European Commission

EUROPEAN MASTER IN OFFICIAL STATISTICS

EMOS

Welcome to EMOS Webinar
26 April 2017
16.30-18.00

Introduction to survey sampling

Ralf Münnich
University of Trier
Economic and Social Statistics

EMOS Webinar — Introduction to Survey Sampling

Ralf Münnich
Economic and Social Statistics
Trier University

Business of statistics

One major aim (business) of statistics, especially for official statistics, is gathering and provision of data for administrative, political, sociological, economic, and other research purposes. This information is obtained in universes (or samples) of units (persons, households, establishments, machinery etc.).
The data are gained in total (complete inventory) or via pre-specified principles partially as samples (randomly, quota). The following shall be considered:

- Primary or secondary statistics (e.g. register data)
- Problems of adequacy
- Errors of different kinds
- Data protection (*anonymisation*)

Surveys and Samples

Definition of a *survey*

A survey, in general, is the methodology for gathering information via samples on persons, households, or other units. In a survey

- planning and development,
- pretest,
- survey design,
- implementation,
- data gathering, editing and processing, as well as
- analysis

play a vital role.

http://www.whatisasurvey.info/

https://www.whatisasurvey.info/downloads/pamphlet_current.pdf

Surveys and Samples

Definition of a *survey*

A survey, in general, is the methodology for gathering information via samples on persons, households, or other units. In a survey

- planning and development,
- pretest,
- survey design,
- implementation,
- data gathering, editing and processing, as well as
- analysis

play a vital role.

http://www.whatisasurvey.info/

https://www.whatisasurvey.info/downloads/pamphlet_current.pdf

Areas of applications in survey statistics

- Economics statistics and empirical economic research
- Social statistics and empirical social research
- Demography and socio-demographic research
- Market and opinion research
- Quality control
- Medical statistics and biometry
- Meteorology
- Environmental statistiscs
- Forestry and environmental control
- Traffic control
- Inventory based on samples
- Natural science research and measurement

Data gathering in surveys

Different kinds of data

- Cross sectional or longitudinal data (point of time, period)
  - Random samples
  - Systematic samples
- Time series
- Panel data (cross sectional and longitudinal)

Data acquisition

- (Computer assisted) interviewing
- Telephone interviewing
- Use of register data
- Snowball sampling
- Quota sampling

Examples for important surveys

EU-SILC: Statistics on Income and Living Conditions in Europe. `http://ec.europa.eu/eurostat/web/income-and-living-conditions/data/database`).

ESS: European Social Survey
The European Social Survey (the ESS) is an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. Now in its third round, the survey covers over 20 nations and employs the most rigorous methodologies. It is funded via the European Commission's 5th and 6th Framework Programmes, the European Science Foundation and national funding bodies in each country. Source: `http://www.europeansocialsurvey.org/`

The German Microcensus (MC)

- Survey of households via interviewer
  - Structure of the population
  - ILO unemployment and labour participation
  - *Special aspects* (changing)
- 1% sample since 1957
- Reform 1990
- Microcensus law from 17. January 1996
  Programme of questions, mandatory survey, max. 4 years
- Since 2005: short-term sampling
- Microcensus as *rotational panel*

`http://www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/index.htm`

Microcensus law 2005:

`http://www.destatis.de/download/d/stat_ges/bevoe/054a.pdf`

Arrangement of statistical units

214 regional classes Federal states, districts

Regional classes: min. 200.000 to 250.000 inh.

5 house size classes Census 1987 and statistics on building activity

1 Small buildings: 1 to 4 dwellings

2 Mid size buildings: 5 to 10 dwellings

3 Large buildings: minimum 11 dwellings

4 Institutions: dw. $=0$ or indiv. $\geq (\text{dw.} + 4) \cdot 4$

6 New buildings

Selection domain sequential arrangement

1 Approx. 12 dwellings ($\sim 10 - 13$); maximal 70 inhabitants

2 One building each

3 6 – 9 dwellings; floor-wise selection

4 Initial letter of last name; approx. 15 inhabitants

6 Selection according 1 – 4 depending on the type of new building

Sampling design of the Microcensus

- *4 -stage* design
  Strata – strata – strata – clusters
- 1% sample
  One selection domain from each zone
  Approx. 1% of individuals and households
  Semi-systematic random selection
- 4 zones $\longrightarrow$ block (rotation quarters)
  Replacement of 1 (4) rotation quarters each year

Note: The dropped rotation quarter yields the basis for acquiring households for the access panel (EU-SILC and ICT).
Further: The Microcensus will be rearranged as core sample for the new European integrated household survey system, while having integrated the LFS and SILC.

Data quality: Eurostat definition

Relevance of the statistical concept:
End-user, *user needs*, hierarchical structure and contents

Accuracy and reliability:

- Sampling errors: standard error, CI coverage
- Non-sampling errors: nonresponse, coverage error, measurement errors

Timeliness and punctuality: Time and duration from data acquisition until publication

Coherence and comparability: Preliminary and final statistics, annual and intermediate statistics (regions, domains, time)

Accessibility and clarity: Publication of data, analysis and method reports

`http://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice`

Errors and types of errors in surveys

- Frame error
- Interviewer error
- Processing error
- Coding error
- Nonresponse (NR)
  - Unit-Nonresponse
  - Item-Nonresponse
- Sampling error

Evaluation of samples and surveys

Main principles in survey sampling

Practicability

Costs of a survey

Accuracy of results

- Standard errors
- Confidence interval coverage
- Disparity of subpopulations

Robustness of results

In order to adequately evaluate the estimates from samples, *appropriate* evaluation criteria have to be considered.

Basic principles in statistics for surveys

In addition to the analysis from survey data, we need to investigate the quality of the results. The aim is to generalize results from the sample to the universe.
$\longrightarrow$ inferential statistics

Sampling function and estimator

Properties of estimators

- unbiasedness
- efficiency
- *normality*
- large sample properties (limit theorems)
- small sample properties

Test procedures based on survey data

Details can be drawn from Schaich and Münnich (2001), Chapter 4 or 6, Mittelhammer (2013), or others.

Unemployment in Saarland

Unemployed		14 – 24	25 – 44	45 – 64	65 +	$\sum$
Women	τ	2.387	7.248	4.686	128	14.449
Men	τ	4.172	9.504	10.588	0	24.264
$\sum$	τ	6.559	16.752	15.274	128	38.713

- *True* values in Saarland
- Estimates from the Microcensus
- Is the quality of the cell estimates identical?

Unemployment in Saarland

Unemployed		14 – 24	25 – 44	45 – 64	65 +	$\sum$
Women	τ	2.387	7.248	4.686	128	14.449
	$\mathrm{E}\,\widehat{\tau}$	2.387	7.238	4.684	128	14.436
Men	τ	4.172	9.504	10.588	0	24.264
	$\mathrm{E}\,\widehat{\tau}$	4.172	9.505	10.598	0	24.275
$\sum$	τ	6.559	16.752	15.274	128	38.713
	$\mathrm{E}\,\widehat{\tau}$	6.558	16.743	15.282	128	38.711

- *True* values in Saarland
- Estimates from the Microcensus
- Is the quality of the cell estimates identical?

Unemployment in Saarland

Unemployed		14 – 24	25 – 44	45 – 64	65 +	$\sum$
Women	τ	2.387	7.248	4.686	128	14.449
	$E\,\widehat{\tau}$	2.387	7.238	4.684	128	14.436
Men	τ	4.172	9.504	10.588	0	24.264
	$E\,\widehat{\tau}$	4.172	9.505	10.598	0	24.275
$\sum$	τ	6.559	16.752	15.274	128	38.713
	$E\,\widehat{\tau}$	6.558	16.743	15.282	128	38.711

- *True* values in Saarland
- Estimates from the Microcensus
- Is the quality of the cell estimates identical?

Unemployment in Saarland

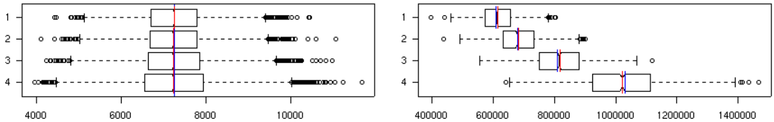
Unemployed		14 – 24	25 – 44	45 – 64	65 +	$\sum$
Women	τ	2.387	7.248	4.686	128	14.449
	$\mathrm{E}\,\widehat{\tau}$	2.387	7.238	4.684	128	14.436
Men	τ	4.172	9.504	10.588	0	24.264
	$\mathrm{E}\,\widehat{\tau}$	4.172	9.505	10.598	0	24.275
$\sum$	τ	6.559	16.752	15.274	128	38.713
	$\mathrm{E}\,\widehat{\tau}$	6.558	16.743	15.282	128	38.711

- *True* values in Saarland
- Estimates from the Microcensus
- Is the quality of the cell estimates identical?

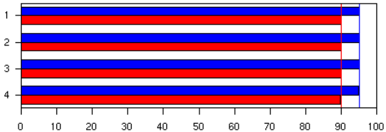
Unemployed women, 25 – 44

Raking estimator

Variance estimator

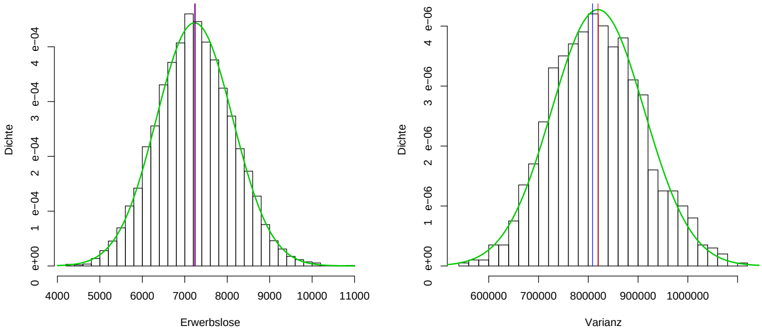


NR rates: 1: 5%, 2: 10%, 3: 25%, 4: 40%



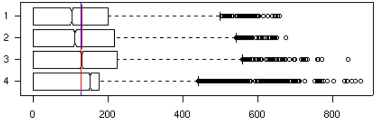
95% 90%

Unemployed women, 25 – 44, distribution of point and Variance estimator (25% NR)

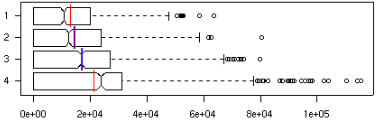


Unemployed women, 65 +

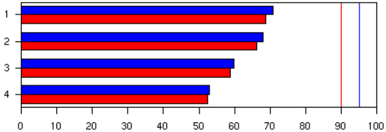
Raking estimator



variance estimator

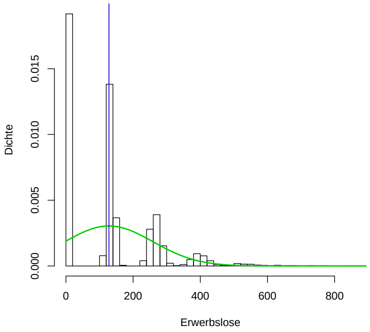


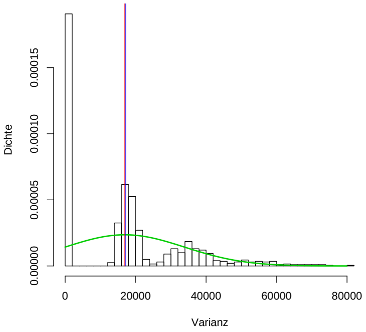
NR rates: 1: 5%, 2: 10%, 3: 25%, 4: 40%



95% 90%

Unemployed women, 65 +, distribution of point and variance estimator (25% NR)





Summary questions for Chapter 1

- How would you *evaluate* survey estimates?
- And what has to be considered using simulations for an evaluation?
- In the last example, normality was not achieved. Is there any opportunity to avoid the lack in normality?
- The theory of SRS is well elaborated. Why should we then consider other sampling designs?

Definitions and notations

Universe (U)		Sample (S)
$\mathcal{U} = \{1, \ldots, N\}$		$\mathcal{S} = \{U_1, \ldots, U_n\}$
Characteristic of interest	Y	y_i
Auxiliary variable(s)	$\mathbf{X}$	$\mathbf{x}_i$
(i-th unit)		
Parameters of the universe		Estimates from the sample
Total	τ	$\widehat{\tau}$
Proportion	θ	$\widehat{\theta}$
Mean	μ	$\widehat{\mu}$
Ratio	$\psi = \frac{\tau_1}{\tau_2}$	$\widehat{\psi}$
General parameter	π	$\widehat{\pi}$

A distinction between random variables and their outcomes follows from the context.

Population parameters

Total $\tau_Y = Y = \sum_{i=1}^{N} y_i$

Mean $\mu_Y = \overline{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i$

Proportion $\theta_Y = P_Y = \frac{1}{N} \sum_{i=1}^{N} y_i$, where $y_i \in \{0; 1\} \forall i$

Variance $\sigma_Y^2 = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \overline{Y}\right)^2$

Variance (dichotomous variable) $\sigma_Y^2 = P_Y \cdot (1 - P_Y)$

Sample values

Total $y = \sum_{i=1}^{n} y_i$

Mean $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Proportion $p_Y = \frac{1}{n} \sum_{i=1}^{n} y_i$, where $y_i \in \{0; 1\} \forall i$

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

Variance (dichotomous variable) $s_y^2 = p_Y \cdot (1 - p_Y)$

Sample selection scheme

A sample $\mathcal{S}$ consists of elements from the universe $\mathcal{U} = \{1, 2, \ldots, N\}$:

Sampling without replacement: The sample is a subset of the universe.

Sampling with replacement: The sample consists of an index set with indices $1, \ldots, N$, in which elements may appear several times.

Definition 1.1

Let $\mathcal{U}$ be a finite population with N elements. A sample $\mathcal{S}$ is generated with the help of a sample selection scheme that selects elements from the universe. The set of all possible samples with respect to the sample selection scheme is denoted by $\mathbb{S}$. The probability of drawing a pre-specified sample j is $P(\mathcal{S}_j)$.

Example 2.1 (cf. Lohr, 1999, p. 25)

A universe $\mathcal{U}$ consists of $N = 4$ elements. Sampling with replacement with a sample size $n = 2$ yields:

$$\mathcal{S}_1 = \{1, 2\} \qquad \mathcal{S}_2 = \{1, 3\} \qquad \mathcal{S}_3 = \{1, 4\}$$
$$\mathcal{S}_4 = \{2, 3\} \qquad \mathcal{S}_5 = \{2, 4\} \qquad \mathcal{S}_6 = \{3, 4\}$$

1. All samples are drawn with equal probability.
2. The probabilities for the samples are: $P(\mathcal{S}_1) = 1/3$, $P(\mathcal{S}_2) = 1/6$, $P(\mathcal{S}_6) = 1/2$, $P(\mathcal{S}_3) = P(\mathcal{S}_4) = P(\mathcal{S}_5) = 0$.
3. The probabilities for the samples are: $P(\mathcal{S}_1) = 1/3$, $P(\mathcal{S}_2) = 1/6$, $P(\mathcal{S}_4) = 1/2$, $P(\mathcal{S}_3) = P(\mathcal{S}_5) = P(\mathcal{S}_6) = 0$.

What is the sample selection probability of a pre-specified element from the universe? How do the different sample selection probabilities affect the *inclusion probabilities*?

First order inclusion probabilities

The probability π_i for an element i from the universe to be included in a sample is called inclusion probability:

$$\pi_i = \sum_{j=1}^{|\mathbb{S}|} \mathcal{I}(i \in \mathcal{S}_j) \cdot P(\mathcal{S}_j) \quad .$$

The statistic $\widehat{\pi}$ to estimate the parameter π can be assessed from the set of possible samples $\mathbb{S}$ (which may be tedious in practice). We get the (design-based) expected value $E(\widehat{\pi})$ from

$$E(\widehat{\pi}) = \sum_{i=1}^{|\mathbb{S}|} \widehat{\pi}(\mathcal{S}_j) \cdot P(\mathcal{S}_j)$$

The same procedure yields the (design-based) variance and MSE.
$\longrightarrow$ Designs with unequal probabilities.

Examples for selection schemes

SRS Simple random sampling with replacement

SRSWOR Simple random sampling without replacement

BERN Bernoulli-Sampling (each element of the universe will be selected with probability θ)

SYS Systematic sampling
One element c from $1, ..., a \ll N$ is drawn. Consequently the elements $i \cdot a + c$ will be selected.

Note:

1. SRS and SRSWOR yield fixed sample sizes n
2. BERN yields a random sample size with $E(n) = \theta \cdot N$.

Examples for selection schemes

SRS Simple random sampling with replacement

SRSWOR Simple random sampling without replacement

BERN Bernoulli-Sampling (each element of the universe will be selected with probability θ)

SYS Systematic sampling
One element c from $1, ..., a \ll N$ is drawn. Consequently the elements $i \cdot a + c$ will be selected.

Note:

1. SRS and SRSWOR yield fixed sample sizes n
2. BERN yields a random sample size with $E(n) = \theta \cdot N$.

Simple random sampling

Definition of a random sample

A random process in which successively n elements are selected from a finite universe with N elements ($n < N$) is called random sample selection. The result from a random sample selection is called random sample.

Simple random sample

Drawing samples from an urn with or without replacement is called simple random sampling. The outcome is a simple random sample.

With or without replacement sampling

- N^n many different simple random samples with replacement can be drawn. Each of them is drawn with identical probability. Each element can be drawn 0 to n times. The draws are stochastically independent.
- $N!/(N-n)!$ many different simple random samples without replacement can be drawn. Each of them is drawn with identical probability. Each element can be drawn 0 or 1 times. The draws are stochastically dependent.
- The first order inclusion probabilities are

$$\pi_i = \frac{n}{N} \qquad \text{(SRSWOR)}$$

$$\pi_i = 1 - \left(1 - \frac{1}{N}\right)^n \qquad \text{(SRSWR)}$$

(small sample fractions: little difference).

The sample mean $\widehat{\mu}$ for simple random sampling

1. $E(\widehat{\mu}) = \mu$ (WR/WOR)
2. $V(\widehat{\mu}) = \dfrac{\sigma^2}{n}$ (WR) and $V(\widehat{\mu}) = \dfrac{\sigma^2}{n} \cdot \dfrac{N-n}{N-1}$ (WOR) resp.
3. If $U \sim N(\mu; \sigma^2)$ (WR): $\widehat{\mu} \sim N(\mu; \dfrac{\sigma^2}{n})$
4. If U is normal (WR): $\dfrac{\widehat{\mu}-\mu}{S}\sqrt{n}$ with $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is t_{n-1} distributed.
5. Without loss of generality for large n

   $\frac{\widehat{\mu}-\mu}{\sigma}\sqrt{n}$; $\quad \frac{\widehat{\mu}-\mu}{\sigma\sqrt{\frac{N-n}{N-1}}}\sqrt{n}$ $\quad (N-n$ large); $\quad \frac{\widehat{\mu}-\mu}{S}\sqrt{n}$; $\quad \frac{\widehat{\mu}-\mu}{S\sqrt{\frac{N-n}{N-1}}}\sqrt{n}$

   are approximately standard normal.

The estimator $\widehat{\mu}$ is

- unbiased for μ (WR/WOR)
- consistent for μ (only WR is relevant)
- is BLUE for μ
- is ML estimator for μ if U is normally distributed (WR)

Similar results are derived for totals $\tau = N \cdot \mu$ via the corresponding estimator $\widehat{\tau} = N \cdot \widehat{\mu}$.

For SRSWOR, the random variables are slightly negatively correlated:

$$Cov X_i; X_j = -\frac{\sigma^2}{N-1}$$

Confidence intervals for the mean (WR)

CI for μ; U normally distributed; σ^2 known:

$$\left[\widehat{\mu} - z(1-\alpha/2)\cdot\frac{\sigma}{\sqrt{n}};\ \widehat{\mu} + z(1-\alpha/2)\cdot\frac{\sigma}{\sqrt{n}}\right]$$

CI for μ; U normally distributed:

$$\left[\widehat{\mu} - t(1-\alpha/2; n-1)\cdot\frac{s}{\sqrt{n}};\ \widehat{\mu} + t(1-\alpha/2; n-1)\cdot\frac{s}{\sqrt{n}}\right]$$

General scheme of confidence intervals:

$$\left[\widehat{\mu} - z(1-\alpha/2)\cdot\sqrt{\widehat{\mathrm{V}}(\widehat{\mu})};\ \widehat{\mu} + z(1-\alpha/2)\cdot\sqrt{\widehat{\mathrm{V}}(\widehat{\mu})}\right]$$

given the normality assumption holds!

Beware of skewed variables and outliers!

Necessary sample size I

CI estimation for μ with confidence level $(1-\alpha)$; e is given; WR; no distributional assumptions are made.
e is the half CI length for *central* CIs; this leads to

$$e = z(1-\alpha/2) \cdot \frac{\sigma}{\sqrt{n}} \quad .$$

With given prior information $\hat{\sigma}$ for σ we get

$$\sqrt{n} = z(1-\alpha/2) \cdot \frac{\hat{\sigma}}{e},$$

and finally

$$n \geq (z(1-\alpha/2))^2 \cdot \frac{\hat{\sigma}^2}{e^2} \qquad .$$

Necessary sample size II

As before, but now considering WOR sampling:

Analogously we have

$$e = z(1-\alpha/2) \cdot \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad .$$

With $(N-1 \approx N)$ we get

$$n = (z(1-\alpha/2))^2 \cdot \frac{\hat{\sigma}^2}{e^2} \cdot \frac{N-n}{N}$$

$$\Leftrightarrow \qquad n(1 + \frac{1}{N}(z(1-\alpha/2))^2 \cdot \frac{\hat{\sigma}^2}{e^2}) \geq (z(1-\alpha/2))^2 \cdot \frac{\hat{\sigma}^2}{e^2} \quad ,$$

and finally

$$n \geq \frac{(z(1-\alpha/2))^2 \cdot N \cdot \hat{\sigma}^2}{(z(1-\alpha/2))^2 \cdot \hat{\sigma}^2 + Ne^2} \quad .$$

SRS and beyond ...

Stratified random sampling

The universe $\mathcal{U}$ is decomposed into L pairwise disjoint non-empty subpopulations (primary sampling units) $G_1, \ldots, G_L$:

$$G_{q_1} \cap G_{q_2} = \emptyset \quad \text{for all } q_1, q_2 = 1, \ldots, L; q_1 \neq q_2 \text{ and } \bigcup_{q=1}^{L} G_q = G$$

In stratified random sampling the subpopulations are called strata. For the qth stratum we get:

$$\mu_q = \frac{1}{N_q} \sum_{i=1}^{Nq} y_{qi} \qquad \text{mean of stratum } q$$

$$\sigma_q^2 = \frac{1}{N_q} \sum_{i=1}^{Nq} (y_{qi} - \mu_q)^2 \qquad \text{variance of stratum } q$$

Sampling from stratified populations is called stratified random sampling (StrRS).

For the universe $\mathcal{U}$ holds ($\gamma_q := N_q/N$):

$$\mu = \sum_{q=1}^{L} \gamma_q \mu_q \quad ; \quad \tau = N \cdot \mu = \sum_{q=1}^{L} N_q \cdot \mu_q$$

$$\theta = \sum_{q=1}^{L} \gamma_q \theta_q$$

$$\sigma^2 = \sum_{q=1}^{L} \gamma_q \sigma_q^2 + \sum_{q=1}^{L} \gamma_q (\mu_q - \mu)^2 = \sigma_w^2 + \sigma_b^2$$

$$\sigma_e^2 = \frac{1}{L} \sum_{q=1}^{L} (N_q \mu_q - \frac{N\mu}{L})^2 = \frac{1}{L} \sum_{q=1}^{L} N_q^2 \mu_q^2 - \frac{N^2 \mu^2}{L^2}$$

$$= \frac{1}{L} \sum_{q=1}^{L} N_q^2 \mu_q^2 - (\frac{1}{L} \sum_{q=1}^{L} N_q \mu_q)^2.$$

Parameters in the sample

Sample allocation

Given a decomposition of the universe $\mathcal{U}$ into L primary sampling units, the breakdown of the total sample size n into L stratum-specific sample sizes $n_1, \ldots, n_q, \ldots, n_L$ with $\sum_q n_q = n$ is called *allocation*.

For the qth primary sampling unit we get

$$\bar{y}_q = \frac{1}{n_q} \sum_{i=1}^{n_q} y_{qi} \qquad \text{stratum mean}$$

$$s_q^2 = \frac{1}{n_q - 1} \sum_{i=1}^{n_q} (y_{qi} - \bar{y}_q)^2 \qquad \text{stratum variance}$$

Estimating means, totals, and proportions in StrRS

Sample mean:

$$\widehat{\mu}_{\text{StrRS}} = \sum_{q=1}^{L} \gamma_q \widehat{\mu}_q = \sum_{q=1}^{L} \gamma_q \frac{1}{n_q} \cdot \sum_{i=1}^{n_q} y_{iq} \overset{!}{=} \widehat{\mu}_{\text{StrRSWOR}}$$

The point estimates do not differ from WR and WOR in StrRS.

Lemma 2.1

The estimator $\widehat{\mu}_{\text{StrRS}}$ is unbiased for μ (WR and WOR). The variance of the estimator $\widehat{\mu}$ is:

$$V(\widehat{\mu}_{\text{StrRS}}) = \sum_{q=1}^{L} \gamma_q^2 \cdot \frac{\sigma_q^2}{n_q} \qquad \text{(WR)}$$

$$V(\widehat{\mu}_{\text{StrRSWOR}}) = \sum_{q=1}^{L} \gamma_q^2 \cdot \frac{\sigma_q^2}{n_q} \cdot \frac{N_q - n_q}{N_q - 1} \qquad \text{(WOR)}$$

- Since $E(S_q^2) = \sigma_q^2$, we get

$$\widehat{V}(\widehat{\mu}_{\text{StrRS}}) = s_{\widehat{\mu}}^2 = \sum_{q=1}^{L} \gamma_q^2 \cdot \frac{s_q^2}{n_q} \qquad \text{(WR)}$$

as an unbiased estimate for $V(\widehat{\mu}_{\text{StrRS}})$.

- Since $E(S_q^2) = \frac{N}{N-1} \cdot \sigma_q^2$,

$$\widehat{V}(\widehat{\mu}_{\text{StrRSWOR}}) = s_{\widehat{\mu}}^2 = \sum_{q=1}^{L} \gamma_q^2 \cdot \frac{s_q^2}{n_q} \cdot \frac{N_q - n_q}{N_q} \qquad \text{(WOR)}$$

is an unbiased estimate for $V(\widehat{\mu}_{\text{StrRSWOR}})$.

- In case of *large* stratum-specific sample sizes n_q, the estimator $\widehat{\mu}_{\text{StrRS}}$ is approximately normal (CLT). This yields the $(1-\alpha) \cdot 100\%$ - CI for μ:

$$\left[\widehat{\mu}_{\text{StrRS}} - z(1-\alpha/2) \cdot \sqrt{\widehat{V}(\widehat{\mu}_{\text{StrRS}})}; \widehat{\mu}_{\text{StrRS}} + z(1-\alpha/2) \cdot \sqrt{\widehat{V}(\widehat{\mu}_{\text{StrRS}})}\right]$$

Example 2.2

A universe is split into 4 strata with stratum sizes 10000, 40000, 30000 and 20000. Drawing a stratified random sample yields the following table:

Stratum	n_q	$\bar{y}_q$	s_q
1	50	40.5	9.8
2	200	60.8	7.5
3	150	70.3	6.3
4	100	31.9	11.8

a) An unbiased estimate for μ is derived as

$$\widehat{\mu}_{\text{StrRS}} = \sum \gamma_q \bar{y}_q = \frac{10}{100} \cdot 40.5 + \cdots + \frac{20}{100} \cdot 31.9 \approx 55.84 \qquad .$$

Example 2.2 (continued)

b) In sampling WR we get

$$\begin{aligned}\widehat{V}(\widehat{\mu}_{\text{StrRS}}) &= \sum_{q=1}^{L} \gamma_q^2 \cdot \frac{s_q^2}{n_q} \\ &= 0.1^2 \cdot \frac{9.8^2}{50} + 0.4^2 \cdot \frac{7.5^2}{200} + 0.3^2 \cdot \frac{6.3^2}{150} + 0.2^2 \cdot \frac{11.8^2}{100} \\ &\approx 0.143718\end{aligned}$$

and thus $\sqrt{\widehat{V}(\widehat{\mu}_{\text{StrRS}})} \approx 0{,}3791$. The 95%-CI for μ is then:

$$\left[55.84 - 1.96 \cdot 0.3791; 55.84 + 1.96 \cdot 0.3791\right] = [55.09; 56.59]$$

In the case of small sample fractions n_q/N_q $(q = 1, \ldots, L)$, sampling WOR yields (approximately) the same figures.

Example 2.3 (cf. Example 2.2)

The sample results remain now out of consideration. We assume known true stratum-specific variances $\sigma_1^2 = 100$, $\sigma_2^2 = 55$, $\sigma_3^2 = 40$ and $\sigma_4^2 = 150$. Calculate $V(\mu_{\mathrm{StrRS}})$ under WR

a) using the allocation given in example 2.2 and
b) with stratum-specific sample sizes $n_1 = 60$, $n_2 = 180$, $n_3 = 110$, and $n_4 = 150$.
Further, comment on the results.

In a) we get

$$V(\widehat{\mu}_{\mathrm{StrRS}}) = 0.1^2 \cdot \frac{100}{50} + 0.4^2 \cdot \frac{55}{200} + 0.3^2 \cdot \frac{40}{150} + 0.2^2 \cdot \frac{150}{100} \approx 0.148$$

and for b)

$$V(\widehat{\mu}_{\mathrm{StrRS}}) = 0.1^2 \cdot \frac{100}{60} + 0.4^2 \cdot \frac{55}{180} + 0.3^2 \cdot \frac{40}{110} + 0.2^2 \cdot \frac{150}{150} \approx 0.138\,.$$

2.4.1 Equal allocation

$$n_q = \frac{n}{L} \qquad q = 1, \ldots, L$$

$$\widehat{\mu}_{\text{StrRS,eq}} = \frac{1}{L} \sum_{q=1}^{L} \bar{y}_q$$

$$V(\widehat{\mu}_{\text{StrRS, eq}}) = \frac{L}{n} \sum_{q=1}^{L} \gamma_q^2 \cdot \sigma_q^2 \qquad \text{(WR)}$$

$$V(\widehat{\mu}_{\text{StrRS, eq}}) = \frac{L}{n} \sum_{q=1}^{L} \gamma_q^2 \cdot \sigma_q^2 \cdot \frac{N_q - n_q}{N_q - 1} \qquad \text{(WOR)}$$

2.4.2 Proportional allocation (Bowley)

$$n_q = \gamma_q \cdot n \qquad q = 1, \dots, L$$

$$\widehat{\mu}_{\text{StrRS, prop}} = \frac{1}{n} \sum_{q=1}^{L} n_q \cdot \bar{y}_q$$

$$V(\widehat{\mu}_{\text{StrRS, prop}}) = \frac{1}{n} \sum_{q=1}^{L} \gamma_q \cdot \sigma_q^2 \qquad \text{(WR)}$$

$$V(\widehat{\mu}_{\text{StrRSWOR, prop}}) = \frac{1}{n} \sum_{q=1}^{L} \gamma_q \cdot \sigma_q^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{N_q}{N_q - 1} \qquad \text{(WOR)}$$

$$= \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \sum_{q=1}^{L} \gamma_q \cdot \sigma_q^2 \qquad (N_q \approx N_q - 1)$$

2.4.3 Optimal allocation (Neyman-Tschuprov)

$$n_q = n \cdot \frac{\gamma_q \sigma_q}{\sum\limits_{h=1}^{L} \gamma_h \sigma_h} = n \cdot \frac{N_q \sigma_q}{\sum\limits_{h=1}^{L} N_h \sigma_h} \quad q = 1, \ldots, L$$

$$V(\widehat{\mu}_{\text{StrRS, opt}}) = \frac{1}{n}\left(\sum_{q=1}^{L} \gamma_q \sigma_q\right)^2 \qquad \text{(WR)}$$

$$V(\widehat{\mu}_{\text{StrRS, opt}}) = \frac{1}{n}\left(\sum_{q=1}^{L} \gamma_q \sigma_q\right)^2 - \frac{1}{N}\sum_{q=1}^{L} \gamma_q \sigma_q^2 \text{(WOR)}$$

Example 2.4

N_q	10000	40000	30000	20000
σ_q^2	100	55	40	150

Given a total sample size of $n = 500$, we calculate the Neyman-Tschuprov allocation via

$$\sum_{q=1}^{L} \gamma_q \sigma_q = 0.1 \cdot 10 + 0.4 \cdot \sqrt{55} + 0.3 \cdot \sqrt{40} + 0.2 \cdot \sqrt{150} \approx 8.3135$$

as $n_1 = 60,\ n_2 = 179,\ n_3 = 114$ and $n_4 = 147.$ The variance estimate $\widehat{\mu}$ results in (WR)

$$\begin{aligned} V(\widehat{\mu}_{\text{StrRS, opt}}) &= 0.1^2 \cdot \frac{100}{60} + 0.4^2 \cdot \frac{55}{179} + 0.3^2 \cdot \frac{40}{114} + 0.2^2 \cdot \frac{150}{147} \\ &\approx 0.13822 \end{aligned}$$

This variance is smaller than the variance in Example 2.3!

Problems in deriving the optimal allocation

- In sampling WOR one should consider the WOR correction:

$$n_q = n \cdot \frac{N_q \sigma_q \cdot \sqrt{\frac{N_q}{N_q - 1}}}{\sum_{h=1}^{L} N_h \sigma_h \cdot \sqrt{\frac{N_h}{N_h - 1}}}$$

- In sampling WOR, an *overallocation* may result. In practice, the corresponding stratum is sampled completely and the remaining sample size will be reallocated optimally. Theory: box-constrained optimal allocation (later).
- Theoretically, the optimal allocation is an integer-optimization problem under constraints. Rounding may result in small errors since the sum of the rounded sample sizes is not equal to the total sample size.

2.4.4 Cost-optimal allocation (Yates-Zacopanay)

$$n_q = \frac{\gamma_q \sigma_q \Big/ \sqrt{c_q}}{\sum\limits_{h=1}^{L} \gamma_h \sigma_h \sqrt{c_h}} \cdot (C - C_0) \qquad q = 1, \ldots, L$$

Subject to a linear cost function!

$$V(\widehat{\mu}_{\text{StrRS, YZ}}) = \frac{1}{C - C_0} \Big(\sum_{q=1}^{L} \gamma_q \sigma_q \sqrt{c_q} \Big)^2 \qquad \text{(WR)}$$

$$V(\widehat{\mu}_{\text{StrRS, YZ}}) = \frac{1}{C - C_0} \Big(\sum_{q=1}^{L} \gamma_q \sigma_q \sqrt{c_q} \Big)^2 - \frac{1}{N} \sum_{q=1}^{L} \gamma_q \sigma_q^2 \qquad \text{(WOR)}$$

Example 2.5

q	N_q	σ_q^2	C_q
1	10000	100	25
2	40000	55	3
3	30000	40	2
4	20000	150	36

A universe is split into $L = 4$ strata according to the table alongside. Calculate the cost-optimal allocation given $C_0 = 2000$ and $C = 10000$.

We calculate

$$n_1 = 8000 \cdot \frac{0.2}{5 + 5.1381 + 2.6833 + 14.6969} = \frac{1600}{27.5183} \approx 58$$

$$n_2 = \frac{8000}{27.5183} \cdot 1.7127 \approx 498$$

$$n_3 = \frac{8000}{27.5183} \cdot 1.3416 \approx 390$$

$$n_4 = \frac{8000}{27.5183} \cdot 0.4082 \approx 119$$

The total sample size results in $n = 1065$.

Problems in deriving the cost-optimal allocation

- Again, an overallocation may result in case of sampling WOR.
- As in the case of the optimal allocation, we have a integer-valued optimization under constraints.
- Rounding may lead to inadequate sample sizes that are (slightly) too expensive.

Accuracy comparisons I

$$V(\widehat{\mu}_{\text{SRS}}) = \frac{\sigma^2}{n} = \frac{1}{n} \cdot (\sigma_w^2 + \sigma_b^2) = \frac{1}{n} \cdot \Big(\sum_{q=1}^{L} \gamma_q \cdot \sigma_q^2 + \sigma_b^2 \Big)$$

$$= V(\widehat{\mu}_{\text{StrRS, prop}}) + \frac{1}{n} \cdot \sigma_b^2$$

$$\Longrightarrow \qquad V(\widehat{\mu}_{\text{SRS}}) \geq V(\widehat{\mu}_{\text{StrRS, prop}})$$

After multiplying out we get

$$\sum_{q=1}^{L} \gamma_q \cdot \Big(\sigma_q - \sum_{\iota=1}^{L} \gamma_\iota \sigma_\iota \Big)^2 = \underbrace{\sum_{q=1}^{L} \gamma_q \sigma_q^2}_{n \cdot V(\widehat{\mu}_{\text{StrRS, prop}})} - \underbrace{\Big(\sum_{q=1}^{L} \gamma_q \sigma_q \Big)^2}_{n \cdot V(\widehat{\mu}_{\text{StrRS, opt}})}$$

and finally $V(\widehat{\mu}_{\text{StrRS, prop}}) \geq V(\widehat{\mu}_{\text{StrRS, opt}})$

Accuracy comparisons II

Given the sample size n the following holds:

$$V(\widehat{\mu}_{\text{StrRS, opt}}) \leq V(\widehat{\mu}_{\text{StrRS, prop}}) \leq V(\widehat{\mu}_{\text{SRS}})$$

Note:

1. The more homogeneous strata are, the higher is the gain in efficiency by using StrRS instead of SRS. This results from σ_w^2 (variance within) being considerably small in contrast to σ_b^2 (variance between). This is called the effect of stratification.
2. The more heterogeneous the stratum-specific variances are, the higher is the gain in efficiency while applying the optimal rather than the proportional allocation.

Box-constrained optimal allocation

Aim: Minimise 2-norm of the RRMSE-vector:

$$||\mathbf{RRMSE}_{<\cdot>}(\widehat{\tau})||_2 = \sqrt[2]{\sum_{g=1}^{G} \mathrm{RRMSE}(\tau_{<g>})^2}$$

subject to:

- Lower and upper sampling fractions are given in each stratum
- Maximum number of sampling units

Solution via a specialised *box-constraints optimization* algorithm with *small area extension*:

Exact: Gabler, Ganninger and Münnich (2012), Metrika
Numerical: Münnich, Sachs and Wagner (2012), AStA
Integer: Friedrich, Münnich, de Vries und Wagner (2015), CSDA

Stratification

Stratification is the unique division of the universe into strata. Two problems arise:

1. The number of strata is given;
2. The stratum boundaries are given.

- Aim: Gain in efficiency of the estimator
- Problem: Prior information on universe level
  - Official Statistics
  - Sampling inventory
- Note: In many cases the number of strata is given in practice.

Stratification

Stratification is the unique division of the universe into strata. Two problems arise:

1. The number of strata is given;
2. The stratum boundaries are given.

- Aim: Gain in efficiency of the estimator
- Problem: Prior information on universe level
  - Official Statistics
  - Sampling inventory
- Note: In many cases the number of strata is given in practice.

Dalenius stratification

- Continuous variable of interest Y with density $f(y)$;
- Number of strata L is given;
- Variable of interest is used for stratification;
- Strata are non-overlapping (division of $\mathbb{D}_Y$)

Stratification problem: The $L-1$ inner stratum boundaries ξ_q $(q = 1, \ldots, L-1)$ are of interest, such that

$$V(\widehat{\mu}_{\text{StrRS}}; \xi_1, \ldots, \xi_{L-1}) \longrightarrow \min \quad .$$

Under proportional allocation we get:

$$V(\widehat{\mu}_{\text{StrRS}}; \xi_1, \ldots, \xi_{L-1}) = \frac{1}{n} \sum_{q=1}^{L} \gamma_q(\xi_1, \ldots, \xi_{L-1}) \cdot \sigma_q^2(\xi_1, \ldots, \xi_{L-1})$$

$$\longrightarrow \min_{\xi_1, \ldots, \xi_{L-1}} \qquad \text{(cf. Münnich, 1997, pp. 78)} \qquad .$$

Approximate solution

cum $\sqrt{f}$ rule (Dalenius and Hodges, 1957)

The stratum boundaries will be set such that the cumulative values $\sqrt{f} \cdot (\xi_q - \xi_{q-1})$ are approximately equal while applying the optimal allocation.

equal aggregate σ rule (Wright, 1983)

The stratum boundaries will be set such that $\sum_{i \in \mathcal{I}_q} \sigma_i = \frac{1}{L} \sum_{i=1}^{N} \sigma_i$ approximately holds. $\mathcal{I}_q$ is the set of all indices in stratum q and σ_i the individual standard deviation which are expected to be monotonous. This model-based approach uses the equal allocation (which is optimal for the variable of interest assuming the existence of an auxiliary variable).

Example 2.5

Class over	Class until	rel. freq.	$\sqrt{\text{rel. freq.}}$	cum.
0	2	9.63	3.10	3.10
2	4	7.72	2.78	5.88
4	6	6.58	2.57	8.45
6	8	6.54	2.56	11.00
8	10	5.49	2.34	13.35
10	12	5.46	2.34	15.68
12	14	6.83	2.61	18.30
14	16	6.16	2.48	20.78
16	18	6.52	2.55	23.33
18	20	5.38	2.32	25.65
20	22	4.69	2.17	27.82
22	24	4.54	2.13	29.95
24	26	4.32	2.08	32.03
26	28	3.88	1.97	34.00
28	30	3.50	1.87	35.87
30	32	3.40	1.84	37.71
32	34	3.33	1.82	39.54
34	36	3.21	1.79	41.33
36	38	1.94	1.39	42.72
38	40	0.88	0.94	43.66

Empirical frequencies of the variable of interest (20 classes) with $L = 5$ strata according to the Dalenius-Hodges approximation.

Here, we have 43.66 / 5 = 873.
The stratum boundaries can be allocated at 8,73; 17,64; 26,19; and 34,92.

In case the class widths differ from each other, the computation has to be corrected.

Example 2.6

20 units were observed in a survey (Y is the variable of interest):

i	$\sigma_{I,i}$	$\sum \sigma_{I,i}$	$\sigma_{II,i}$	$\sum \sigma_{II,i}$
1	0.04	0.04	0.00	0.00
2	0.13	0.17	0.02	0.02
3	0.44	0.61	0.19	0.21
4	0.62	1.23	0.38	0.60
5	0.62	1.85	0.38	0.98
6	0.74	2.59	0.55	1.53
7	0.86	3.45	0.74	2.27
8	1.37	4.82	1.88	4.15
9	1.43	6.25	2.04	6.19
10	1.82	8.07	3.31	9.50
11	2.05	10.12	4.20	13.70
12	2.08	12.20	4.33	18.03
13	2.51	14.71	6.30	24.33
14	3.03	17.74	9.18	33.51
15	3.54	21.28	12.53	46.04
16	3.66	24.94	13.40	59.44
17	4.26	29.20	18.15	77.59
18	4.92	34.12	24.21	101.79
19	5.55	39.67	30.80	132.60
20	6.04	45.71	36.48	169.08

The aim is to build 4 strata. The assumptions $\sigma_I \propto y$ and $\sigma_{II} \propto y^2$ are to be considered as true.

We get 45.71 and 169.08 as *aggregate* σ. This yields the stratum boundaries 11.43, 22.86 and 34.28 as well as 42.27, 84.54 and 126.81 respectively.

Note: the information was obtained from extremely few observations (teaching example!). The equal allocation will be applied.

Poststratification

The stratification is used *after* drawing a simple random sample (Lohr, 1999)

- SRS yields approximately equal proportions with respect to a stratification
- Correction of over- and under-representation of the categories (strata)
- Variance formulae of the proportional allocation are applied
- Poststratification is applied when the necessary information for stratification is not available in the sampling frame but from other sources (e.g. register).
  Example: Nationality (German, non-German) and gender
- Note: Stressing poststratification with many variables may lead to erroneous results

Example 2.7

A simple random sample yields:

	n	$\overline{y}$	s^2
W	52	544	112^2
M	48	618	124^2

The proportion of women in the universe is 60%. Hence:

SRS $\mu_{\text{SRS}} = \frac{1}{100} \cdot (52 \cdot 544 + 48 \cdot 618) = 579.52$

$$\widehat{V}(\widehat{\mu}_{\text{SRS}}) = \frac{1}{100} \cdot \frac{100}{99} \cdot \left(\frac{51 \cdot 112^2 + 52 \cdot 544^2 + 47 \cdot 124^2 + 48 \cdot 618^2}{100} - 579.52^2 \right) = 151.42$$

PostStr $\mu_{\text{PostStr}} = 0.6 \cdot 544 + 0.4 \cdot 618 = 573.6$

$$\widehat{V}(\widehat{\mu}_{\text{PostStr}}) = \sum_{q=1}^{L} \gamma_q^2 \cdot \frac{s_q^2}{n_q} = 138.10$$

The variability of the sample size in the denominator of $\widehat{V}(\widehat{\mu}_{\text{PostStr}})$ was not considered (may be ignored in practice if not too low)!

Example 2.7

A simple random sample yields:

	n	$\bar{y}$	s^2
W	52	544	112^2
M	48	618	124^2

The proportion of women in the universe is 60%. Hence:

SRS $\mu_{\text{SRS}} = \frac{1}{100} \cdot (52 \cdot 544 + 48 \cdot 618) = 579.52$

$$\widehat{V}(\widehat{\mu}_{\text{SRS}}) =$$
$$\frac{1}{100} \cdot \frac{100}{99} \cdot \left(\frac{51 \cdot 112^2 + 52 \cdot 544^2 + 47 \cdot 124^2 + 48 \cdot 618^2}{100} - 579.52^2 \right) = 151.42$$

PostStr $\mu_{\text{PostStr}} = 0.6 \cdot 544 + 0.4 \cdot 618 = 573.6$

$$\widehat{V}(\widehat{\mu}_{\text{PostStr}}) = \sum_{q=1}^{L} \gamma_q^2 \cdot \frac{s_q^2}{n_q} = 138.10$$

The variability of the sample size in the denominator of $\widehat{V}(\widehat{\mu}_{\text{PostStr}})$ was not considered (may be ignored in practice if not too low)!

Example 2.7

A simple random sample yields:

	n	$\bar{y}$	s^2
W	52	544	112^2
M	48	618	124^2

The proportion of women in the universe is 60%. Hence:

SRS $\mu_{\text{SRS}} = \frac{1}{100} \cdot (52 \cdot 544 + 48 \cdot 618) = 579.52$

$$\widehat{V}(\widehat{\mu}_{\text{SRS}}) =$$
$$\frac{1}{100} \cdot \frac{100}{99} \cdot \left(\frac{51 \cdot 112^2 + 52 \cdot 544^2 + 47 \cdot 124^2 + 48 \cdot 618^2}{100} - 579.52^2 \right) = 151.42$$

PostStr $\mu_{\text{PostStr}} = 0.6 \cdot 544 + 0.4 \cdot 618 = 573.6$

$$\widehat{V}(\widehat{\mu}_{\text{PostStr}}) = \sum_{q=1}^{L} \gamma_q^2 \cdot \frac{s_q^2}{n_q} = 138.10$$

The variability of the sample size in the denominator of $\widehat{V}(\widehat{\mu}_{\text{PostStr}})$ was not considered (may be ignored in practice if not too low)!

Further remarks on stratified random sampling

- In general, stratification is set up as regards content (cf. microcensus)
- Stratification and estimation variable have to be different (cf. Dalenius)
- Multidimensional allocation (cf. Schaich and Münnich, 1993): Possible conflict of targets while applying the optimal allocation $\longrightarrow$ decision problem
  Exact: Friedrich, Münnich and Rupp (submitted)
  Uncertain: Nomani, Burgard, Dürr, Münnich (in submission)
- Number of strata:
  - Theory: L high ($\rightarrow$ variance reduction)
    integer valued solution!
  - Practice: 5 - 8 strata
  - High sensitivity to data (and estimators)

Single stage cluster sampling

A universe is split into L primary sampling units (PSU; clusters). The selection of PSUs will follow SRS at the first stage; all secondary sampling units (SSU) in the selected PSUs (indicated by a superscript s) are sampled totally at the second stage.

- At the first-stage, l PSUs are selected by SRSWOR (single stage cluster sampling: SIC).
- The total sample size of sampling units is random.

$$\widehat{\mu}_{\mathsf{SIC}} = \frac{L}{l}\sum_{q=1}^{l}\gamma_q^{\mathsf{sel}}\cdot\mu_q^{\mathsf{sel}} = \frac{L}{l}\sum_{q=1}^{l}\frac{N_q^{\mathsf{sel}}}{\sum\limits_{r=1}^{L}N_r}\cdot\frac{1}{N_q^{\mathsf{sel}}}\sum_{i=1}^{N_q^{\mathsf{sel}}}Y_{iq} \quad .$$

is an unbiased estimator for the mean μ in the universe (here: $\mu_q^{\mathsf{sel}} = \widehat{\mu}_q^{\mathsf{sel}}$)

Variance of SIC

Single stage cluster sampling equals random sampling without replacement at the first stage; hence, the observed values can easily be summed up. The resulting *survey* is equivalent to stratified random sampling. It follows:

$$V(\widehat{\mu}_{\mathrm{SIC}}) = \frac{L^2}{N^2} \cdot \frac{\sigma_e^2}{l} \cdot \frac{L-l}{L-1} \qquad ,$$

which can be (asymptotically unbiasedly) estimated via

$$\widehat{V}(\widehat{\mu}_{\mathrm{SIC}}) = \frac{L^2}{N^2} \cdot \frac{s_e^2}{l} \cdot \frac{L-l}{L} \qquad ,$$

where

$$s_e^2 = \frac{1}{l-1} \cdot \sum_{r=1}^{l} \left(N_r^{\mathrm{sel}} \mu_r^{\mathrm{sel}} - \frac{N \cdot \widehat{\mu}_{\mathrm{SIC}}}{L} \right)^2 \qquad .$$

Estimation of totals under SIC

We have $\widehat{\tau}_{\text{SIC}} = N \cdot \widehat{\mu}_{\text{SIC}}$ and hence

$$V(\widehat{\tau}_{\text{SIC}}) = L^2 \cdot \frac{\sigma_e^2}{l} \cdot \frac{L-l}{L-1} \qquad \text{or} \qquad \widehat{V}(\widehat{\tau}_{\text{SIC}}) = L^2 \cdot \frac{s_e^2}{l} \cdot \frac{L-l}{L}$$

respectively.

Estimation of proportions under SIC

Here, we have $\mu_q^r = p_q^r$. Then, the relation of interest follows from

$$\frac{\sigma_{e;\theta}^2}{N^2} = \frac{1}{L} \sum_{q=1}^{L} (\gamma_q \theta_q - \frac{\theta}{L})^2 \quad .$$

Accuracy comparisons I

In general, we assume ($I \ll L$):

$$V(\widehat{\mu}_{\text{SRS}}) \leq V(\widehat{\mu}_{\text{SIC}})$$

This yields

$$V(\widehat{\mu}_{\text{SIC}}) \approx \frac{1}{I} \cdot \sigma_b^2 = \frac{1}{I}(\sigma^2 - \sigma_w^2)$$

1. If $\sigma_b^2 \approx 0$ (and hence $\sigma_w^2 = \sigma^2$), the variance $V(\widehat{\mu}_{\text{SIC}})$ becomes considerably small. SIC is then approximately equivalent to SRSWOR.
2. In contrast, if $\sigma_w^2 \approx 0$, the variance $V(\widehat{\mu}_{\text{SIC}})$ may become very large.

This effect is called cluster effect.

Accuracy comparisons II

Whenever the PSUs are of approximately the same size ($N_q \equiv \overline{N}$), one can show that

$$V(\widehat{\mu}_{\text{SIC}}) = \left(1 - \frac{l}{L}\right) \cdot \frac{\overline{N}}{L-1} \frac{\text{SSB}}{l} \qquad \text{, where}$$

$$\text{SSTOT} = \sum_{q=1}^{L} \sum_{i=1}^{N_q} (Y_{qi} - \overline{Y}_q)^2 + \sum_{q=1}^{L} N_q \cdot (\overline{Y}_q - \overline{Y})^2 \quad = \text{SSW} + \text{SSB} \quad .$$

This leads to the intraclass correlation coefficient (ICC)

$$\text{ICC} = 1 - \frac{\overline{N}}{\overline{N}-1} \cdot \frac{\text{SSW}}{\text{SSTOT}}$$

and finally to

$$V(\widehat{\mu}_{\text{SIC}}) = \frac{L\overline{N}-1}{\overline{N}(L-1)} \cdot \left(1 + (\overline{N}-1) \cdot \text{ICC}\right) \cdot V(\widehat{\mu}_{\text{SRSWOR}})$$

Accuracy comparisons III

- Applying WR (SRS and SIC on the first stage) yields

$$V(\widehat{\mu}_{\text{SIC}}) = \left(1 + (\overline{N} - 1) \cdot \text{ICC}\right) \cdot V(\mu_{\text{SRS}})$$

- In general, the cluster effect (using SIC instead of SRSWOR) yields a loss in efficiency. We have $-\frac{1}{\overline{N}-1} \leq \text{ICC} \leq 1$.
- The ratio of the variances

$$\text{Deff} = \frac{V(\widehat{\mu}_{\text{SIC}})}{V(\widehat{\mu}_{\text{SRSWOR}})} = \frac{L\overline{N} - 1}{\overline{N}(L-1)} \cdot \left(1 + (\overline{N} - 1) \cdot \text{ICC}\right)$$

  is called design effect of SIC related to SRSWOR while estimating the mean of the universe. SRSWOR is used as a reference design.
- Systematic random sampling can be viewed as a special case of SIC with $l = 1$.

Definitions in two-stage cluster sampling (TSC)

The first stage, total, and second stage sampling fractions are denoted by $f_I = \frac{l}{L}$, $f_{II} = \frac{n}{N} = \sum_{r=1}^{l} n_r^{\text{sel}} \Big/ \sum_{q=1}^{L} N_q$ and $f_r = \frac{n_r^{\text{sel}}}{N_r^{\text{sel}}}$ respectively.

The following drawing schemes are considered $(r = 1, \ldots, l)$:

- $n_r^{\text{sel}} = \frac{n}{l}$
- $n_r^{\text{sel}} = \dfrac{N_r^{\text{sel}}}{\sum\limits_{\kappa=1}^{l} N_\kappa^{\text{sel}}} \cdot n$
- $n_r^{\text{sel}} = f_r \cdot N_r^{\text{sel}}$ with f_r constant

Estimation of means in TSC

In TSC, the mean estimator

$$\widehat{\mu}_{\mathrm{TSC}} = \frac{L}{l} \sum_{q=1}^{l} \gamma_q^{\mathrm{sel}} \cdot \widehat{\mu}_q^{\mathrm{sel}}$$

for WR (first stage selection) and WR/WOR (second stage selection) is an unbiased estimator for μ for all three sampling schemes. Further, the variance is given by (WR)

$$V(\widehat{\mu}_{\mathrm{TSC}}) = \frac{1}{N^2} \cdot \left(L^2 \cdot \frac{\sigma_e^2}{l} \cdot \frac{L-l}{L-1} + \frac{L}{l} \sum_{q=1}^{L} N_q^2 \cdot \frac{\sigma_q^2}{n_q} \right)$$

and (WOR)

$$V(\widehat{\mu}_{\mathrm{TSCWOR}}) = \frac{1}{N^2} \cdot \left(L^2 \cdot \frac{\sigma_e^2}{l} \cdot \frac{L-l}{L-1} + \frac{L}{l} \sum_{q=1}^{L} N_q^2 \cdot \frac{\sigma_q^2}{n_q} \cdot \frac{N_q - n_q}{N_q - 1} \right)$$

Variance estimation under TSC

The variance $V(\widehat{\mu}_{\text{TSC}})$ can be estimated as:

$$\widehat{V}(\widehat{\mu}_{\text{TSC}}) = \frac{1}{N^2} \cdot \left(L^2 \cdot \frac{s_e^2}{l} \cdot \frac{L-l}{L} + \frac{L}{l} \sum_{q=1}^{l} {N_q^{\text{sel}}}^2 \cdot \frac{{s_q^{\text{sel}}}^2}{n_q^{\text{sel}}} \right)$$

for WR at the second stage and

$$\widehat{V}(\widehat{\mu}_{\text{TSCWOR}}) = \frac{1}{N^2} \cdot \left(L^2 \cdot \frac{s_e^2}{l} \cdot \frac{L-l}{L} + \frac{L}{l} \sum_{q=1}^{l} {N_q^{\text{sel}}}^2 \cdot \frac{{s_q^{\text{sel}}}^2}{n_q^{\text{sel}}} \cdot \frac{N_q^{\text{sel}} - n_q^{\text{sel}}}{N_q^{\text{sel}}} \right)$$

WOR at the second stage respectively.
Estimating totals and proportions is analogous to SIC.

Example 2.9

A universe is split into $L = 8$ subpopulations with equal sizes ($N_q = 1000$, $q = 1, \dots, L$). Each subpopulation consists of two observations with the following values:

q	1	2	3	4	5	6	7	8
$Y_{q,1}$	10	20	16	25	20	30	28	30
$N_{q,1}$	500	400	500	400	600	600	800	200
$Y_{q,2}$	110	140	104	170	160	180	188	180
$N_{q,2}$	500	600	500	600	400	400	200	800

TSC has to be considered.

a) At the first stage $l = 4$ PSUs are selected. At the second stage, $n_r^{\text{sel}} = 100$ SSUs shall be selected from each selected PSU. Both stages assume WOR. Derive $V(\widehat{\mu}_{\text{TSCWOR}})$.

b) A sample yields PSUs 1, 3, 5, 8 with $\widehat{\mu}_r^{\text{sel}} = 58; 55.6; 83; 153$ and ${s_r^{\text{sel}}}^2 = 2521.21; 1936; 4900; 3354, 55$ respectively. Calculate the total estimate as well as its variance.

Accuracy comparisons of SRS and TSC

Given the assumptions

$$N_1 = \cdots = N_L = \overline{N} = \frac{N}{L} \tag{1}$$

and

$$n_{q_1}^{\mathrm{sel}} = \cdots = n_{q_l}^{\mathrm{sel}} = \frac{n}{l} = \overline{n} \tag{2}$$

as well as l/L and $\overline{n}/\overline{N}$ small leads to

$$V(\widehat{\mu}_{\mathrm{TSC}}) = V(\widehat{\mu}_{\mathrm{SRS}}) \cdot (1 + (\overline{n} - 1) \cdot \mathrm{ICC}) \quad ,$$

where

$$\mathrm{ICC} = \frac{1}{\sigma^2} \cdot \left(\sigma_b^2 - \frac{\sigma_w^2}{\overline{N} - 1}\right)$$

is the intraclass correlation coefficient.

Remark: (2) follows from (1) for all three selection schemes!

Example 2.10 (cf. example 2.9)

Calculate $V(\widehat{\mu}_{\text{SRSWOR}})$ with $n = 400$. The universe with $L = 8$ subpopulations is now split according to the following scheme:

q	1	2	3	4	5	6	7	8
$Y_{q,1}$	10	110	20	25	140	170	28	188
$N_{q,1}$	500	500	400	400	600	600	800	200
$Y_{q,2}$	16	104	30	20	160	180	30	180
$N_{q,2}$	500	500	600	600	400	400	200	800

The following values can be derived:

μ_q	13	107	26	22	148	174	28,4	181,6
$N^2\mu_q^2$ $[10^6]$	169	11449	676	484	21904	30276	806,56	32978,56
σ_q^2	9	9	24	6	96	24	0,64	10,24

Calculate again the variance $V(\widehat{\mu}_{\text{TSCWOR}})$.

Summary questions for Chapter 2

- What design would you prefer by what reason?
- Do we always have *one finite population of interest*?
- Where did we use auxiliary information?
- Could this use be extended?

3.1 Fundamental ideas of regression estimation

Example 3.1: After conducting a census, in the subsequent year, only a sample may be drawn. The information on the variable of interest is still available from the census but will also be available for the sample. Two cases occur:

- the census information is only available as totals (e.g. due to disclosure control reasons);
- unit identifiers allow to match sample and census information.

The auxiliary variable will be the target variable observed within the census (earlier time).

The main problem occurs due to entries and exits of elementary units.

Example 3.2: The livestock of pig breedings is surveyed by a complete inventory of all breedings via questionnaires. An additional sample inventory is performed on real count basis by specialists in order to evaluate the declaration error. The real count defines the target variable and the questionnaire counts the auxiliary variable.

See pig counts in Belgium by Heinrich Strecker and Rolf Wiegert.

Example 3.3: Sample inventory: The target variable consists of real values from all units in an inventory. As auxiliary values, the book values from an inventory management system could be taken, which are available as complete inventory.

One has to distinguish three models:

1. the difference estimator for additive models;
2. the ratio estimator for multiplicative models;
3. and the linear regression estimator for linear models.

Difference estimation

Assumptions:

- $\mu_X = \overline{x}$ is known;
- the pairs of variates $(x_i; y_i)$ are sampled by SRS or SRSWOR $(i = 1, \ldots, n)$.

The difference estimator is

$$\widehat{\mu}_{\text{Diff, SRS}} = \frac{1}{n}\sum_{i=1}^{n} y_i + B \cdot (\frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{n}\sum_{i=1}^{n} x_i) = \widehat{\mu}_Y + B \cdot (\mu_X - \widehat{\mu}_X)$$

where B is a an appropriate predetermined constant.
The difference estimator $\widehat{\mu}_{\text{Diff, SRS}}$ (and $\widehat{\mu}_{\text{Diff, SRSWOR}}$ respectively) is unbiased for μ_Y.
The total estimate is given by $\widehat{\tau}_{\text{Diff, SRS}} = N \cdot \widehat{\mu}_{\text{Diff, SRS}}$.

Variance of the difference estimator

The variance of the difference estimator is given by

$$\begin{aligned} V(\widehat{\mu}_{\text{Diff, SRS}}) &= V(\widehat{\mu}_{Y,\text{SRS}}) + B^2 V(\widehat{\mu}_{X,\text{SRS}}) - 2Cov(\widehat{\mu}_{Y,\text{SRS}}, B \cdot \widehat{\mu}_{X,\text{SRS}}) \\ &= \frac{1}{n} \cdot \left(\sigma_Y^2 + B^2 \cdot \sigma_X^2 - 2 \cdot B \cdot \sigma_{XY}\right) \end{aligned}$$

and

$$V(\widehat{\mu}_{\text{Diff, SRSWOR}}) = \frac{1}{n} \cdot \frac{N-n}{N-1} \cdot \left(\sigma_Y^2 + B^2 \cdot \sigma_X^2 - 2 \cdot B \cdot \sigma_{XY}\right) \quad .$$

The variance $V(\widehat{\mu}_{\text{Diff, SRS}})$ is minimal for $B = \frac{\sigma_{XY}}{\sigma_X^2}$ which equals the regression coefficient of the slope in the universe. This assignment yields for WR

$$V(\widehat{\mu}_{\text{Diff, SRSWR, min}}) = \frac{1}{n} \cdot \sigma_Y^2 \cdot \left(1 - \varrho_{XY}^2\right)$$

(WOR analogously).

Variance estimation for difference estimation

An unbiased estimate for $V(\widehat{\mu}_{\text{Diff, SRS}})$ is given by

$$\widehat{V}(\widehat{\mu}_{\text{Diff, SRS}}) = \frac{1}{n} \cdot \left(s_y^2 - 2 \cdot B \cdot s_{xy} + B^2 \cdot s_x^2 \right)$$

and for $V(\widehat{\mu}_{\text{Diff, SRSWOR}})$ by

$$\widehat{V}(\widehat{\mu}_{\text{Diff, SRSWOR}}) = \frac{1}{n} \cdot \left(s_y^2 - 2 \cdot B \cdot s_{xy} + B^2 \cdot s_x^2 \right) \cdot \frac{N - n}{N} \quad .$$

Remark:

In practice, B has to be determined appropriately (in many cases close to 1). Alternatively, B can be estimated from the sample which leads to linear regression estimation.

Example 3.4:

An inventory with $N = 16.000$ units is given. The book values are denoted with X and true values Y. The book values yielded $\mu_X = 80{,}5$ and $\sigma_X^2 = 8466$. From a sample with size $n = 400$ (WOR) the means $(\bar{x}, \bar{y}) = (81{,}2; 91{,}4)$ were gained.

a) The correlation coefficient between book and true values was assigned to $\varrho_{XY} = 0{,}95$; further, the assumption $\sigma_Y^2 = \sigma_X^2$ was made. Determine the pre-assigned value B under these conditions.

b) Calculate the estimate $\widehat{\mu}_{\text{Diff, SRSWOR}}$ for the mean of the true inventory values.

c) Further, from the sample the values $(s_x^2, s_y^2) = (8580; 8230)$ and $s_{xy} = 7950$ were gained. Determine the variance estimate $\widehat{V}(\widehat{\mu}_{\text{Diff, SRSWOR}})$.

Ratio estimation

The ratio estimator is given by

$$\widehat{\mu}_{\text{Ratio, SRS}} = \frac{\overline{y}}{\overline{x}} \cdot \mu_X = \frac{\widehat{\mu}_Y}{\widehat{\mu}_X} \cdot \mu_X =: \widehat{R} \cdot \mu_X \quad ,$$

where $r := \widehat{R} = \widehat{\tau}_Y / \widehat{\tau}_X$ for $R = \tau_Y / \tau_X = \mu_Y / \mu_X$.
The variance of $\widehat{\mu}_{\text{Ratio, SRSWOR}}$ is given by

$$V(\widehat{\mu}_{\text{Ratio, SRSWOR}}) = \frac{1}{n} \cdot \frac{N-n}{N-1} \cdot \left(\sigma_Y^2 + R^2 \cdot \sigma_X^2 - 2 \cdot R \cdot \sigma_{XY} \right)$$

and is estimated by

$$\widehat{V}(\widehat{\mu}_{\text{Ratio, SRSWOR}}) = \frac{1}{n} \cdot \frac{N-n}{N} \cdot \left(s_Y^2 + r^2 \cdot s_X^2 - 2 \cdot r \cdot s_{XY} \right) \quad ,$$

where $r := \widehat{R}$. In case of WR the finite population correction terms are removed.

Example 3.5 (cf. Example 3.4)

From

$$\widehat{R} = r = \frac{91{,}4}{81{,}2} = 1{,}1256$$

we get as ratio estimate

$$\widehat{\mu}_{\text{Ratio, SRSWOR}} = 1{,}1256 \cdot 80{,}5 = 90{,}6121$$

with variance estimate

$$\widehat{V}(\widehat{\mu}_{\text{Ratio, SRSWOR}}) = \frac{1}{400} \cdot \frac{15600}{16000} \cdot \Big(8230 + 1{,}1256^2 \cdot 8580 - 2 \cdot 1{,}1256 \cdot 7950\Big) = 2{,}9338 \quad .$$

The difference estimator yielded $\widehat{\mu}_{\text{Diff, SRS}} = 90{,}735$ and $\widehat{V}(\widehat{\mu}_{\text{Diff, SRSWOR}}) = 2{,}1168$.

Linear regression estimation

Estimating the coefficient B of the difference estimation via a linear regression model delivers the linear regression estimator. The estimate is achieved by applying the least squares (LS) method and results in

$$\widehat{B} = \frac{\widehat{\sigma}_{XY}}{\widehat{\sigma}_X^2} = \frac{s_{XY}}{s_X^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad .$$

This yields

$$\widehat{\mu}_{\text{Reg, SRS}} = \overline{y} + \widehat{B} \cdot \left(\overline{X} - \overline{x}\right) = \widehat{\mu}_{Y,SRS} + \widehat{B} \cdot \left(\mu_X - \widehat{\mu}_{X,SRS}\right)$$

as the (linear) regression estimator (SRSWOR analogously).

Properties of the regression estimator

- $\widehat{\mu}_{\text{Reg, SRS}}$ is asymptotically unbiased (SRS and SRSWOR);
- The variance of the regression estimator is given by

$$V\widehat{\mu}_{\text{Reg, SRS}} = \frac{1}{n} \cdot \sigma_Y^2 \cdot \left(1 - \varrho_{XY}^2\right) \qquad \text{or}$$

$$V\widehat{\mu}_{\text{Reg, SRSWOR}} = \frac{1}{n} \cdot \frac{N-n}{N-1} \cdot \sigma_Y^2 \cdot \left(1 - \varrho_{XY}^2\right)$$

- The variance can be estimated by

$$\widehat{V}\widehat{\mu}_{\text{Reg, SRS}} = \frac{1}{n} \cdot \left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right) \qquad \text{or}$$

$$\widehat{V}\widehat{\mu}_{\text{Reg, SRSWOR}} = \frac{1}{n} \cdot \frac{N-n}{N} \cdot \left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right)$$

respectively.

Model-assisted estimation

- Under the assumptions of a linear regression model $y = \alpha + \beta \cdot x + \varepsilon$
  - $E(\varepsilon_i) = 0 \quad \forall i$
  - $\sigma^2_{\varepsilon_i}$ is constant
  - $\sigma_{\varepsilon_i,\varepsilon_j} = 0 \quad \forall i \neq j$

  $\widehat{\mu}_{\text{Reg, SRS}}$ is (model-) unbiased for μ_Y;
- The variance of the regression estimator then becomes

$$V(\widehat{\mu}_{\text{Reg, SRSWOR}}) = \sigma^2_\varepsilon \cdot \left(\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{(\mu_X - \widehat{\mu}_X)^2}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}\right) \quad .$$

- Using $s^2_\varepsilon = \dfrac{1}{n-2} \cdot \sum_{i=1}^{n} \Big((y_i - \overline{y}) - b \cdot (x_i - \overline{x})\Big)^2$ as model-unbiased estimator for σ^2_ε leads to the residual variance estimator of the regression estimator.

Example 3.6 (cf. Example 3.4)

From

$$\widehat{B} = \frac{7950}{8580} = 0{,}9266$$

we get as regression estimate

$$\begin{aligned}\widehat{\mu}_{\text{Reg, SRSWOR}} &= 91{,}4 + 0{,}9266 \cdot (80{,}5 - 81{,}2) \\ &= 90{,}7514\end{aligned}$$

with estimated variance

$$\widehat{V}(\widehat{\mu}_{\text{Reg, SRSWOR}}) = \frac{1}{400} \cdot \frac{15600}{16000} \cdot \left(8230 - \frac{7950^2}{8580}\right) = 2{,}1054 \quad .$$

Before, we had for the difference estimator $\widehat{\mu}_{\text{Diff, SRS}} = 90{,}735$ and $\widehat{V}(\widehat{\mu}_{\text{Diff, SRSWOR}}) = 2{,}1168$ and for the ratio estimator $\widehat{\mu}_{\text{Ratio, SRSWOR}} = 90{,}6121$ and $\widehat{V}(\widehat{\mu}_{\text{Ratio, SRSWOR}}) = 2{,}9338$.

Remarks in linear regression estimation

- Difference and ratio estimator can be derived as special cases of regression estimation;
- One has to differ between design and model properties. In general, preferable properties of the estimators are derived under the model assumptions, which give the difference between design-based and model-assisted estimation!
- The ratio estimator, in general, uses a heteroscedastic model, which certainly does not fulfil classical LS linear regression assumptions;
- In practice, one should prefer the residual variance estimator, which is more stable especially for small sample sizes (cf. Section 3.6);

Stratified regression estimation

The estimators in Section 3.2 to 3.4 were applied to SRS (WR/WOR). Of special interest is now StrRS. One has to distinguish mainly between two (three) assumptions:

- For all strata, only one regression line has to be estimated and, hence, only one $\widehat{B}$, which is used in all strata for regression estimation. This approach is referred to as combined regression estimation.
- In case that a regression line is estimated for each stratum separately, a regression coefficient $\widehat{B}_q$ $(q = 1, \ldots, L)$ results in each stratum. A weighted estimate using all L regression coefficients results in the separate regression estimator.
- In cases of curvilinear interactions between the variable of interest and the auxiliary variable, one may prefer the linear spline regression estimation (cf. Münnich, 1997). Cf. non-parametric model-assisted estimation.

Generalized regression estimation (GREG)

The linear regression estimator can be extended to the generalized case with *many* covariates. The resulting estimator is called generalized regression estimator, which leads to

$$\widehat{\mu}_{Y,\mathrm{GREG}} = \widehat{\mu}_y + \left(\mu_{\mathbf{X}} - \widehat{\mu}_{\mathbf{x}}\right)' \cdot \widehat{\mathbf{B}} \quad ,$$

(cf. Särndal et al. 1992, pp. 225) with

$$\widehat{\mathbf{B}} = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \cdot \left(\sum_{i=1}^{n} \mathbf{x}_i y_i\right) \quad .$$

As an appropriate variance estimator for the GREG, the residual variance estimator is applied while using s_e^2 with $e_i = y_i - \mathbf{x}_i' \cdot \widehat{B}$, rather than $s_Y^2 \cdot (1 - \varrho^2)$.

Summary questions for Chapter 3

- Which estimator would you prefer by what reason?
- How do we consider a change of population while using the regression estimator?
- How do the classical sampling ideas from Chapters 1 and 2 interact with the regression estimator?
- Are there any other ideas incorporating auxiliary information? Hint: consider business statistics and, here, the impact of influential units (of high concentration).