estimator

Variance of $(\widehat{\theta})$:

$$V(\widehat{\theta}) = E\big([\widehat{\theta} - E(\widehat{\theta})]^2\big) = \sum_{s \in \Omega} \Pr(s)[\widehat{\theta}(s) - E(\widehat{\theta})]^2$$

- An estimator $\widehat{\theta}$ is **precise** if $V(\widehat{\theta})$ is small

Example

(cont.)

$$\begin{aligned} V(\hat{t}) &= \frac{1}{6}(6-40)^2 + \cdots + \frac{1}{6}(14-40)^2 + \frac{2}{6}(10-40)^2 \\ &= 906.67 \end{aligned}$$

Accuracy of an estimator

Mean squared error of $(\widehat{\theta})$:

$$\begin{aligned} \text{MSE}(\widehat{\theta}) &= E\big([\widehat{\theta} - \theta]^2\big) \\ &= V(\widehat{\theta}) + \big[\text{Bias}(\widehat{\theta})\big]^2 \end{aligned}$$

- An estimator $\widehat{\theta}$ is **accurate** if $\text{MSE}(\widehat{\theta})$ is small

Remarks

- An estimator with big (unknown) bias is useless in practice no matter how small variance it has
- If the bias is small compared to the variance it might be worthwhile to consider
- If *bias/variance* is large, the confidence interval may have a coverage significantly differ from the confidence level

bias/$\sqrt{variance}$	Coverage (95% confidence level)
0.00	0.9500
0.05	0.9497
0.10	0.9489
0.30	0.9396
0.50	0.9210
1.0+	0.8300

Session 3: Simple random sampling (SRS)

Sampling design

Example

Let $U = \{1, 2, 3, 4\}$ of size $N = 4$; all possible subsets by size $0 < n \leq N$:

n	Sample subset	Number
1	$\{(1), (2), (3), (4)\}$	4
2	$\{(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)\}$	6
3	$\{(1,2,3), (1,2,4), (1,3,4), (2,3,4)\}$	4
4	$\{(1,2,3,4)\}$	1

- Under **SRSWOR**: all possible subsets of size $0 < n < N$ have the same probability of selection

Sampling design (cont.)

- There are $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ possible samples of size n
- $\Pr(s) = \frac{1}{\binom{N}{n}}$ for any $s \subset U$ of size n

Sampling design: assigns same probability of being included in the sample to every possible subset $s \subset U$ of size n

Inclusion probability and sampling weight

Inclusion probability

$$\begin{aligned} \pi_i = \Pr(i \in s) &= \frac{\text{number of samples containing } i}{\text{number of all possible samples}} \\ &= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} \\ &= \frac{n}{N} \quad (\textbf{sampling fraction}) \end{aligned}$$

Sampling weight

$$d_i = \frac{1}{\pi_i} = \frac{N}{n}$$

Sample mean

- **Sample mean**:

$$\bar{y}_s \overset{\text{or}}{=} \bar{y} = \frac{1}{n} \sum_{i \in s} y_i$$

is an **unbiased** estimator of the *population mean*

$$\bar{Y} = \frac{1}{N} \sum_{i \in U} y_i$$

over all possible samples of a given size selected with SRSWOR from the given finite population

Sample variance

- Define the *population variance* as

$$S^2 = \frac{1}{N-1}\sum_{i \in U}(y_i - \bar{Y})^2$$

- For fixed n, as $N \to \infty$, S^2 *approximately equal* to the variance of the y_i values in the finite population: $\sigma^2 = \frac{1}{N}\sum_{i \in U}(y_i - \bar{Y})^2$
- Under SRSWOR with sample size n,

$$s^2 = \frac{1}{n-1}\sum_{i \in s}(y_i - \bar{y})^2$$

is an **unbiased** estimator for S^2

Sampling variance of $\bar{y}$

- **Sampling variance** of $\bar{y}$:

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

$f = n/N \Rightarrow$ **sampling fraction**

  - $f = n/N \to 0$ as $N \to \infty$ for fixed n
  - n has usually bigger effect on $\widehat{\mathrm{SE}}(\bar{y})$ than f does

- *Estimated* sampling variance:

$$\widehat{V}(\bar{y}) = (1 - f)\frac{s^2}{n}$$

Example

Example

Consider three SRSWOR designs given below

N	n	f	$V(\bar{y})$
4 000	400	0.1	$0.00225\,S^2$
300	30	0.1	$0.03\,S^2$
300 000 000	3 000	0.001	$0.000333\,S^2$

Source: Exercise 10 in Lohr (2019, p.63)

- Sample size n has higher effect on the *precision* than that of sampling fraction f

Standard error (SE)

- **Standard error** of $\bar{y}$:

$$\mathrm{SE}(\bar{y}) = \sqrt{(1-f)\frac{S^2}{n}}$$

- *Estimated* standard error:

$$\widehat{\mathrm{SE}}(\bar{y}) = \sqrt{(1-f)\frac{s^2}{n}}$$

  - A *measure* of **sampling error**

Coefficient of variation (CV)

- **Coefficient of Variation (CV)** of $\bar{y}$:

$$\mathrm{CV}(\bar{y}) = \frac{\mathrm{SE}(\bar{y})}{\bar{Y}} = \sqrt{1-f}\,\frac{S}{\sqrt{n}\bar{Y}}$$

- *Estimated* CV of $\bar{y}$:

$$\widehat{\mathrm{CV}}(\bar{y}) = \frac{\widehat{\mathrm{SE}}(\bar{y})}{\bar{y}} = \sqrt{1-f}\,\frac{s}{\sqrt{n}\bar{y}}$$

  - Free of *scale* of measurement
  - More *stable* over repeated surveys

Example

Example

Let $\{y_1, y_2, y_3, y_4\} = \{16, 10, 30, 25\}$, $n = 3$, $s = \{1, 2, 4\}$

$$
\begin{aligned}
\bar{y} &= (16 + 10 + 25)/3 = 17 \\
s^2 &= \frac{1}{3-1}[(16-17)^2 + (10-17)^2 + (25-17)^2] = 57 \\
\widehat{V}(\bar{y}) &= \left(1 - \frac{3}{4}\right)\frac{57}{3} = 4.75 \\
\widehat{\mathrm{SE}}(\bar{y}) &= \sqrt{4.75} = 2.18 \\
\widehat{\mathrm{CV}}(\bar{y}) &= \frac{2.18}{17} = 0.13 = 13\%
\end{aligned}
$$

Estimation of population total

- **Population total**: $Y = \sum_{i \in U} y_i = N\bar{Y}$
- The **unbiased** estimator of Y:

$$\hat{t} = N\bar{y} = \sum_{i \in s} d_i y_i = \frac{N}{n} \sum_{i \in s} y_i$$

- $V(\hat{t}) = N^2 V(\bar{y}) = N^2(1-f)\frac{S^2}{n}$
- *Estimated* **variance** of $\hat{t}$:

$$\widehat{V}(\hat{t}) = N^2 \widehat{V}(\bar{y}) = N^2(1-f)\frac{s^2}{n}$$

Estimation of population total (cont.)

- *Estimated* **standard error** of $\hat{t}$:

$$\widehat{\mathrm{SE}}(\hat{t}) = \sqrt{N(1-f)\frac{s^2}{n}}$$

- The **CV** of $\hat{t}$ **same** as that of $\bar{y}$:

$$\mathrm{CV}(\hat{t}) = \frac{\sqrt{N^2 V(\bar{y})}}{N\bar{y}} = \mathrm{CV}(\bar{y})$$

Estimation of population proportion

- **Proportion** of units with certain characteristics:
  $p = \sum_{i \in U} = y_i/N$, with $y_i = 1$ if unit i has characteristics, and $y_i = 0$ otherwise
- **Sample proportion**: **unbiased** for p

$$\hat{p} = \frac{\sum_{i \in s} y_i}{n}$$

- $S^2 = Np(1-p)/(N-1)$; $S^2 \approx p(1-p)$ for large N
- $s^2 = n\hat{p}(1-\hat{p})/(n-1)$; $s^2 \approx \hat{p}(1-\hat{p})$ for large n

$$\widehat{\mathrm{SE}}(\hat{p}) = \sqrt{(1-f)\frac{\hat{p}(1-\hat{p})}{n-1}}$$

CV of $\hat{p}$

$$\mathrm{CV}(\hat{p}) = \frac{1}{p}\sqrt{\frac{(1-f)Np(1-p)}{(N-1)n}} \approx \sqrt{\frac{(1-f)(1-p)}{np}}$$

- The standard error of $\hat{p}$ increases until $p = 0.5$ and then decreases
- Coefficient of variation **decreases monotonically** in p
- May not be sensible to use CV for **small** proportions as they will yield **large CV**

Sampling design

- Selecting n *independent* samples of size 1 from population U with probabilities $1/N$
- Same units may appear in the sample **several** times
- **Sampling distribution**: $\Pr(s) = 1/N$
- Sample mean $\bar{y} = \sum_{i=1}^{n} y_i/n$, where n is the number of draws and y_i is the value observed in the ith draw
  - $\bar{y}$ is **unbiased** for $\bar{Y}$

Sampling variance of $\bar{y}$

- **Sampling variance**:

$$V(\bar{y}) = \frac{\sigma^2}{n},$$

where σ^2 is the **population variance** defined by

$$V(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2 = \sigma^2$$

- *Estimated* sampling variance:

$$\widehat{V}(\bar{y}) = \frac{s^2}{n}$$

- s^2 is **unbiased** for σ^2 under SRSWR

Comparison: SRSWOR vs SRSWR

- $V_{srswor}(\bar{y}) \leq V_{wr}(\bar{y})$:

$$(1-f)\frac{S^2}{n} \leq \frac{\sigma^2}{n} = \frac{(N-1)S^2}{Nn}$$

- **Finite population correction (fpc)**: $1-f$
- If $(1-f) \to 1$, or equivalently $f \to 0$, SRSWOR and SRSWR have the **same efficiency** since $S^2 \to \sigma^2$ as $(N-1)/N \to 1$ with large N

Confidence interval

- **Central Limit Theorem (CLT)**: The *sampling distribution* of $\bar{y}$ is *approximately normal* if $n, N, N-n$ are all sufficiently large: $\bar{y} \sim N\big(\bar{Y}, \widehat{\mathrm{SE}}(\bar{y})\big)$
- 95% **Confidence Interval (CI)**:

$$[\bar{y} - 1.96\,\widehat{\mathrm{SE}}(\bar{y}), \bar{y} + 1.96\,\widehat{\mathrm{SE}}(\bar{y})]$$

- **Interpretation**: The true population mean is expected to be included 95% of the time over all possible samples if the CLT holds
- The probability that a given CI(s) includes the population mean is either 0 or 1 as $\bar{Y}$ is fixed

Example

Example

Let $\{y_1, y_2, y_3, y_4\} = \{16, 10, 30, 25\}$, $\bar{Y} = 20.25$, $n = 3$, $\delta(s) = 1$ if $\bar{Y} \in CI(s)$, and $\delta(s) = 0$ otherwise

s	$\Pr(s)$	95% CI(s)	$\delta(s)$
(1,2,3)	1/4	[12.74, 24.47]	1
(1,2,4)	1/4	[12.73, 21.27]	1
(1,3,4)	1/4	[19.65, 27.68]	1
(2,3,4)	1/4	[15.78, 27.56]	1

- The **true coverage**: $\frac{1}{4}(1+1+1+1) = 100\%$
- **Asymptotically**, as $n, N \to \infty$, coverage $\to$ *nominal level* if the **CLT** holds

Precision constraints

- Possible **precision constraints**:

$$\mathrm{SE}(\bar{y}) \leq c, \text{ or } z_{\alpha/2}\mathrm{SE}(\bar{y}) \leq c, \text{ or } \mathrm{CV}(\bar{y}) \leq c$$

- **Margin-of-Error (MoE)**: $z_{\alpha/2}\mathrm{SE}(\bar{y})$ (e.g. $z_{\alpha/2} = 1.96$ for 95% confidence level): *half-length of the CI*

Example

Find n under SRSWOR satisfying $1.96\,\mathrm{SE}(\bar{y}) \leq c$
$(\iff (1.96)^2 V(\bar{y}) \leq c^2)$

$$(1.96)^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n} = (1.96)^2\left(\frac{1}{n} - \frac{1}{N}\right)S^2 \leq c^2$$

Solving this for n gives: $n \geq \frac{(1.96)^2 S^2}{c^2 + \frac{(1.96)^2 S^2}{N}}$

Precision constraints (cont.)

- For **SRSWR**: $n \geq \frac{(1.96)^2 S^2}{c^2}$
  - This can also be used for **SRSWOR** if f is small which leads to $(1-f) \to 1$

Remark

A precision criteria based on CV leads to huge sample sizes for small proportions

e.g. Under SRSWR or with fpc $= (1-f) \to 1$, $n \geq \mathrm{CV}(\hat{p})/c^2$ ($\mathrm{CV}(\hat{p}) = \sqrt{(1-p)/p}$ for $n = 1$).

c	p	$\mathrm{CV}(\hat{p})$	n
0.05	0.5	1	400
0.05	0.1	3	3 600
0.05	0.01	9.9	39 600
0.05	0.001	31.6	399 600

When fpc matters?

Example

Proportion is of interest. Let $z_{\alpha/2} = 1.96$. Sample sizes for a margin-of-error $c = 4\%$. As p is unknown, one can use $p = 0.05$ which provides the maximum possible value for S^2, i.e. $S^2 = 0.25$. For example, for City A:

$n_{srswor} \geq 1.96^2(0.25)/\left(0.04^2 + 1.96^2(0.25)/5\,000\right) = 535.91$

$n_{srswr} \geq 1.96^2(0.25)/(0.04^2) = 600.25$

City	N	n_{srswor}	n_{srswr}
A	5 000	536	601
B	60 000	595	601
C	1 500	429	601
D	1 000 000	600	601
E	150 000	598	601

Source: Adapted from Exercise 19 in Lohr (2019, p.65)

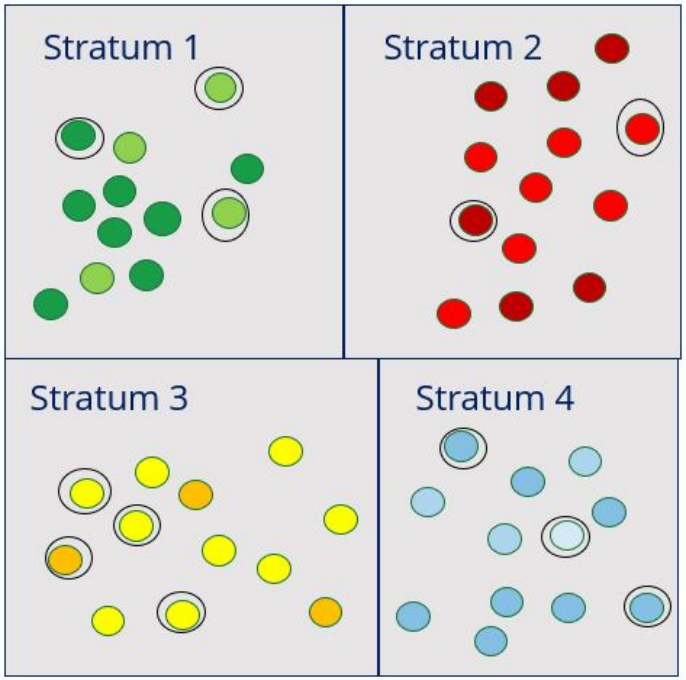
Session 4: Stratified sampling

Stratified sampling

- Population is divided into H **mutually exclusive** and **exhaustive** sub-populations called **strata**: $h = 1, \ldots, H$

  e.g. region, age, sex, economical activities, etc.

- Samples are selected **independently** between strata
- U_h: population for stratum h, such that $U = U_1 \cup \cdots \cup U_H$
- N_h: size of U_h, such that $N = N_1 + \cdots + N_H$



Why stratified sampling?

- To avoid very **unlucky samples** which are possible to get with SRSWOR
- **Precision requirements** for certain sub-populations or domains: regional level precision
- Different **data collection** methods for different strata: e.g. web survey, telephone interview
- Different **sampling frames** for different strata due to different data sources
- Different **sampling and estimation** methods for different strata
- **Representation** of certain groups in the sample
- To improve **efficiency**

Sampling design

- A sample s_h of size n_h selected from U_h of size N_h with **SRSWOR**
- $s = \bigcup_{h=1}^{H} s_h$ and $n = \sum_{h=1}^{H} n_h$
- **Inclusion probabilities**:

$$\pi_i = \frac{n_h}{N_h}, \quad \text{for } i \in U_h$$

- **Sampling weights**:

$$d_i = \pi_i^{-1} = \frac{N_h}{n_h}, \quad \text{for } i \in s_h$$

Population quantities

- **Within-stratum**:

  Total: $t_h = \sum_{i \in U_h} y_{hi}$

  Mean: $\bar{Y}_h = \frac{t_h}{N_h}$

  Variance: $S_h^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_{hi} - \bar{Y}_h)^2$

- **Whole population**:

  Total: $t = \sum_{h=1}^{H} t_h$

  Mean: $\bar{Y} = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{Y}_h = \sum_{h=1}^{H} W_h \bar{Y}_h$, with $W_h = N_h / N$

  Variance: $S^2 = \frac{1}{N-1} \sum_{h=1}^{H} \sum_{i \in U_h} (y_{hi} - \bar{Y})^2$

Stratified SRSWOR

- **Variance = within-stratum variance + between-strata variance**

$$S^2 = \frac{1}{N-1}\sum_{h=1}^{H}(N_h - 1)S_h^2 + \frac{1}{N-1}\sum_{h=1}^{H} N_h(\bar{Y}_h - \bar{Y})^2$$

Aim

More alike within stratum and more different between strata

Estimation of within-stratum population quantities

- Population quantities (mean, total, variance, etc.) are estimated **separately** for each stratum and then the results are **combined** to get estimates for the whole population
- **Within-stratum** estimators:

Mean: $\bar{y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_{hi}$

Total: $\hat{t}_h = N_h \bar{y}_h$

Variance: $s_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} (y_{hi} - \bar{y}_h)^2$

- All point and variance estimators are **unbiased**
- Formulas for **stratified SRSWR** can be obtained analogously

Estimation of whole population quantities

- Estimators for the **whole population**:

Mean: $\hat{\bar{Y}}_{str} = \sum_{h=1}^{H} W_h \bar{y}_h$, with $W_h = N_h/N$

Sampling variance:
$\widehat{V}(\hat{\bar{Y}}_{str}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{s_h^2}{n_h}, \quad f_h = \frac{n_h}{N_H}$

Total: $\hat{t}_{str} = \sum_{h=1}^{H} N_h \bar{y}_h$

Sampling variance: $\widehat{V}(\hat{t}_{str}) = N^2 \widehat{V}(\hat{\bar{Y}}_{str})$

- All point and variance estimators are **unbiased**
- Formulas for **stratified SRSWR** can be obtained analogously

Proportional allocation

- Suppose **total sample size** n and **strata** given
- How to allocate n to strata?
- **Proportional allocation**:

$$n_h = n\frac{N_h}{N}$$

- Stratum sample size proportional to stratum population size: $n_h \propto N_h = n\,W_h$
- **Inclusion probabilities**:

$$\frac{n_h}{N_h} = \frac{n}{N} \quad \rightarrow \textbf{constant}$$

Proportional allocation (cont.)

- **Equal Probabilities SElection Method (EPSEM)** design: if π_i are equal for all $i \in U$
  - Or, **self-weighting** design: if $d_i = \pi_i^{-1}$ are equal for all $i \in U$
- **SRSWOR** and **stratified SRSWOR** with proportional allocation are **EPSEM** designs
- As a result of EPSEM design:

$$\hat{\bar{Y}}_{str} = \sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \bar{y}$$

Proportional allocation (cont.)

- As a result of EPSEM design:

$$V(\hat{\bar{Y}}_{str}) = (1-f)\frac{\sum_{h=1}^{H} W_h S_h^2}{n}, \quad f = \frac{n}{N}$$

- Compared to $\bar{y}$ under SRSWOR: S^2 replaced by $\sum_{h=1}^{H} W_h S_h^2$
- Using **variance decomposition**:
  $S^2 = \sum_{h=1}^{H}(N_h - 1)S_h^2/(N-1) +$ **between-strata variance**

$$\sum_{h=1}^{H} W_h S_h^2 < S^2, \quad \text{for large } N_h \text{ and } \bar{Y}_h \neq \bar{Y}$$

$\Rightarrow$ Stratified SRSWOR with proportional allocation is **more efficient** than SRSWOR

Optimal allocation

- **Budget** an important issue in surveys
- Let $C = c_0 + \sum_{h=1}^{H} c_h n_h$ be the **total cost** function

Aim

Minimise $V(\hat{\bar{Y}}_{str}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) S_h^2 / n_h$ **subject to**

$$\sum_{h=1}^{H} c_h n_h = C - c_0 \quad \text{and} \quad \sum_{h=1}^{H} n_h = n$$

$$n_h = n \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} W_h S_h / \sqrt{c_h}}$$

Optimal allocation (cont.)

- **Neyman allocation**: if $c_h = c$

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$$

- **Proportional allocation**: if $c_h = c$ and S_h are all equal: $n_h = nN_h/N$
- For **fixed** n and **equal** costs:

$$V_{opt}(\hat{\bar{Y}}_{str}) \leq V_{prop}(\hat{\bar{Y}}_{str}) \leq V_{srswor}(\bar{y})$$

- The more **homogeneity** within the strata, or **variability** between the strata, the more improvement in **precision**

Neyman allocation: practical issues

1. S_h often **unknown**
   - Estimate from **pilot** survey or **previous** surveys if any
   - Use **auxiliary** variable, X, if available
     - Use S_{xh} as proxies to S_{yh}, or
     - Predict y_{hi} using x_{hi} for all $i \in U$ and then estimate S_{yh}
     - Use an **explicit** model for the variance
   - If the target statistics **proportion**: proportional allocation **nearly** Neyman allocation unless p_h very small or large

     e.g. $S_h = \sqrt{p_h(1-p_h)} = 0.5$ if $p_h = 0.5$, or $S_h = 0.46$ if $p_h = 0.3$

Neyman allocation: practical issues (cont.)

2. Usually **many variables**
   - Create a **linear combination** of the key variables and then apply the allocation method on this new variable
   - Choose **few key** variables and then take an average of all the allocations
   - Use **proportional** allocation instead (especially in household surveys) as it is still more efficient than SRSWOR
3. $n_h > N_h \Rightarrow \pi_{hi} > 1$ for some strata
   - Take all units (i.e. $n_h = N_h$) in such strata, so-called **certainity** or **take-all** or **self-representing** stratum, and then allocate $n - N_h$ to the remaining strata

Neyman allocation: practical issues (cont.)

4. $n_h = 1$ for some strata
   - Set a **minimum** sample size; e.g. $n_{hmin} = 2$ or 3
   - Put such strata together with larger similar strata for the purpose of variance estimation
5. If strata **boundaries** not given for **continuous variables** (e.g. turnover, age, number of employees) if any used
   - Use **algorithms** forming strata based on efficiency criteria: **e.g. Cum-SqRoot-f rule** (Dalenius and Hodges (1959)), Lavallée and Hidiroglou (1988)'s method

Post-stratification

- Stratifying variables may **not be known** throughout the population, or **available** at the time of sampling
  - Stratum variables are **collected** in the sample: **e.g.** household size, ethnicity, education
  - Stratum **population sizes** are available for **estimation**

Definition

Post-stratification: May use SRSWOR as the sampling design, but form the strata later in the sample, and estimate as if these were the design strata

Post-stratified weights

- Post-stratification involves **reweighting**
- **Sampling (design or base) weight**: inverse of the selection probabilities, i.e., $d_i = \pi_i^{-1}$
  - Under SRSWOR: $d_i = N/n$, as $\pi_i = n/N$, yielding in fact, $\hat{t} = \sum_{i \in s} d_i y_i = N\bar{y}$, $\bar{y} = \sum_{i \in s} d_i y_i / N$
- **Post-stratified weights**: **modified** sampling weights

$$w_{hi} = \frac{N_h}{n_h}$$

  for **realised** n_h
- Use post-stratified weights in **estimation**

Post-stratified estimators

- **Post-stratified estimators**:

$$\hat{\bar{Y}}_{pst} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in s_h} w_{hi}y_{hi} = \sum_{h=1}^{H} W_h \bar{y}_h$$

$$\hat{t}_{pst} = \sum_{h=1}^{H}\sum_{i\in s_h} w_{hi}y_{hi} = \sum_{h=1}^{H} N_h \bar{y}_h$$

- Under SRSWOR, conditional on **realised** $\mathbf{n} = \{n_1, \ldots, n_H\}$,

$$V(\hat{\bar{Y}}_{pst} \mid \mathbf{n}) = \sum_{h=1}^{H} W_h^2 (1 - f_h)\frac{S_h^2}{n_h} = V(\hat{\bar{Y}}_{str})$$

Remarks

- Under stratified SRSWOR, n_h **fixed**
- **Unconditional** sampling variance $\approx$ stratified SRSWOR with proportional allocation + a **lower-order** term vanishing with large n_h
- For **sufficiently large** n_h: **Strategy** := SRSWOR + $\hat{\bar{Y}}_{pst}$ **nearly as efficient as** **Strategy** := Stratified SRSWOR + $\hat{\bar{Y}}_{str}$
- Replace S_h^2 with s_h^2 to estimate the variance $V(\hat{\bar{Y}}_{pst} \mid \mathbf{n})$

Session 5: Estimation with auxiliary information

Ratio estimation

- **Auxiliary information** often used in surveys: **e.g. stratification**
- Used in **estimation** to increase **precision** of the estimators of population means and totals
- Ratio may be the **parameter of interest**

e.g. Percentage of total time of watching football matches:

$$\frac{\text{total time of watching football mathches}}{\text{total time of watching TV}}$$

e.g. Unemployment rate:

$$\frac{\text{total number of unemployed people}}{\text{total number of unemployed and employed people}}$$

Example: Laplace's population estimation

Example

Population size estimation (Laplace, 1814)

- Population size in a sample of 30 communes: $t_{ys} = 2\,037\,615$
- Annual registered births in the same communes specified to be: $t_{xs} = 71\,866.33$
- Annual registered births for whole France: $t_x = 1\,000\,000$
- Population size for whole France estimated by: $\hat{t}_r = t_x\, t_{ys}/t_{xs}$

Source: Lohr (2019, p.117)

Notation

- X: **Auxiliary** variable
- t_x: population **total** of X, i.e., $\sum_{i \in U} x_i$
- $\bar{X}$: population **mean** of X, i.e., $\bar{X} = t_x / N$
- $\bar{x}$: sample mean, i.e., $\bar{x} = \sum_{i \in s} x_i / n$
- $\hat{t}_x$: sample estimate for t_x, i.e., $\hat{t}_x = N \bar{x}$
- $B = \frac{t_y}{t_x} = \frac{\bar{Y}}{\bar{X}}$: population **ratio**

Ratio estimators

- Population **ratio** estimated by:

$$\widehat{B} = \frac{\hat{t}_y}{\hat{t}_x} = \frac{\bar{y}}{\bar{x}}$$

- **Ratio estimator** of population **mean**:

$$\hat{\bar{Y}}_{rt} = \frac{\bar{y}}{\bar{x}} \bar{X}$$

- **Ratio estimator** of population **total**:

$$\hat{t}_r = \frac{\bar{y}}{\bar{x}} t_x$$

- Note that $\bar{X}$ and t_x as well as x_i values in the sample must be **known** to estimate mean and total

Variance of the ratio estimators

- **Variance** of $\widehat{B}$:

$$V(\widehat{B}) = (1-f)\frac{1}{\bar{X}^2}\frac{S_e^2}{n}, \quad S_e^2 = \frac{1}{N-1}\sum_{i\in U} e_i^2,\ e_i = y_i - Bx_i$$

- **Variance** of $\hat{\bar{Y}}_{rt}$:

$$V(\hat{\bar{Y}}_{rt}) = \bar{X}^2 V(\widehat{B}) = (1-f)\frac{S_e^2}{n}$$

- **Variance** of $\hat{t}_r$:

$$V(\hat{t}_r) = t_x^2 V(\widehat{B}) = N^2(1-f)\frac{S_e^2}{n}$$

Estimated variances of the ratio estimators

- *Estimated* **variance** of $\widehat{B}$:

$$\widehat{V}(\widehat{B}) = (1-f)\frac{1}{\bar{x}^2}\frac{s_e^2}{n}, \quad s_e^2 = \frac{1}{n-1}\sum_{i\in s}\hat{e}_i^2,\ \hat{e}_i = y_i - \widehat{B}x_i$$

- *Estimated* **variance** of $\hat{\bar{Y}}_{rt}$:

$$\widehat{V}(\hat{\bar{Y}}_{rt}) = \bar{X}^2\widehat{V}(\widehat{B}) = (1-f)\left(\frac{\bar{X}}{\bar{x}}\right)^2\frac{s_e^2}{n}$$

- *Estimated* **variance** of $\hat{t}_r$:

$$\widehat{V}(\hat{t}_r) = N^2\widehat{V}(\hat{\bar{Y}}_{rt}) = N^2(1-f)\left(\frac{\bar{X}}{\bar{x}}\right)^2\frac{s_e^2}{n}$$

Remarks

- In **large samples**, $\bar{X}/\bar{x} \to 1$, thus the variances of $\hat{\bar{Y}}_{rt}$ and $\hat{t}_r$ can be *alternatively* estimated without $\bar{X}/\bar{x}$ in the formulas above
- With **sufficiently large** sample size, approximate 95% CIs constructed by:

$$\widehat{B} \pm 1.96\,\widehat{\mathrm{SE}}(\widehat{B}), \quad \hat{\bar{Y}}_{rt} \pm 1.96\,\widehat{\mathrm{SE}}(\hat{\bar{Y}}_{rt}), \quad \hat{t}_r \pm 1.96\,\widehat{\mathrm{SE}}(\hat{t}_r)$$

- **Ratio estimator** has a smaller variance $\iff S_e^2 < S_y^2$
- Improvement in **precision** if X is highly correlated with Y.
- Let ρ be the **correlation coefficient** between X and Y, which measures how Y and X **vary jointly**

Remarks (cont.)

$$V(\hat{\bar{Y}}_{rt}) \leq V(\bar{y}) \iff \rho \geq \frac{\mathrm{CV}(\bar{X})}{2\mathrm{CV}(\bar{Y})}$$

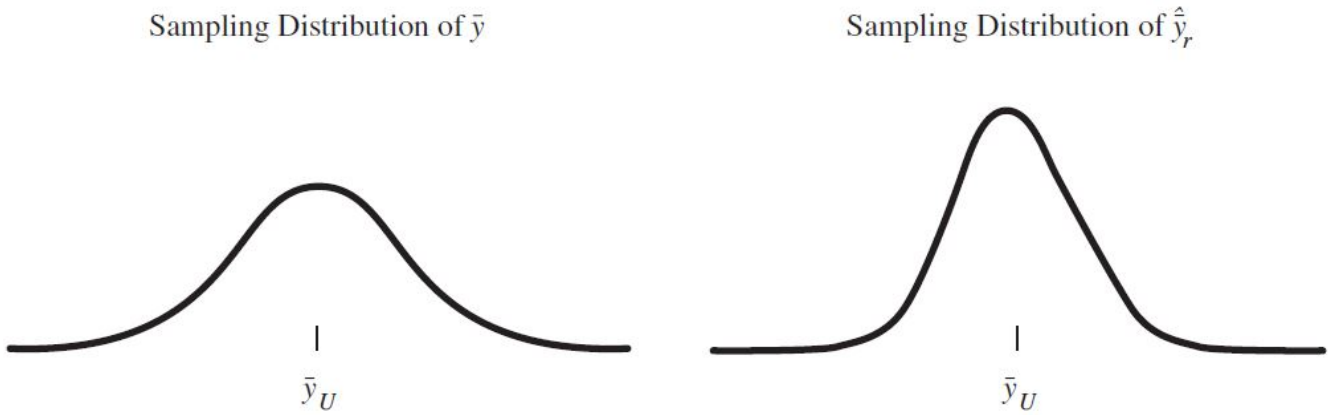
- If the CVs are **approximately equal**, the ratio estimator is **more efficient** when $\rho \geq 1/2$
- Gain in precision is paid off with **bias**

Remark

Ratio estimator is not unbiased. However, the bias vanishes with large n.

Remarks (cont.)

- For **large samples** the sampling distributions of simple **sample mean** $\bar{y}$ and the **ratio estimator** of mean, $\hat{\bar{Y}}_{rt}$, will be approximately **normal**
- Given high positive correlation between X and Y, the **bias** and **variance** of the two estimators:



Source: Lohr (2019, p.124)

Ratio estimation using weight adjustments

- The estimated population total using **sampling weights**: $\hat{t} = \sum_{i \in s} d_i y_i$
- The **ratio estimation** of the total:

$$\hat{t}_r = \frac{t_x}{\hat{t}_x} \hat{t}_y = \frac{t_x}{\hat{t}_x} \sum_{i \in s} d_i y_i = \sum_{i \in s} d_i g_i y_i, \quad g_i = \frac{t_x}{\hat{t}_x}$$

- The **base weights** d_i are adjusted by g_i
- The **weight adjustments** g_i **calibrate** the estimates on the X variable:

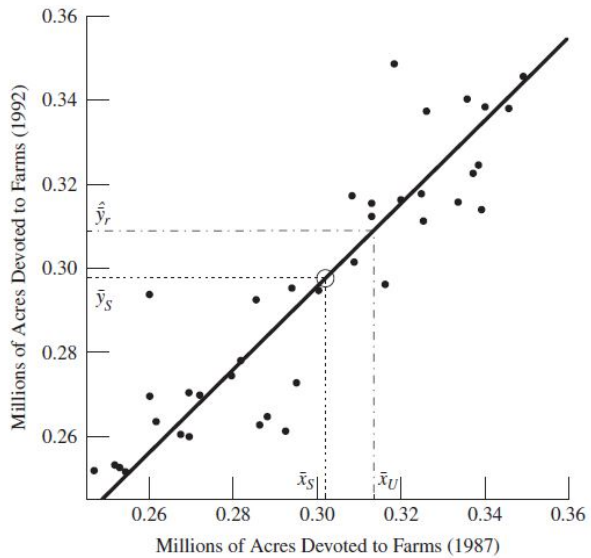
$$\sum_{i \in s} d_i g_i x_i = \frac{t_x}{\hat{t}_x} \sum_{i \in s} d_i x_i = t_x$$

- g_i: **calibration factors**; $w_i = d_i g_i$: **calibration weights** (**adjusted weights**)

"Working model" for ratio estimation

- **Ratio estimation** works well if a **straight line** through the origin summarises the relationship between y_i and x_i (i.e., $y_i = 0$ whenever $x_i = 0$), and the variance of y_i about the line is proportional to x_i

$$y_i = Bx_i + e_i, \quad V(e_i) = x_i\sigma^2$$

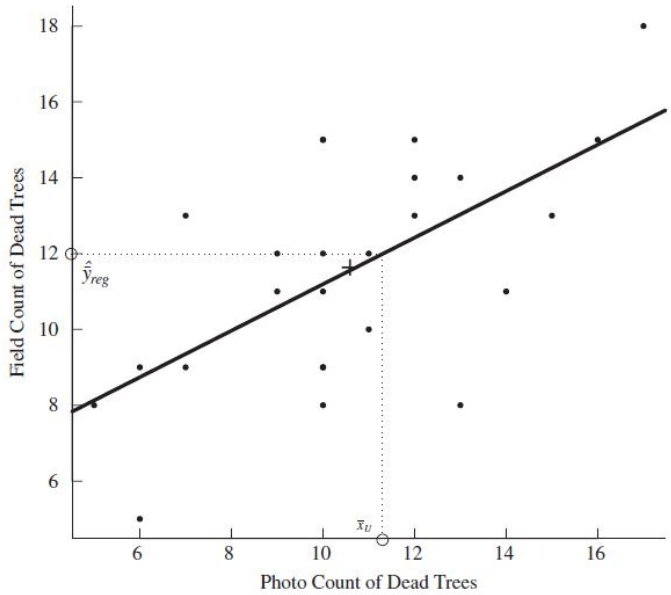


Source: Lohr (2019, p.123)

"Working model" for regression estimation

- **Regression estimation** works well if data appear to be evenly scattered about a **straight line** not passing through the origin

$$y_i = B_0 + B_1 x_i + e_i, \quad V(e_i) = \sigma^2$$



Source: Lohr (2019, p.140)

Regression estimator of population mean

- **Auxiliary values**: $(1, x_i)$ for each $i \in U$
- Sample estimates of B_0 and B_1:

$$\widehat{B}_1 = \frac{\sum_{i \in s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in s}(x_i - \bar{x})^2},$$

$$\widehat{B}_0 = \bar{y} - \widehat{B}_1 \bar{x}$$

- **Regression estimator** of population **mean**:

$$\begin{aligned} \hat{\bar{Y}}_{reg} &= \widehat{B}_0 + \widehat{B}_1 \bar{X} \\ &= (\bar{y} - \widehat{B}_1 \bar{x}) + \widehat{B}_1 \bar{X} \\ &= \bar{y} + \widehat{B}_1 (\bar{X} - \bar{x}) \end{aligned}$$

Regression estimator of population total

- **Regression estimator** of population **total**:

$$\begin{aligned} \hat{t}_{reg} &= \widehat{B}_0 + \widehat{B}_1 t_x \\ &= \hat{t}_y + \widehat{B}_1 (t_x - \hat{t}_x) \end{aligned}$$

- The regression estimator **adjusts** $\bar{y}$ and $\hat{t}_y$ by the amounts $\widehat{B}_1(\bar{X} - \bar{x})$ and $\widehat{B}_1(t_x - \hat{t}_x)$, respectively
- If X and Y positively correlated and $\bar{x} < \bar{X}$, then $\bar{y}$ would be expected to be smaller than $\bar{Y}$

Variance of the regression estimators

- **Variance** of $\hat{\bar{Y}}_{reg}$:

$$V(\hat{\bar{Y}}_{reg}) = (1-f)\frac{S_e^2}{n},$$

$$S_e^2 = \frac{1}{N-1}\sum_{i\in U} e_i^2, \quad e_i = y_i - B_0 - B_1 x_i$$

- **Variance** of $\hat{t}_{reg}$:

$$V(\hat{t}_{reg}) = N^2 V(\hat{\bar{Y}}_{reg}) = N^2(1-f)\frac{S_e^2}{n}$$

Estimated variance of the regression estimator

- *Estimated* **variance** of $\hat{\bar{Y}}_{reg}$:

$$\widehat{V}(\hat{\bar{Y}}_{reg}) = (1-f)\frac{s_e^2}{n},$$

$$s_e^2 = \frac{1}{n-1}\sum_{i\in s}\hat{e}_i^2, \quad \hat{e}_i = y_i - \widehat{B}_0 - \widehat{B}_1 x_i$$

- *Estimated* **variance** of $\hat{t}_{reg}$:

$$\widehat{V}(\hat{t}_{reg}) = N^2\widehat{V}(\hat{\bar{Y}}_{reg}) = N^2(1-f)\frac{s_e^2}{n}$$

Regression estimator of mean vs sample mean

- The regression estimator is **biased** like the ratio estimator
- The **bias** often vanishes with **large** n: said to be **approximately unbiased**
- $V(\hat{\bar{Y}}_{reg})$ can be rewritten as follows:

$$V(\hat{\bar{Y}}_{reg}) = (1-f)\frac{S_y^2}{n}(1-\rho^2) = V(\bar{y})(1-\rho^2)$$

  - Comparing to $\bar{y}$, improvement in **precision** achieved with $\hat{\bar{Y}}_{reg}$ when there is a **high correlation** between X and Y

Regression estimation with weight adjustments

- The **regression estimation** of the total:

$$
\begin{aligned}
\hat{t}_{reg} &= \hat{t}_y + \widehat{B}_1(t_x - \hat{t}_x) \\
&= \sum_{i\in s} d_i y_i + \frac{\sum_{i\in s}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i\in s}(x_i-\bar{x})^2}(t_x - \hat{t}_x) \\
&= \sum_{i\in s} d_i y_i + \frac{\sum_{i\in s} d_i(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i\in s} d_i(x_i-\bar{x})^2}(t_x - \hat{t}_x) \\
&= \sum_{i\in s} d_i y_i \left[1 + \frac{(x_i-\bar{x})}{\sum_{i\in s} d_i(x_i-\bar{x})^2}(t_x - \hat{t}_x)\right] \\
&= \sum_{i\in s} d_i g_i y_i,
\end{aligned}
$$

- $g_i = 1 + \frac{(x_i-\bar{x})}{\sum_{i\in s}(x_i-\bar{x})^2}(t_x - \hat{t}_x)$, **vary** by **sample**

Regression estimation with weight adjustments (cont.)

- The **sampling weights** $d_i = N/n$ are adjusted by g_i
- The **weight adjustments** g_i **calibrate** the estimates on (N, t_x):

$$\sum_{i \in s} d_i g_i x_i = N, \quad \text{for } x_i = 1,$$

$$\sum_{i \in s} d_i g_i x_i = \hat{t}_x + \frac{\sum_{i \in s} d_i x_i (x_i - \bar{x})}{\sum_{i \in s} d_i (x_i - \bar{x})^2} (t_x - \hat{t}_x) = t_x,$$

since $\sum_{i \in s} d_i x_i (x_i - \bar{x}) = \sum_{i \in s} d_i (x_i - \bar{x})^2$

- g_i: **calibration factors**; $w_i = d_i g_i$: **calibration weights** (**adjusted weights**)

Summary: the use of auxiliary information

- **Sampling design**: SRSWOR; **target**: $\bar{Y}$

Method	Estimator	Residual (e_i)
Sample mean	$\bar{y}$	$y_i - \bar{y}$
Ratio	$\frac{\bar{y}}{\bar{x}}\bar{X}$	$y_i - \widehat{B}x_i$
Regression	$\bar{y} + \widehat{B}_1(\bar{X} - \bar{x})$	$y_i - \widehat{B}_0 - \widehat{B}_1 x_i$

- **Sampling design**: SRSWOR; **target**: t_y

Method	Estimator	Residual (e_i)
Expansion	$\hat{t}_y$	$y_i - \bar{y}$
Ratio	$\frac{\hat{t}_y}{\hat{t}_x}t_x$	$y_i - \widehat{B}x_i$
Regression	$N[\bar{y} + \widehat{B}_1(\bar{X} - \bar{x})]$	$y_i - \widehat{B}_0 - \widehat{B}_1 x_i$

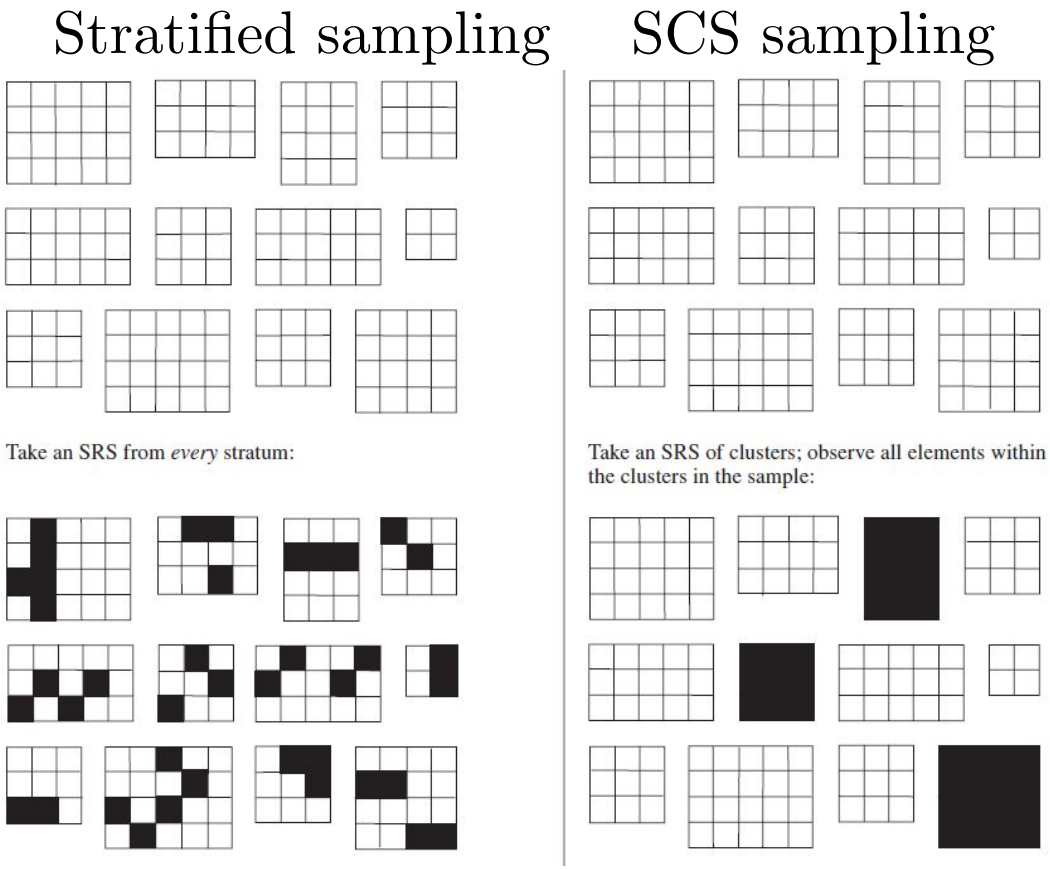
Session 6: Cluster sampling

Cluster sampling

1. Population elements are grouped into **clusters**
2. Select a **sample** of clusters under a sampling design
3. All or some of the elements in sample clusters are included in the sample
   - **One (single)-stage cluster sampling (SCS)**: all elements are included in the sample
   - **Multi-stage cluster sampling (MCS)**: some elements are selected within sample clusters

- Under **stratified sampling**: all strata are represented in the sample
- Under **cluster sampling**: Only some clusters are represented in the sample

Stratified sampling vs SCS



Source: Lohr (2019, p.167)

Nested structure

- Clusters often exist **naturally**
- Population elements grouped **hierarchically**

Example

Survey	Cluster	Element
LFS	Address block, household	Individuals
PISA	School, class	Teacher, pupil
Wage survey	Establishment	Employees
Agricultural survey	Areas	Farms
Health survey	Medical centres	Patients

Nested structure (cont.)

Stage 1: select a sample of **primary sampling units (PSUs)**

Stage 2: select a sample of **secondary-SUs (SSUs)** within sample **clusters** from **stage 1**

Stage 3: select a sample of **tertiary-SU (TSUs)** within sample **clusters** from **stage 2**

⋮ ⋮

Stage k: select a sample of **ultimate-SU (USUs)** within sample **clusters** from **stage (k-1)**

Why cluster sampling?

1. **No access** to the **frame** of **ultimate sampling units (USUs)** or **no frame** available for USUs, but a frame **primary sampling units (PSU)s** available
2. To **reduce** travel costs: population may be widely spread geographically

- Cluster sampling generally decreases **precision**, unlike stratified sampling, due to the **homogeneity** within the clusters
  - Members of same household may have similar political views
  - Students at/in same school/class tend to have similar level of reading & mathematics achievement

Notation

	Population	Sample
PSU-level	U	s
Index for **PSUs**	$i \in \{1, \ldots, N\}$	$i \in \{1, \ldots, n\}$
Within PSU i	U_i	s_i
Index for **SSUs** in PSU i	$j \in \{1, \ldots, M_i\}$	$j \in \{1, \ldots, m_i\}$
Number of **PSUs**	N	n
Number of **SSUs** in PSU i	M_i	m_i
Total number of **SSUs**	$M_0 = \sum_{i \in U} M_i$	$m_0 = \sum_{i \in s} m_i$
Average number of **SSUs** per PSU	$\bar{M} = M_0/N$	$\bar{m} = m_0/n$

Notation (cont.)

- y_{ij}: **Outcome variable** for **element** j of **cluster** i

Population quantities:	PSU-level
Value	$t_i = \sum_{j \in U_i} y_{ij}$
Total	$t_y = \sum_{i \in U} t_i$
Mean per **cluster**	$\bar{Y}_c = t_y / N$
Variance of the PSU totals	$S_t^2 = \frac{1}{N-1} \sum_{i \in U} (t_i - \bar{Y}_c)^2$

Notation (cont.)

Population quantities	SSU-level
Value	y_{ij}
Total	$t_y = \sum_{i \in U} \sum_{j \in U_i} y_{ij}$
Mean	$\bar{Y} = \frac{t_y}{M_0}$
Mean within PSU i	$\bar{Y}_i = \frac{t_i}{M_i}$
Variance	$S^2 = \frac{1}{M_0 - 1} \sum_{i \in U} (y_{ij} - \bar{Y})^2$
Variance within PSU i	$S_i^2 = \frac{1}{M_i - 1} \sum_{j \in U_i} (y_{ij} - \bar{Y}_i)^2$

Sampling design

- Clusters selected by **SRSWOR**
  - 1st stage **inclusion probabilities**: $\pi_i = n/N$
- All elements in sample clusters taken: $m_i = M_i$
  - 2nd stage **inclusion probabilities** (given $i \in s$): $\pi_{j|i} = 1$
- **Inclusion probabilities** of element j in the ith PSU: $\Pr(i \in s, j \in s_i) = \pi_{ij} = \pi_i \pi_{j|i} = n/N$
- **Sampling weight** for SSU j in PSU i: $d_{ij} = \pi_{ij}^{-1} = N/n$

Unbiased estimation under SCS

- **Unbiased** *estimator* for population **mean** per cluster, $\bar{Y}_c$:

$$\bar{y}_c = \frac{\sum_{i \in s} t_i}{n}$$

- **Unbiased** *estimator* for population **total**, t_y:

$$\hat{t}_y = N \frac{\sum_{i \in s} t_i}{n} = N \bar{y}_c = \sum_{i \in s} \sum_{j \in s_i} d_{ij} y_{ij}$$

- **Unbiased** *estimator* for population **variance**, S_t^2:

$$s_t^2 = \frac{1}{n-1} \sum_{i \in s} (t_i - \bar{y}_c)^2$$

Unbiased estimation under SCS (cont.)

- **Variance** of $\bar{y}_c$: $V_{scs}(\bar{y}_c) = (1-f)S_t^2/n$
- *Estimated* **variance** of $\bar{y}_c$:

$$\widehat{V}_{scs}(\bar{y}_c) = (1-f)\frac{s_t^2}{n}$$

- **Variance** of $\hat{t}_y$: $V_{scs}(\hat{t}_y) = N^2 V_{scs}(\bar{y}_c)$
- *Estimated* **variance** of $\hat{t}_y$:

$$\widehat{V}_{scs}(\hat{t}_y) = N^2\widehat{V}_{scs}(\bar{y}_c) = N^2(1-f)\frac{s_t^2}{n}$$

Unbiased estimation under SCS (cont.)

- Suppose M_0 is **known**
- The **unbiased** estimator of the population mean, $\bar{Y}$:

$$\bar{y} = \frac{\hat{t}_y}{M_0} = \frac{N\bar{y}_c}{M_0}$$

- **Variance** of $\bar{y}$:

$$V_{scs}(\bar{y}) = \left(\frac{N}{M_0}\right)^2 V_{scs}(\bar{y}_c) = \frac{1}{\bar{M}^2}(1-f)\frac{S_t^2}{n}$$

- The **unbiased variance** *estimator* of $V_{scs}(\bar{y})$:

$$\widehat{V}_{scs}(\bar{y}) = \left(\frac{N}{M_0}\right)^2 \widehat{V}_{scs}(\bar{y}_c) = \frac{1}{\bar{M}^2}(1-f)\frac{s_t^2}{n}$$

Ratio estimation under SCS

- Suppose M_0 is **known**
- Let $x_i = M_i$ (or $x_{ij} = 1$), and so that:

$$t_x = \sum_{i \in U} x_i = M_0, \quad t_{xs} = \sum_{i \in s} x_i = m_0$$

- The population **ratio**: $B = t_y/t_x = t_y/M_0 = \bar{Y}$ can be **estimated** by

$$\widehat{B} = \frac{t_{ys}}{t_{xs}} = \frac{\sum_{i \in s} t_i}{m_0} = \hat{\bar{Y}}_{rt}$$

- **Ratio-to-size** estimator of total t_y:

$$\hat{t}_r = \widehat{B} t_x = \frac{t_{ys}}{t_{xs}} t_x = \hat{\bar{Y}}_{rt} M_0$$

Ratio estimation under SCS (cont.)

- **Variance** of $\hat{t}_r$ is simply the variance of a **ratio estimator**:

$$V_{scs}(\hat{t}_r) \approx N^2(1-f)\frac{S_e^2}{n}, \quad S_e^2 = \frac{1}{N-1}\sum_{i\in U} e_i^2,$$

with $e_i = t_i - \bar{Y}M_i$

- *Estimated* **variance** of $\hat{t}_r$:

$$\widehat{V}_{scs}(\hat{t}_r) = (1-f)\left(\frac{M_0}{\bar{m}}\right)^2 \frac{s_e^2}{n}, \quad s_e^2 = \frac{1}{n-1}\sum_{i\in s} \hat{e}_i^2,$$

with $\hat{e}_i = t_i - \hat{\bar{Y}}_{rt}M_i$

Ratio estimation under SCS (cont.)

- Regardless of M_0 **known** or **unknown**, the population mean, $\bar{Y}$, can be estimated by a **ratio estimator** as follows:

$$\hat{\bar{Y}}_{rt} = \frac{\hat{t}_y}{\widehat{M}_0} = \frac{\sum_{i\in s}\sum_{j\in s_i} d_{ij}y_{ij}}{\sum_{i\in s}\sum_{j\in s_i} d_{ij}} = \frac{t_{ys}}{m_0} = \widehat{B}$$

- **Variance** of $\hat{\bar{Y}}_{rt}$:

$$V_{scs}(\hat{\bar{Y}}_{rt}) = \frac{1}{M_0^2} V_{scs}(\hat{t}_r) = (1-f)\frac{1}{\bar{M}^2}\frac{S_e^2}{n}$$

- *Estimated* **variance** of $\hat{\bar{Y}}_{rt}$:

$$\widehat{V}_{scs}(\hat{\bar{Y}}_{rt}) = \frac{1}{M_0^2} \widehat{V}_{scs}(\hat{t}_r) = (1-f)\frac{1}{\bar{m}^2}\frac{s_e^2}{n}$$

Comparison: estimators of population total

1. $\hat{t}_y = N\bar{y}_c$ **unbiased** whereas $\hat{t}_r$ **approximately unbiased** for t_y
2. $\hat{t}_r$ often **more precise** than $\hat{t}_y$: $V_{scs}(\hat{t}_r) < V_{scs}(\hat{t}_y)$

$$\begin{aligned} V_{scs}(\hat{t}_r) &\approx N^2(1-f)\frac{1}{n}\frac{1}{N-1}\sum_{i\in U}(t_i - \bar{Y}M_i)^2 \\ &= N^2(1-f)\frac{1}{n}\frac{1}{N-1}\sum_{i\in U}M_i^2(\bar{Y}_i - \bar{Y})^2 \end{aligned}$$

- If $\bar{Y}_i \approx \bar{Y}$ for all $i \in U$, then $V_{scs}(\hat{t}_r) \approx 0$

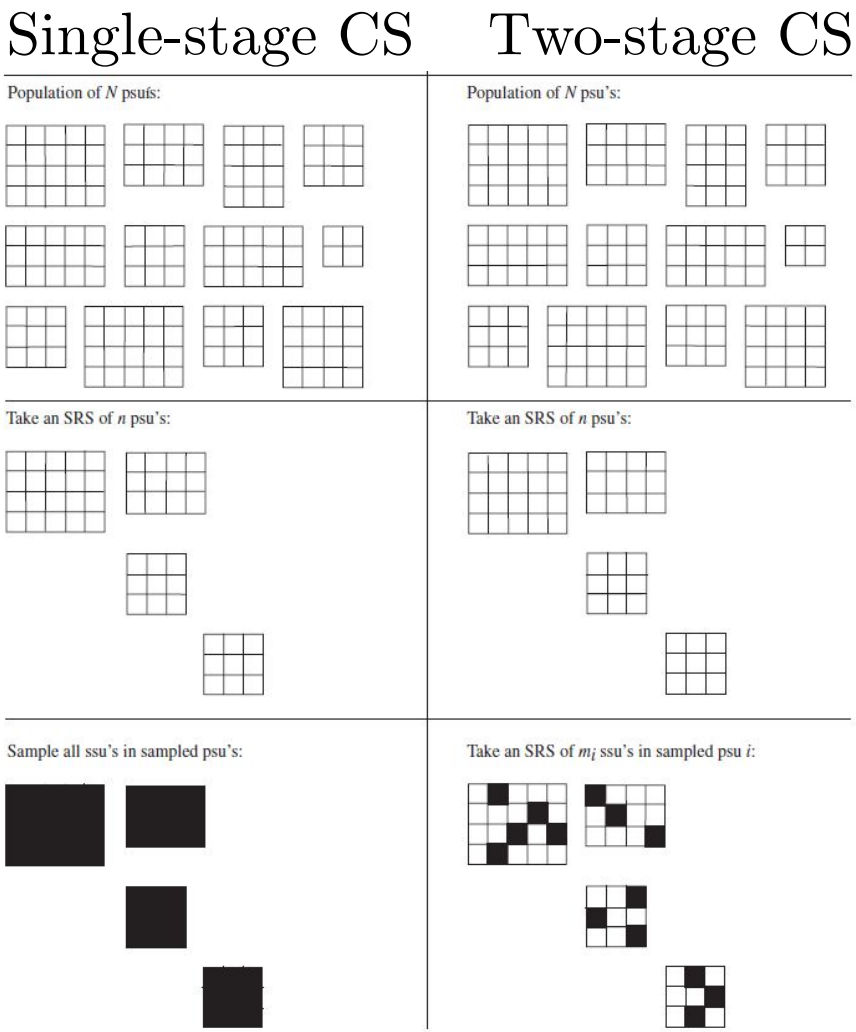
Comparison: estimators of population total (cont.)

3. $\bar{Y}_i$ usually **less variable** than t_i across clusters Remember: $V_{scs}(\hat{t}_y) = (1-f)n^{-1}\sum_{i\in U}(t_i - \bar{Y}_c)^2/(N-1)$
4. $\hat{t}_r$ often **preferable** due to the possible **gain in precision**
5. $\hat{t}_r$ requires knowledge of M_0, while $\hat{t}_y$ does not
6. If M_0 **unknown**, $\hat{t}_r$ cannot be used

Comparison: estimators of population mean

1. $\bar{y}$ **unbiased** whereas $\hat{\bar{Y}}_{rt}$ **approximately unbiased** for population mean $\bar{Y}$
2. $\hat{\bar{Y}}_{rt}$ usually **more precise** then $\bar{y}$ for the same reasons as estimators of total
3. $\bar{y}$ requires knowledge of M_0, while $\hat{\bar{Y}}_{rt}$ does not
4. $\hat{\bar{Y}}_{rt}$ almost always **preferable** even M_0 known
5. If M_0 **unknown**, $\bar{y}$ cannot be used

SCS vs TCS



Source: Lohr (2019, p.183)

Sampling design

- Observing all the elements in sample clusters may not be **cost-effective**, as some clusters may be quite large
- Given **homogeneity** within clusters and a limited **budget**, selecting more clusters and taking some of the elements within sample clusters may be a more efficient strategy than SCS
- Select n PSUs with **SRSWOR**: $\pi_i = n/N$
- Select m_i SSUs with **SRSWOR** from the ith sample PSU: $\pi_{j|i} = m_i/M_i$

Sampling design (cont.)

- **Inclusion probabilities** of jth SSU in the ith PSU:

$$\Pr(i \in s, j \in s_i) = \pi_{ij} = \pi_i \pi_{j|i} = \frac{n}{N} \frac{m_i}{M_i}$$

- **Sampling weight** for SSU j in PSU i:

$$d_{ij} = \frac{1}{\pi_{ij}} = \frac{NM_i}{nm_i}$$

- **EPSEM** design if: $\frac{n}{N} \frac{m_i}{M_i}$ = **constant**, which is the case if m_i/M_i = **constant**
- However, $m_0 = \sum_{i \in s} m_i$ **random**
- If $m_i = m$, total **sample size** nm **fixed**, but not an EPSEM design since nm/NM_i not **constant**

Estimation under TCS (cont.)

- **Unbiased** *estimator* of population total, t_y:

$$\hat{t}_y = \sum_{i \in s} \sum_{j \in s_i} d_{ij} y_{ij} = \sum_{i \in s} \sum_{j \in s_i} \frac{N M_i}{n m_i} y_{ij}$$

- **Variance** of $\hat{t}_y$:

$$\begin{aligned} V_{tcs}(\hat{t}_y) &= N^2(1-f)\frac{S_t^2}{n} + \frac{N}{n}\sum_{i \in U} M_i^2 \left(1 - \frac{m_i}{M_i}\right)\frac{S_i^2}{m_i} \\ &= V_{psu} + V_{ssu} \end{aligned}$$

V_{psu}: **variance** due to the **1st-stage sampling**
V_{ssu}: **variance** due to the **2nd-stage sampling**

- V_{psu} often dominates V_{ssu} (when N is large)

Estimation under TCS (cont.)

- For the estimation of the **variance** of $\hat{t}_y$, let

$$s_t^2 = \frac{1}{n-1} \sum_{i \in s} \left(\hat{t}_i - \frac{\hat{t}_y}{N} \right)^2,$$

be the **sample variance** among the **estimated** PSU totals, and

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

be the **sample variance** of the SSUs in the ith PSU

Estimation under TCS (cont.)

- $\hat{t}_i$: **estimated** total for PSU i:

$$\hat{t}_i = \sum_{j \in s_i} \frac{M_i}{m_i} y_{ij}, \quad (\hat{t}_i = t_i \text{ under SCS, i.e., } m_i = M_i)$$

- $\bar{y}_i$: **sample mean** for PSU i: $\bar{y}_i = \sum_{j \in s_i} y_{ij}/m_i$
- *Estimated* **variance** of $\hat{t}_y$:

$$\widehat{V}_{tcs}(\hat{t}_y) = N^2(1-f)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i \in s} M_i^2 \left(1 - \frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}$$

Estimation under TCS (cont.)

- The **population mean** $\bar{Y}$ can be estimated by the **ratio estimator** ($x_i = M_i$):

$$\hat{\bar{Y}}_{rt} = \frac{\hat{t}_y}{\widehat{M}_0} = \frac{\sum_{i \in s} \hat{t}_i}{\sum_{i \in s} M_i}$$

- *Estimated* **variance** of $\hat{\bar{Y}}_{rt}$:

$$\widehat{V}_{tcs}(\hat{\bar{Y}}_{rt}) = (1-f)\frac{1}{\bar{m}^2}\frac{s_e^2}{n} + \frac{1}{nN\bar{m}^2}\sum_{i \in s} M_i^2 \left(1 - \frac{m_i}{M_i}\right)\frac{s_i^2}{m_i},$$

$$s_e^2 = \frac{1}{n-1}\sum_{i \in s}(\hat{t}_i - \hat{\bar{Y}}_{rt} M_i)^2 = \frac{1}{n-1}\sum_{i \in s} M_i^2 (\bar{y}_i - \hat{\bar{Y}}_{rt})^2$$

Design effect (DEFF)

Definition

Compares variance of a sampling strategy (Sampling design + Estimator) to that of SRSWOR of elements, assuming the same sample size of elements

Example

Sampling strategy: SCS by SRS with **ratio** estimator, $\hat{\bar{Y}}_{rt}$
Reference sampling strategy: SRSWOR of m_0 elements, equal to that under SCS, and **sample mean** estimator $\bar{y}$

$$\text{DEFF}(\text{SCS}, \hat{\bar{Y}}_{rt}) = \frac{V_{scs}(\hat{\bar{Y}}_{rt})}{V_{srs}(\bar{y})}$$

DEFF (cont.)

- **Design factor** (DEFT): $\text{DEFT} = \sqrt{\text{DEFF}}$
- $\text{DEFF} > 1$: **loss of efficiency** due to using a sampling strategy different from (SRSWOR + $\bar{y}$)
- $\text{DEFF} < 1$: **gain of efficiency** due to using a sampling strategy different from (SRSWOR + $\bar{y}$)
- $\text{DEFF}(\text{SCS}, \hat{\bar{Y}}_{rt})$ can be estimated by

$$\widehat{\text{DEFF}}(\text{SCS}, \hat{\bar{Y}}_{rt}) = \frac{\widehat{V}_{scs}(\hat{\bar{Y}}_{rt})}{\widehat{V}_{srs}(\bar{y})}$$

DEFF (cont.)

- If cluster sizes are **equal**: $M_i = M_0/N = \bar{M}$, then

$$\mathrm{DEFF}(\mathrm{SCS}, \hat{\bar{Y}}_{rt}) \approx 1 + (\bar{M} - 1)\rho,$$

  ρ is the **intra-cluster correlation coefficient**, which measures how vary the values within the same cluster

- Often $\rho > 0$ as elements within the same cluster tend to be **more homogeneous** than across the whole population
- Cluster sampling usually **less efficient**

DEFF (cont.)

- Design effect very useful in specifying of **sample size** under complex designs
  - Calculate sample size under SRSWOR given a precision criteria, n_{srs}
  - Multiply n_{srs} with DEFF to calculate the sample size needed under the **complex design** to achieve the same level of efficiency as that under SRS

$$n_{design} = n_{srs} * \text{DEFF}$$

Session 7: Systematic sampling

Introduction

- Can be used for cases when **sampling frame** is **unavailable**
- Population elements **sorted** in some way (**randomly** or **nonrandomly**), a random starting point r picked, and thereafter *every* kth unit taken in the sample
  - Visit every 10th house
  - Select every 100th person in a telephone book
  - Interview every 50th tourist arriving in Kyiv Boryspil Airport
- Commonly applied at the **final stage** of selection in **cluster sampling**
- Very **easy** to implement

Sampling design

- Let $U = \{1, \dots, N\}$ and $N = nk + c$, with $0 \leq c < n$
- Pick a random integer r, $1 \leq r \leq k$, with **probability** $1/k$
- Select rth unit and *every* kth unit thereafter:

$$s_r = \{r, r + k, r + 2k, \dots\}$$

- k: **sampling interval**

Example

Let $N = 9$ and $n = 3$. $\Rightarrow k = 3$ and $c = 0$
For $r = 1$: $s_1 = \{1, 4, 7\} \Rightarrow n_1 = 3$
For $r = 2$: $s_2 = \{2, 5, 8\} \Rightarrow n_2 = 3$
For $r = 3$: $s_3 = \{3, 6, 9\} \Rightarrow n_3 = 3$
There are $k = 3$ possible **distinct samples**, each of size 3

Examples

Example

Let $N = 13$ and $n = 3$. $\Rightarrow k = 4$ and $c = 1$
For $r = 1$: $s_1 = \{1, 5, 9, 13\} \Rightarrow n_1 = 4$
For $r = 2$: $s_2 = \{2, 6, 10\} \Rightarrow n_2 = 3$
For $r = 3$: $s_3 = \{3, 7, 11\} \Rightarrow n_3 = 3$
For $r = 4$: $s_4 = \{4, 8, 12\} \Rightarrow n_4 = 3$
There are $k = 4$ possible **distinct samples**, three of size 3 and one of size 4

Example

Let $N = 13$ and $n = 5$. $\Rightarrow k = 2$ and $c = 3$
For $r = 1$: $s_1 = \{1, 3, 5, 7, 9, 11, 13\} \Rightarrow n_1 = 7$
For $r = 2$: $s_2 = \{2, 4, 6, 8, 10, 12\} \Rightarrow n_2 = 6$
There are $k = 2$ possible **distinct samples**, one of size 7 and one of size 6

Sampling design (cont.)

- If $c > 0$, **actual** sample size **varies**, and one may never **achieve** the desired sample size
- There are k possible **distinct** samples with a selection probability of $\Pr(s_r) = 1/k$
- Can be envisaged as a special case of **cluster sampling**: one PSU chosen with SRSWOR from k PSUs
- Unlike SRSWOR, there are many subsets of U with **zero probability** of selection
- Every element can only be in **one sample**
- **Inclusion probability**: $\pi_i = 1/k$
- **Sampling weight**: $d_i = k$

Estimation of mean and its variance

- The **unbiased** estimator of the population mean $\bar{Y}$:

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{i\in s} d_i y_i = \frac{k}{N} t_{ys}, \quad t_{ys} = \sum_{i\in s} y_i$$

  - If $N = nk$, then $\hat{\bar{Y}} = \bar{y}$ (**sample mean**)

- **Variance** of $\hat{\bar{Y}}$:

$$V_{sys}(\hat{\bar{Y}}) = \frac{1}{k}\sum_{s}\left(\frac{k}{N} t_{ys} - \bar{Y}\right)^2$$

- There does <u>not</u> exist an **unbiased** estimator for $V_{sys}(\hat{\bar{Y}})$

Systematic sampling vs SRSWOR

- If the population elements sorted **randomly** (i.e. ordering **unrelated** to the variable of interest: **alphabetic** order, ordering by the last four digits of telephone numbers, etc.), it can be treated as if units were selected with **SRSWOR**, and the associated variance estimator can be used
- **Increasing** or **decreasing** order: yields approximately **proportionate stratified sample**
- Let $N = nk$; select one unit per stratum at random

Units	Implicit strata
$1,2,\cdots,k$	1
$k+1,\cdots,2k$	2
$\vdots$	$\vdots$
$(n-1)k+1,\cdots nk$	n

Systematic sampling vs SRSWOR (cont.)

- If units within implicit strata **homogeneous**, systematic sampling **more efficient** than SRSWOR
  - Financial records with increasing or decreasing order
- **Periodic pattern**: the sampling interval coincides with a multiple of the period: systematic sampling will be **less efficient** than SRSWOR

Example

Let population values be ordered as such: {1,2,3,1,2,3,1,2,3,1,2,3}, and $k = 3$. $\bar{Y} = 2$ and $S^2 = 0.73$

$s_1 = \{1, 1, 1, 1\}$, $s_1 = \{2, 2, 2, 2\}$, $s_3 = \{3, 3, 3, 3\}$

$V_{sys}(\hat{\bar{Y}}) = \frac{1}{3}[(1-2)^2 + (2-2)^2 + (3-2)^2] = 0.67$

$V_{srs}(\bar{y}) = (1 - 4/12)S^2/4 = 0.12$

Session 8: Unequal probability sampling

PPS sampling

- Cluster sampling with **equal probabilities** may lead to **large variance** for the **design-unbiased** estimators of the population total and mean
- Let $n = 1$ PSU be selected from N PSUs in the population with **selection probability proportional to cluster-size**:

$$p_i = \frac{M_i}{\sum_{i \in U} M_i} = \frac{M_i}{M_0}$$

- **Unbiased estimator** of the population total under **SCS**:

$$\hat{t}_y = \sum_{i \in s} d_i t_i = \sum_{i \in s} \frac{t_i}{p_i}$$

Example: PPS sampling vs SRSWOR

Example

A population of supermarkets in a town. Select $n = 1$ store. Estimate the total amount of sales. (Lohr, 2019, pp.222-224)

Store	Size (M_i)	p_i	$\pi_{i;srs}$	t_i	$\hat{t}_{pps}$	$\hat{t}_{srs}$
A	100	1/16	1/4	11	176	44
B	200	2/16	1/4	20	160	80
C	300	3/16	1/4	24	128	96
D	1 000	10/16	1/4	245	392	980
Total	1 600	1	1	300		

Under **PPS**: $V_{pps}(\hat{t}_{pps}) = \frac{1}{16}(176 - 300)^2 + \cdots + \frac{10}{16}(392 - 300)^2 = 14.25$
Under **SRSWOR**:
$V_{srs}(\hat{t}_{srs}) = \frac{1}{4}(44 - 300)^2 + \cdots + \frac{1}{4}(980 - 300)^2 = 154.45$

SCS with $n = 1$

Example

Cluster	M_i	p_i	Cumulative M_i	Range
1	43	0.0811	43	1- 43
2	98	0.1849	141	44-141
3	20	0.0377	161	142-161
4	121	0.2283	282	162-282
5	160	0.3019	442	283-442
6	88	0.1660	530	443-530

Let $n = 1$. Pick a **random number** between 1 and $M_0 = 530$: f.exp., $r = 300$. Find the range including r. Select the cluster corresponding to that range: **Cluster 5**.

SCS with $n > 1$

- If $n > 1$, **repeat** the process, drawing one cluster at a time, **independently** until reaching n draws
- Let i be the cluster selected on the kth draw, with $k = 1, \ldots, n$
- **Unbiased** estimator of the population total on the kth draw: $\hat{t}_y(k) = t_i/p_i$
- With n draws, **independent** sample: $s : \{\hat{t}_y(1), \ldots, \hat{t}_y(n)\}$
- s may contain **repeated** elements

Estimation under SCS

- **Unbiased** estimator of the population total:

$$\hat{t}_y = \frac{1}{n} \sum_{i \in s} \frac{t_i}{p_i}$$

- Given **independent and identically distributed (iid)** sample, **unbiased** estimator of the variance of $\hat{t}_y$:

$$\widehat{V}_{pps}(\hat{t}_y) = \frac{s_t^2}{n} = \frac{1}{n(n-1)} \sum_{i \in s} \left(\frac{t_i}{p_i} - \hat{t}_y \right)^2$$

Estimation under SCS (cont.)

- If $p_i = M_i/M_0$:

$$\widehat{M}_0 = \frac{1}{n}\sum_{i\in s}\frac{M_i}{p_i} = M_0$$

- The **ratio** estimator becomes **unbiased** for the population **mean**:

$$\hat{\bar{Y}} = \frac{\hat{t}_y}{\widehat{M}_0} = \frac{1}{M_0}\frac{1}{n}\sum_{i\in s}\frac{M_0 t_i}{M_i} = \frac{1}{n}\sum_{i\in s}\bar{y}_i$$

- *Estimated* **variance**:

$$\widehat{V}_{pps}(\hat{\bar{Y}}) = \frac{s^2}{n} = \frac{1}{n(n-1)}\sum_{i\in s}(\bar{y}_i - \hat{\bar{Y}})^2$$

s^2: the **sample variance** of the PSU means $\bar{y}_i$

TCS

- **Two-stage cluster** sampling with **unequal-probabilities** with replacement almost the same as those for SCS
- PSUs are selected independently on n draws with probabilities p_i
- **Some** of the elements in **sample PSUs** are selected, for example, with SRSWOR or systematic sampling

Two requirements:

1. **Invariance**: every time a PSU is selected, the **same** sampling design used in selection of the SSUs from that PSU
2. **Independence**: if a PSU selected more than once, then the **sub-sampling** of the **SSUs** from each replicate of the sample PSU, must be done **independently**: generate a different set of random numbers before the selection

Estimation under TCS

- Let $\hat{t}_i$ be an **unbiased** estimator for the PSU total t_i
- **Unbiased** estimator of the population **total**:

$$\hat{t}_y = \frac{1}{n} \sum_{i \in s} \frac{\hat{t}_i}{p_i}$$

- **Unbiased** estimator of the **variance** of $\hat{t}_y$:

$$\widehat{V}_{pps}(\hat{t}_y) = \frac{1}{n(n-1)} \sum_{i \in s} \left(\frac{\hat{t}_i}{p_i} - \hat{t}_y \right)^2$$

An EPSEM design for TCS

- Let PSUs be selected with **replacement** in n draws with probabilities $p_i = M_i/M_0$ and $m_i = m$ SSUs be selected with **SRSWOR** from each sample PSU
- **Ultimate inclusion probabilities**, π_{ij}, all equal: **EPSEM** design:

$$\pi_{ij} = \frac{M_i}{M_0}\frac{m}{M_i} = \frac{m}{M_0} = \textbf{constant}$$

- Total **sample size** is fixed: nm
- PPS sample is also called **self-weighting** as d_{ij} all equal:

$$d_{ij} = \frac{M_0}{nM_i}\frac{M_i}{m} = \frac{M_0}{nm} = \textbf{constant}$$

PPS sampling without replacement

- Generally sampling with replacement **less efficient** than sampling without replacement
- In practice, samples selected without replacement
- The theory of **unequal-probability sampling without replacement** much more complicated
- Probability of selection at each draw adjusted depending on which units selected in the previous draws: not **independent**
- Units selected with **probabilities proportional** to some **size measures**, x_i,:

$$\pi_i \propto x_i \Rightarrow \pi_i = c\, x_i, \quad \text{for some constant } c$$

Inclusion probabilities

- For **fixed** sample size designs:

$$\sum_{i \in U} \pi_i = n$$

$$\sum_{i \in U} c\, x_i = c\, X = n \Rightarrow c = \frac{n}{X}$$

- Hence,

$$\pi_i = n \frac{x_i}{X}$$

- $\pi_i \leq 1 \Rightarrow x_i \leq X/n$ must be **satisfied**
- If not, take elements with $x_i > X/n$ with **certainty**, and apply the selection procedure to the rest of the population units

Horvitz-Thompson (HT) estimator

- Suppose the **inclusion probabilities** $\pi_i = \Pr(i \in s)$ and the **joint inclusion probabilities** $\pi_{ik} = \Pr(i \in s, k \in s)$ are known
- Under *any* sampling **without replacement**, the **Horvitz-Thompson** (HT) (1952) estimator of the population total:

$$\hat{t}_{\text{HT}} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

unbiased for t.

HT variance estimator

- **Variance** of the **HT-estimator**:

$$V(\hat{t}_{\mathrm{HT}}) = \sum_{i \in U} \sum_{k \in U} (\pi_{ik} - \pi_i \pi_k) \frac{y_i}{\pi_i} \frac{y_k}{\pi_k}$$

- There are two **unbiased** estimators of the variance $V(\hat{t}_{\mathrm{HT}})$, both of which require $\pi_{ik} > 0$
- The **HT-estimator** of the variance:

$$\widehat{V}_{\mathrm{HT}}(\hat{t}_{\mathrm{HT}}) = \sum_{i \in s} \sum_{k \in s} \frac{(\pi_{ik} - \pi_i \pi_k)}{\pi_{ik}} \frac{y_i}{\pi_i} \frac{y_k}{\pi_k}$$

Sen-Yates-Grundy (SYG) variance estimator

- The **Sen-Yates-Grundy (SYG) estimator** of the variance:

$$\widehat{V}_{\text{SYG}}(\hat{t}_{\text{HT}}) = \sum_{i \in s} \sum_{k \neq i \in s} \frac{(\pi_{ik} - \pi_i \pi_k)}{\pi_{ik}} \left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k} \right)^2$$

- Under **SCS** where **PSUs** selected with unequal-probabilities without replacement, treat t_i as **observational units** and replace y_i with t_i in the estimators for total and its variance
- Both variance estimators requires **joint inclusion probabilities** known, which may be challenging or not possible to calculate exactly in practice if the π_i unequal

Remarks

- The **HT or SYG variance estimators** may yield **negative** variance estimates under some unequal probability designs, when $\pi_{ik} < \pi_i \pi_k$ for some units (i, k)
- They can also be very **unstable**: vary a lot from sample to sample
- **Approximations** to π_{ik} that yield **positive** variance estimates available (e.g. **Hájek (1964)'s approximation**)
- If N **large** and **sampling fraction** negligible, $n/N \to 0$, without replacement design can be **approximated** by that with replacement

With replacement approximation, SCS

- Assuming $\pi_i = np_i$ 1st-stage inclusion probabilities under design without replacement, the **with replacement** variance estimator:

$$\widehat{V}_{\mathrm{WR}}(\hat{t}_{\mathrm{HT}}) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in s}\left(\frac{t_i}{p_i} - \hat{t}_{\mathrm{HT}}\right)^2 = \frac{n}{n-1}\sum_{i\in s}\left(\frac{t_i}{\pi_i} - \frac{\hat{t}_{\mathrm{HT}}}{n}\right)^2$$

- $\widehat{V}_{\mathrm{WR}}(\hat{t}_{\mathrm{HT}})$ **positively biased** for $V(\hat{t}_{\mathrm{HT}})$ under design without replacement: more conservative variance estimates obtained since **fpc**$= 1 - f$ not included in $\widehat{V}_{\mathrm{WR}}(\hat{t}_{\mathrm{HT}})$
- **Joint inclusion** probabilities not required
- Variance estimates **always positive**

TCS

- Under **TCS**, the PSU total t_i replaced with its **unbiased estimator** $\hat{t}_i$

$$\hat{t}_{\text{HT}} = \sum_{i \in s} \frac{\hat{t}_i}{\pi_i}$$

- Its variance: $V(\hat{t}_{\text{HT}}) = V_{psu} + V_{ssu}$

$$V(\hat{t}_{\text{HT}}) = \sum_{i \in U} \sum_{k \in U} (\pi_{ik} - \pi_i \pi_k) \frac{t_i}{\pi_i} \frac{t_k}{\pi_k} + \sum_{i \in U} \frac{V(\hat{t}_i)}{\pi_i}$$

- The second term due to **estimating** the t_i rather than measuring them exactly

Variance estimators, TCS

- The **HT variance estimator** of $\hat{t}_{\mathrm{HT}}$:

$$\widehat{V}_{\mathrm{HT}}(\hat{t}_{\mathrm{HT}}) = \sum_{i \in s} \sum_{k \in s} \frac{(\pi_{ik} - \pi_i \pi_k)}{\pi_{ik}} \frac{\hat{t}_i}{\pi_i} \frac{\hat{t}_k}{\pi_k} + \sum_{i \in s} \frac{\widehat{V}(\hat{t}_i)}{\pi_i}$$

- The **SYG variance estimator** of $\hat{t}_{\mathrm{HT}}$:

$$\widehat{V}_{\mathrm{SYG}}(\hat{t}_{\mathrm{HT}}) = \sum_{i \in s} \sum_{k \neq i \in s} \frac{(\pi_{ik} - \pi_i \pi_k)}{\pi_{ik}} \left(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_k}{\pi_k} \right)^2 + \sum_{i \in s} \frac{\widehat{V}(\hat{t}_i)}{\pi_i}$$

Ultimate cluster approach, TCS

- Provided **negligible** 1st-stage **sampling fraction**, $n/N \to 0$:
  - The effect of the variance at the second stage to the total variance becomes **negligible**
  - The with replacement variance estimator can be used for a design without replacement with probabilities $\pi_i = np_i$:

$$\widehat{V}_{\mathrm{WR}}(\hat{t}_{\mathrm{HT}}) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in s}\left(\frac{n\hat{t}_i}{\pi_i} - \hat{t}_{\mathrm{HT}}\right)^2 = \frac{n}{n-1}\sum_{i\in s}\left(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_{\mathrm{HT}}}{n}\right)^2$$

- A huge **practical advantage**

Remarks

- If $f = n/N$ not **negligible**, the with replacement variance estimator can be **adjusted** with the **fpc**:

$$\widehat{V}^*(\hat{t}_{\mathrm{HT}}) = \left(1 - \frac{n}{N}\right) \widehat{V}_{\mathrm{WR}}(\hat{t}_{\mathrm{HT}})$$

- $\widehat{V}^*$ may be **negatively biased** for $V(\hat{t}_{\mathrm{HT}})$
- Variance of the HT-estimator will be **smaller** if $\pi_i \propto y_i$, (or $\pi_i \propto t_i$ under SCS)
- However, if the **inclusion probabilities** not related with the **variable of interest**, the HT-estimator may provide very **bad estimates**

When to avoid sampling weights?: Basu's elephants

- A circus owner has 50 **elephants**
- Interested in estimating the **total weight** of the elephants
- Wish to select the average elephant: '**Sambo**' (**purposive** sampling)
- Wish to estimate the total by $\hat{t} = 50 y_{sambo}$
- The circus statistician horrified when he learns this!!!
- He makes a sampling plan that allows Sambo to be selected with almost **certainty**

When to avoid sampling weights?: Basu's elephants (cont.)

- Assigns a probability $\pi_{sambo} = 99/100$ to Sambo and $\pi_i = 1/4900$ to the other elephants: $\sum_{i=1}^{50} \pi_i = 1$
- The statistician says $50y_{sambo}$ cannot be used, and he suggests using the **HT-estimator**: $\hat{t}_{\text{HT}} = \sum_{i \in s} \pi_i^{-1} y_i$, which gives $\hat{t}_{\text{HT}} = \frac{100}{99} y_{sambo}$ when sambo selected in the sample
- What if elephant '**Jumbo**', which is the heaviest elephant in the herd, is selected? asks the circus owner
- The statistician replies: $\hat{t}_{\text{HT}} = 4900 y_{jumbo}$
- The statistician losts his job at the circus afterwards!

Remarks: TCS

- Try to ensure **self-weighting** (**EPSEM**) design
  - For **two-stage** design where PSUs selected with probabilities $\pi_i \propto c\,x_i$, and m_i elements selected from each sample PSU: $\pi_{ij} = c\,x_i m_i / M_i$ **EPSEM** design if either

$$m_i \propto \frac{M_i}{x_i} \quad \text{or} \quad x_i \approx M_i \text{ and } m_i \approx m$$

- In specification of a cluster design, there should be a **balance** between **precision** and **cost**. Given fixed total sample size, nm,
  - Increasing n and decreasing $m \Rightarrow$ increases precision & cost
  - Decreasing n and increasing $m \Rightarrow$ decreases precision & cost

Session 9: Survey nonresponse

Survey nonresponse

- Increasing nonresponse in surveys
- High nonresponse may cause serious **bias** in estimation: Literary Digest Survey (1936) with a response rate of 24%
- Crucial to **prevent** it at the design stage as much as possible
- **Unit nonresponse**: no survey observations (returned questionnaire) at all
- **Item nonresponse**: observe some but not all the survey variables
- Nonrespondents often differ from respondents

Nonrespondents vs respondents

Example

1969 survey on voting behaviour by Statistics Norway. Three calls and a follow-up by mail. Nonresponse rate 9.9%. Norwegian voting register used to find out voting rate in the election.

		Age				
	All	20-24	25-29	30-49	50-69	70-79
Nonrespondents	71	59	56	72	78	74
Selected sample	88	81	84	90	91	84

Source: Lohr (2019, p. 330)

Design to prevent nonresponse

1. **Survey content**: ordering of sensitive questions, randomised response technique for sensitive questions
2. **Timing of survey**: avoid choosing a period that could yield low response rates: summer time for person surveys
3. **Interviewer**: varying response rates among interviewers
4. **Mode of data-collection**: generally telephone, mail and web surveys low response rates than in-person (face-to-face) surveys
5. **Questionnaire design**
6. **Response burden**: split-questionnaire design, negative coordination for business surveys

Design to prevent nonresponse (cont.)

7. **Survey introduction**: the purpose of the data usage, motivation, assured confidentiality
8. **Incentives and disincentives**: Monetary or non-monetary
9. **Call-backs or follow-up**
10. **Two-phase (double) sampling**: initial vs. later respondents, subsampling of the nonrespondents
11. **Responsive survey design** (Groves and Heeringa 2006)

Response propensity (probability)

- Suppose the population divided into two fixed strata: **respondents** and **nonrespondents**

$$R_i = \begin{cases} 1 & \text{if unit } i \in U \text{ responds} \\ 0 & \text{otherwise} \end{cases}$$

- r_i, the **realisation** of R_i, known for those in the sample
- The probability that a unit selected in the sample responds:

$$\phi_i = \Pr(R_i = 1), \quad 0 < \phi_i \leq 1$$

- ϕ_i: **response propensity (probability)**, **unknown**
- There are three types of mechanisms for missing data

Response mechanisms

1. **Missing Completely At Random (MCAR)**: response probabilities do not depend on y_i, x_i (known auxiliary information), or the survey design; that is: $\phi_i = \phi =$ **unknown constant**
   - Nonresponse can be **ignored**: f.exp., under **SRSWOR**, $\bar{y}_r$ (sample mean for the respondents) **approximately unbiased** for the population mean $\bar{Y}$
2. **Missing At Random (MAR)**: ϕ_i depends on x_i, but independent of y_i given x_i
   - The mean of the survey variable for respondents and nonrespondents similar within, f.exp., age, sex, education, ethnicity groups

Response mechanisms (cont.)

3. **Not Missing At Random (NMAR)**: ϕ_i depends on both x_i and y_i
   - Most likely the case in reality
   - Modelling needed: but not reasonable to expect a perfect model that explains the nonresponse mechanism
   - In practice, adjustment by treating the response mechanism as if MAR

Dealing with nonresponse

1. **Reweighting**: design weights adjusted to compensate for nonresponse
   - Generally used for **unit nonreponse**
   - Without nonresponse: $\hat{t} = \sum_{i \in s} d_i y_i = \sum_{i \in s} y_i / \pi_i$
   - With nonresponse: $\Pr(i \in s, R_i = 1) = \pi_i \phi_i$ and $\hat{t} = \sum_{i \in s} r_i w_i y_i$, with $w_i = 1/(\pi_i \hat{\phi}_i)$; $w_i > d_i = \pi_i^{-1}$, where $\hat{\phi}_i$ sample estimate for ϕ_i
2. **Imputation**: values are assigned to the missing items
   - Commonly used for **item nonreponse**
   - Recommended to create a variable in the data set that indicates whether the response measured or imputed

Weighting class adjustment

- Divide sample into C **classes**: $s = s_1 \cup \cdots \cup s_C$ of sizes $n = n_1 + \cdots + n_C$
- Denote the sample of respondents in class c by s_{rc}: $s_r = s_{r1} \cup \cdots \cup s_{rC}$ with $n_r = n_{r1} + \cdots + n_{rC}$
- **Assumption**: Respondents and nonrespondents within the same **weighting adjustment class** have similar characteristics
- **MAR** response mechanism or **MCAR** within weighting classes; that is, for i in class c with $r_i = 1$,

$$\phi_i = \phi_c,$$

  where ϕ_c the **response propensity** for class c, for $c = 1, \ldots, C$
- If $C = 1$, the response mechanism is **MCAR** *everywhere*

Weighting class adjustment (cont.)

- **Response propensity** for class c estimated by:

$$\hat{\phi}_c = \frac{\sum_{i \in s_c} r_i d_i}{\sum_{i \in s_c} d_i}$$

- Under **SRSWOR**: $\hat{\phi}_c = n_{rc}/n_c$, where n_{rc} the number of respondents in class c
- Design weights **adjusted** by $\hat{\phi}_c$:

$$w_i = \frac{d_i}{\hat{\phi}_c}, \quad \text{for } i \in s_{rc}$$

- $w_i = 0$ for nonrespondents, that is, if $r_i = 0$

Weighting class adjustment estimators

- The population **total** $\hat{t}$ estimated by:

$$\hat{t}_{wc} = \sum_{i \in s} r_i w_i y_i$$

- The population **mean** $\bar{Y}$ estimated by:

$$\hat{\bar{Y}}_{wc} = \frac{\hat{t}_{wc}}{\sum_{i \in s} r_i w_i}$$

Weighting class adjustment estimators (cont.)

- Under **SRSWOR**, the **weighting class adjustment** estimator of t:

$$\hat{t}_{wc} = \sum_{c=1}^{C} \sum_{i \in s_c} r_i \frac{N}{n} \frac{n_c}{n_{rc}} y_i = \frac{N}{n} \sum_{c=1}^{C} n_c \bar{y}_{rc}, \quad \bar{y}_{rc} = \frac{\sum_{i \in s_c} r_i y_i}{n_{rc}}$$

- Under **SRSWOR**, the **weighting class adjustment** estimator of $\bar{Y}$:

$$\hat{\bar{Y}}_{wc} = \frac{\hat{t}_{wc}}{\sum_{i \in s} r_i w_i} = \frac{\hat{t}_{wc}}{N} = \frac{1}{n} \sum_{c=1}^{C} n_c \bar{y}_{rc}, \quad \sum_{c=1}^{C} \sum_{i \in s_c} r_i \frac{N}{n} \frac{n_c}{n_{rc}} = N$$

Remarks: weighting class adjustment

- **Unstable** weights and so **large variance** if the class includes many nonrespondents (i.e. large nonresponse weights)
- Avoid weighting cells with **small** number of respondents
- Set a **threshold** for $\hat{\phi}_c$ and combine the neighbouring cells with $\hat{\phi}_c$ below the threshold: f.exp. $\hat{\phi}_c \geq 0.5 \Rightarrow 1/\hat{\phi}_c \leq 2$
- **Nonresponse bias** largely reduced if within a class:
  - the units **similar** in terms of the survey variables of interest
  - the response propensities **homogeneous**
  - the y_i **uncorrelated** with ϕ_i

Post-stratification

- Sample divided into H **post-strata**: $s = s_1 \cup \cdots \cup s_H$ of sizes $n = n_1 + \cdots + n_H$
- Denote the sample of respondents in stratum h by s_{rh}: $s_r = s_{r1} \cup \cdots \cup s_{rH}$ with $n_r = n_1 + \cdots + n_{rH}$
- Design weights **modified** in such a way that:

$$\sum_{i \in s_h} r_i w_i = N_h$$

- The modified weights for the respondent unit i in stratum h:

$$w_i = d_i \frac{N_h}{\sum_{i \in s_h} r_i d_i}$$

Post-stratified estimators under SRSWOR

- The **post-stratified** estimator of t under **SRSWOR**:

$$\hat{t}_{pst} = \sum_{h=1}^{H} \sum_{i \in s_h} r_i d_i \frac{N_h}{\sum_{i \in s_h} r_i d_i} y_i = \sum_{h=1}^{H} N_h \bar{y}_{rh}$$

- The **post-stratified** estimator of $\bar{Y}$ under **SRSWOR**:

$$\hat{\bar{Y}}_{pst} = \frac{\hat{t}_{pst}}{N} = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_{rh}$$

- The N_h **known** in post-stratification, while N_c **unknown** in the weighting class adjustment, $\hat{t}_{wc} = N \sum_{c=1}^{C} n_c \bar{y}_{rc}/n$, and estimated by $N n_c/n$

Remarks: post-stratification

- The **post-stratified** estimator **approximately unbiased** if within a stratum:
  - the units **similar** in terms of the survey variables of interest
  - the response propensities **homogeneous**
  - the y_i **uncorrelated** with ϕ_i
- Many post-strata with small number of respondents may lead to **unstable** estimates

Raking-ratio adjustment

- A special case of post-stratification that can be used in the presence of more than one post-stratification variable and only **marginal** totals known

Example

The sum of the sampling weights given in the cells of the following table. Post-stratification variables: gender and three age groups.

	15-24	25754	75+	$\hat{N}_{gender}$
Female	300	1 200	120	1 620
Male	150	1 080	150	1 380
$\hat{N}_{age}$	450	2 280	270	3 000

Source: Adapted from an Example in Lohr (2019, p.344)

Raking-ratio adjustment (cont.)

Example

- Suppose the marginal population counts are known and given by $N_{female} = 1\,510$, $N_{female} = 1\,490$, $N_{1524} = 600$, $N_{2574} = 2\,120$ and $N_{75+} = 280$
- Multiply the first row estimated counts by $1\,510/1\,620$ and the second row estimated counts by $1\,490/1\,380$:

	15-24	25-54	55+	$\hat{N}_{gender}$
Female	279.63	1 118.52	111.85	1 510
Male	161.96	1 166.09	161.96	1 490
$\hat{N}_{age}$	441.59	2 284.61	273.81	3 000

Raking-ratio adjustment (cont.)

Example

- Now multiply the column estimated counts by 600/441.59, 2 120/2 284.61 and 280/273.81, respectively:

	15-24	25-54	55+	$\hat{N}_{gender}$
Female	379.94	1 037.93	114.38	1 532.25
Male	220.06	1 082.07	165.62	1 467.75
$\hat{N}_{age}$	600.00	2 120.00	280.00	3 000.00

Raking-ratio adjustment (cont.)

- Repeat the procedure until both row and column totals agree with the population counts

Example

- The final estimated cell counts:

	15-24	25-54	55+	$\hat{N}_{gender}$
Female	375.59	1 021.47	112.94	1 510
Male	224.41	1 098.53	167.06	1 490
$\hat{N}_{age}$	600.00	2 120.00	280.00	3 000

The **raking weight adjustment factor**, f.exp., for female aged 25-54: $w_i = 1\,021.47/1\,200 = 0.8512$

- Additional **assumption** to those for post-stratification: the **response propensities** only depend on the row & column totals

Imputation

- Values are assigned to the missing items: y_i^* if $r_i = 0$
- **Deterministic** imputation: if y_i^* **fixed** given s and $s_r = \{i \in s : r_i = 1\}$
- **Random** imputation: if y_i^* **random** given s and $s_r = \{i \in s : r_i = 1\}$
- The **distribution** of the y_i values preserved with random imputation
- However, random imputation causes **extra variation** in addition to sampling and response mechanisms

Imputation methods

- **Deductive imputation**: imputation during the data editing; uses **logical relations** among the survey variables
- **Cell (class)-mean imputation**: similar to **weighting class adjustment**; same under SRSWOR: $y_i^* = \bar{y}_{rc}$ with $i \in s_c$
- **Hot-deck imputation**: imputation made using the same dataset
  - **Random hot-deck**: a **donor** randomly chosen for i, with $r_i = 0$, from the respondents in the same class as i
  - **Nearest-Neighbour hot-deck**: define a **distance measure** between observations, and choose a **donor** that is closest to the sample unit with the missing items
  - Use the same **donor** for all the missing items to preserve multivariate relationships among the survey variables

Imputation methods (cont.)

- **Regression imputation**: prediction of missing values using a **regression model**: $y_i^* = \mathbf{x}_i^\top \hat{\beta}$, with $\hat{\beta}$ estimated from $\{(\mathbf{x}_i, y_i) : i \in s_r\}$
  - **Stochastic regression imputation**: adding random draws of **residuals** from $\{\hat{\epsilon}_i : i \in s_r\}$ to the predicted values: $y_i^* = \mathbf{x}_i^\top \hat{\beta} + \hat{\epsilon}_i$
- **Cold-deck imputation**: imputation made using a different dataset: previous survey, historical data, etc.
- **Multiple imputation**: **independent** random (*stochastic*) imputation $B \geq 2$ times (see Rubin (1987) for details)

Remarks

- **Imputation** creates a 'full' data set
- **Consistent** results by analysing different subsets of the data
- Substantial decrease in **nonresponse bias** due to the item nonresponse, if MAR given the covariates used in the imputation procedure
- Treating the imputed data set as if the 'original' data leads to **underestimation** of the variance
- **Keep records** of the imputed observations and the donor observations

After the survey

- Compute **response rates** in a uniform fashion over time
- Monitor **composition** of causes of nonresponse over time: refusals, not-at-homes, out-of-scopes, address not locatable, post-master returns
- **Transparency** of dissemination, relevance to quality: publish response rate components in survey reports and provide definitions of response rates used
- Encourage **research** for remedies

Session 10: Variance estimation in complex surveys

Complex surveys and complex parameters

- Variance formulas under **SRS** relatively simple
- Formulas become more **complicated** under multi-stage sampling designs without replacement with unequal probabilities
- **Nonresponse and imputation** bring additional terms in estimation of the variance
- Under sampling with unequal probabilities, variance estimates available for estimators of population **totals**
- Often we also interested in **other quantities**: ratios, proportions, percentages, quantiles
- How to estimate the variance of a ratio under an unequal probability sampling?

Parameter: linear combination of totals

- Suppose the parameter of interest is a **linear combination** of population totals
- Let $\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_k$ be the **sample estimates** of population totals $t_1, t_2, \cdots, t_k$, with $\hat{t}_j = \sum_{i \in s} d_i y_{ij}$.
- Define $q_i = \sum_{j=1}^{k} a_j y_{ij}$ as a linear combination of the y_{ij} for any constants a_{ij}
- A linear combination of the t_j, $t_q = \sum_{i \in U} q_i = \sum_{j=1}^{k} a_j t_j$, estimated by

$$\hat{t}_q = \sum_{i \in s} d_i q_i = \sum_{j=1}^{k} a_j \hat{t}_j$$

Example I

- The **variance** of $\hat{t}_q$ estimated by:

$$\widehat{V}(\hat{t}_q) = \sum_{j=1}^{k} a_j^2 \widehat{V}(\hat{t}_j) + \sum_{j=1}^{k} \sum_{j\prime \neq j}^{k} a_j a_{j\prime} \widehat{\mathrm{Cov}}(\hat{t}_j, \hat{t}_{j\prime})$$

Example

Norwegian Labour Force Survey (LFS): Quarterly estimates of the total number of employed and unemployed people computed as a weighted average of the monthly estimates:

$$\hat{t}_Q = \sum_{j=1}^{3} a_{Qj} \hat{t}_{Qj}$$

Example I (cont.)

Example

Norwegian Labour Force Survey (LFS): The a_{Qj} proportional to the number of survey weeks in month j in quarter Q with $\sum_{j=1}^{3} a_{Qj} = 1$. We have $a_{Qj} = 4/13$ and $a_{Qj} = 5/13$ for months with four and five survey weeks, respectively.

$$\widehat{V}(\hat{t}_Q) = \sum_{j}^{k} a_{Qj}^2 \widehat{V}(\hat{t}_{Qj}),$$

as $\text{Cov}(\hat{t}_{Qj}, \hat{t}_{Qj\prime}) = 0$, for $j \neq j\prime$, due to the independence between monthly samples

Example II

Example

Periodical changes: monthly, quarterly, yearly changes often of interest in official statistics

The change in the population total t from time $t = 1$ to time $t = 2$, i.e. $\Delta = t_2 - t_1$, estimated by

$$\hat{\Delta} = \hat{t}_2 - \hat{t}_1$$

leading to $a_2 = 1$ and $a_1 = -1$.

$$\widehat{V}(\hat{\Delta}) = \widehat{V}(\hat{t}_1) + \widehat{V}(\hat{t}_2) - 2\,\widehat{\mathrm{Cov}}(\hat{t}_1, \hat{t}_2)$$

Parameter: non-linear function of totals

- Non-linear statistics, f.exp., ratios, **linearised** using the method of **Taylor series expansion**
- Let $\mathrm{f}(\mathbf{t})$, with $\mathbf{t} = (t_1, \cdots, t_k)^\top$, be a non-linear function of the population totals $t_1, \cdots, t_k$.
- The **first-order Taylor series** approximation of $\mathrm{f}(\hat{\mathbf{t}})$ given by

$$\mathrm{f}(\hat{\mathbf{t}}) \approx \mathrm{f}(\mathbf{t}) + \mathrm{f}'(\mathbf{t})^\top (\hat{\mathbf{t}} - \mathbf{t})$$

- $\mathrm{f}'(\mathbf{t}) = \frac{\partial \mathrm{f}(\hat{\mathbf{t}})}{\partial \hat{\mathbf{t}}}|_{\hat{\mathbf{t}}=\mathbf{t}}$: the vector of **partial derivatives** of $\mathrm{f}(\hat{\mathbf{t}})$ with respect to $\hat{\mathbf{t}}$ evaluated at $\hat{\mathbf{t}} = \mathbf{t}$

$$\widehat{V}[\mathrm{f}(\hat{\mathbf{t}})] \approx \mathrm{f}'(\hat{\mathbf{t}})^\top \widehat{\mathrm{Cov}}(\hat{\mathbf{t}}) \mathrm{f}'(\hat{\mathbf{t}})$$

Example: ratio estimator

Example

Consider the **ratio estimator** of the population total: $\hat{t}_{yr} = \widehat{B} t_x$, with $\widehat{B} = \hat{t}_y / \hat{t}_x$. $\widehat{V}(\hat{t}_{yr}) = t_x^2 \widehat{V}(\widehat{B})$. By **Taylor**'s theorem:

$$\begin{aligned}
\widehat{B} - B &\approx \left(\frac{\partial \hat{t}_{yr}}{\partial \hat{t}_y}\Big|_{\hat{t}_y = t_y} \right) (\hat{t}_y - t_y) + \left(\frac{\partial \hat{t}_{yr}}{\partial \hat{t}_x}\Big|_{\hat{t}_x = t_x} \right) (\hat{t}_x - t_x) \\
&= \frac{1}{t_x}(\hat{t}_y - t_y) - \frac{t_y}{t_x^2}(\hat{t}_x - t_x) \\
&= \frac{1}{t_x}(\hat{t}_y - B\hat{t}_x)
\end{aligned}$$

Replacing population quantities with their estimators:

$$\widehat{B} - B \approx \frac{1}{\hat{t}_x} \sum_{i \in s} d_i (y_i - \widehat{B} x_i) = \frac{1}{\hat{t}_x} \sum_{i \in s} d_i \hat{e}_i = \sum_{i \in s} d_i q_i$$

Example: ratio estimator (cont.)

Example

- *Estimated* variance of $\hat{t}_{yr}$:

$$\widehat{V}(\hat{t}_{yr}) = t_x^2 \widehat{V}(\widehat{B}) = t_x^2 \widehat{V}(\hat{t}_q) = \left(\frac{t_x}{\hat{t}_x}\right)^2 \widehat{V}(\hat{t}_e)$$

- If t_x known, variance **alternatively** estimated by:

$$\widehat{V}_{alt}(\hat{t}_{yr}) = \widehat{V}(\hat{t}_e)$$

- For large samples, $\widehat{V}(\hat{t}_{yr}) \approx \widehat{V}_{alt}(\hat{t}_{yr})$.
- For small samples, $\widehat{V}(\hat{t}_{yr})$ better than $\widehat{V}_{alt}(\hat{t}_{yr})$ in many situations
- Under **SRSWOR**: $\widehat{V}(\hat{t}_{yr}) = N^2(1-f)(\bar{X}/\bar{x})^2 s_e^2/n$

Pros & Cons

- Well **developed** theory
- Applicable in general sampling designs
- Many statistical **software** (SAS, r, Stata, SPSS) compute linearised variance estimates
- **Messy calculation** of complex functions
- **Analytical** expressions of partial derivatives may not always exist; thus, one must calculate them numerically
- Some statistics may not expressed as **smooth** functions of populations totals, such as, the median or other quantiles
- **Sample size** has to be **large** enough for **accuracy** of the linearised variance estimation

Jackknife

- Widely used in sample surveys
- The Jackknife variance estimator **asymptotically (for large n) unbiased** for smooth functions of totals
- May be **computationally cumbersome** for some complex sampling designs and with large datasets
- Let θ be the parameter of interest, given by a function of population totals, and $\widehat{\theta}$ be its estimator
- Let $\widehat{\theta}_{-i}$ be an estimator of the same form as $\widehat{\theta}$, but not using the ith unit

Delete-1 jackknife

- For an **SRS**, the **customary jackknife** (*delete-1 jackknife*) estimator of variance (Tukey 1958):

$$\widehat{V}_{\mathrm{JK}}(\widehat{\theta}) = \frac{n-1}{n} \sum_{i \in s} \left(\widehat{\theta}_{-i} - \frac{1}{n} \sum_{k \in s} \widehat{\theta}_{-k} \right)^2$$

- Under **SRSWR**, $\widehat{V}_{\mathrm{JK}}(\widehat{\theta})$ **consistent** (*asymptotically unbiased*) for $V(\widehat{\theta})$
- $\widehat{V}_{\mathrm{JK}}(\widehat{\theta})$ **biased** for **SRSWOR** with large sampling fraction
- Bias reduced when $\widehat{V}_{\mathrm{JK}}(\widehat{\theta})$ multiplied by the **fpc**:

$$\widehat{V}^*_{\mathrm{JK}}(\widehat{\theta}) = (1-f)\widehat{V}_{\mathrm{JK}}(\widehat{\theta}), \quad f = n/N$$

Example

Example

Consider the jackknife variance estimation of the ratio: $\widehat{B} = \bar{y}/\bar{x}$ under **SRSWOR**. Here, $\widehat{\theta} = \bar{y}/\bar{x}$ and $\widehat{\theta}_{-i} = \widehat{B}_{-i} = \bar{y}_{-i}/\bar{x}_{-i}$

$$\widehat{V}_{\mathrm{JK}}(\widehat{B}) = \frac{n-1}{n} \sum_{j \in s} \left(\widehat{B}_{-i} - \frac{1}{n} \sum_{k \in s} \widehat{B}_{-k}\right)^2$$

i	x_i	y_i	$\bar{x}_{-i}$	$\bar{y}_{-i}$	$\widehat{B}_{-i}$
1	14	37	18.8	32.6	1.7340
2	18	50	18.0	30.0	1.6667
3	15	15	18.6	37.0	1.9892
4	11	22	19.4	35.6	1.8351
5	19	25	17.8	35.0	1.9663
6	31	51	15.4	29.8	1.9351

Example (cont.)

Example

For example, $\bar{x}_{-2}$ is the average of all the x_i except x_1; that is, $\bar{x}_{-2} = (14 + 15 + 11 + 19 + 31)/5 = 18$. We have $\sum_{k \in s} \widehat{B}_{-k}/6 = 1.8544$.

$$\widehat{V}_{\mathrm{JK}}(\widehat{B}) = \frac{6-1}{6} \sum_{i \in s} (\widehat{B}_{-i} - 1.8544)^2 = 0.07276$$

Stratified sampling

- Under **stratified** sampling, the jackknife applied separately in each stratum
- For **stratified SRSWOR**, the following jackknife estimator can be used:

$$\widehat{V}_{\mathrm{JK}}(\widehat{\theta}_{str}) = \sum_{h=1}^{H}(1-f_h)\frac{n_h-1}{n_h}\sum_{i\in s_h}\left(\widehat{\theta}_{(-hi)} - \frac{1}{n_h}\sum_{k\in s_h}\widehat{\theta}_{(-hk)}\right)^2$$

- $\widehat{\theta}_{(-hi)}$ calculated using the weights

$$w_{k(-hi)} = \begin{cases} 0 & \text{if} \quad k = i \\ \frac{n_h}{n_h-1}d_k & \text{if} \quad k \in s_h,\ j \neq i \\ d_k & \text{if} \quad k \notin s_h \end{cases}$$

Stratified TCS

- For **stratified two-stage cluster** sampling, deletion of units happens at the **PSU level**. Once a PSU deleted, all the units belonging to that PSU omitted
- Suppose n_h **PSU**s selected with **SRSWR**
- The **jackknife variance estimator** under **stratified TCS** given by

$$\widehat{V}_{\text{JK}}(\widehat{\theta}_{strtcs}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{i \in s_h} \left(\widehat{\theta}_{(-hi)} - \frac{1}{n_h} \sum_{k \in s_h} \widehat{\theta}_{(-hk)} \right)^2$$

- $\widehat{\theta}_{(-hi)}$: the estimator of the same form as $\widehat{\theta}_{strtcs}$ when PSU i in stratum h omitted

Stratified TCS (cont.)

- $\widehat{\theta}_{(-hi)}$ calculated using the adjusted weights of the **SSU**s:

$$w_{j(-hi)} = \begin{cases} 0 & \text{if SSU } j \text{ is in PSU } i \text{ of stratum } h \\ \frac{n_h}{n_h - 1} d_j & \text{if SSU } j \text{ is in stratum } h \text{ but not in PSU } i \\ d_j & \text{if SSU } j \text{ is not in stratum } h \end{cases}$$

- The jackknife applicable to **unequal probability** sampling (e.g. Campbell 1987, Berger 2007)
- The effect of **imputation** can be taken into account with the jackknife (e.g. Rao and Shao 1992)

Naïve bootstrap

- The **naïve bootstrap** (Efron 1979), which may be called the **Monte Carlo bootstrap**, requires that the observations in the **original sample** s *independently and identically distributed* (**i.i.d.**), the case when the original sample selected with, f.exp., **SRSWR**
- Suppose s an **SRSWR** of size n
- Draw B **sub-samples** of sizes n with replacement from the original sample s
- Typically, $B = 500$ or $B = 1000$, although there may be cases where $B = 100$ would be sufficient (e.g. Wu and Rao 1988)
- For each sub-sample, calculate $\widehat{\theta}_b^*$, for the population parameter θ

Bootstrap variance estimator

- The **Monte Carlo bootstrap** estimator if the variance:

$$\widehat{V}_{\text{boot}}(\widehat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}_b^* - \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_b^*\right)^2.$$

- Efron (1982) suggested to select sub-samples of size $n-1$ for **asymptotically unbiased** estimation of the variance
- If the original sample selected **SRSWOR**, $\widehat{V}_{\text{boot}}$ will overestimate the variance when $f = n/N$ **large**
- One can create a **pseudo-population** by generating N/n **replicates (copies)** of the original sample. Then, B sub-samples taken with SRSWOR from the *pseudo-population*

Stratified TCS

- Suppose n_h **PSU**s selected with **SRSWR**
- The **rescaling bootstrap** method (Rao and Wu 1992) can be used for **stratified two-stage sampling**
- For each **bootstrap replicate** apply the following steps:

1. Select an **SRS** of $n_h - 1$ PSUs with replacement from the n_h sample PSUs in stratum h
2. Calculate the **rescaling weights**:

$$w^*_{hij} = \frac{n_h}{n_h - 1} m_{hi} d_{hij},$$

where m_{hi} the number of times the ith PSU in stratum h selected in a bootstrap replicate and d_{hij} the sampling weight of the jth SSU in PSU i of stratum h

3. Calculate $\widehat{\theta}^*_b$ using the weights w^*_{hij}

Remarks

- The **bootstrap** variance estimator can be used for sampling **WOR** when $f = n/N$ negligible
- Bootstrap applicable to both **smooth** and **non-smooth** functions of totals, unlike the delete-1 jackknife
- Bootstrap sometimes more **computationally intensive** than the jackknife
- 95% CI can be directly calculated by taking the 2.5th and 97.5 **percentiles** from the distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$
- The **standard CI** based on the **normality assumption** can be constructed for linearisation and the resampling methods if $\widehat{\theta}$ a **smooth** function and its distribution **approximately normal**, achieved with **large** sample size

Thank you for your attention ☺