Sampling and Estimation

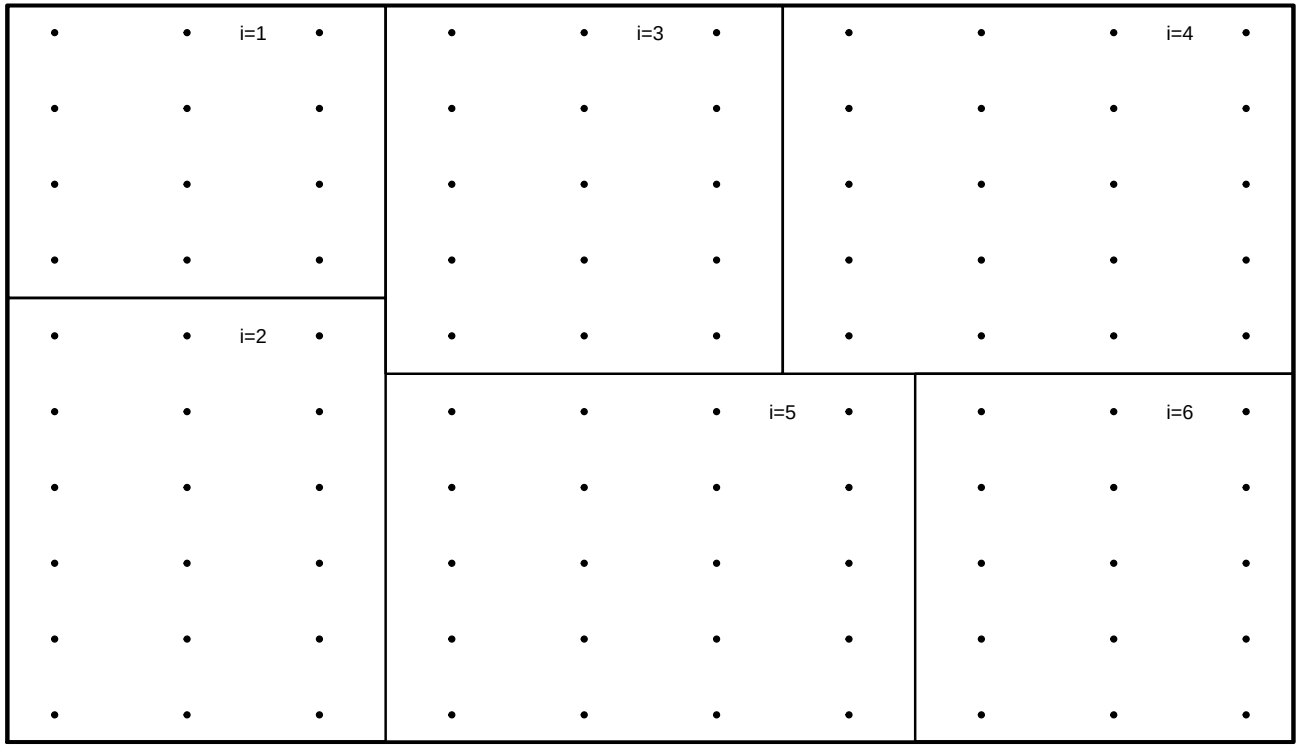
Part 2: Complex Sampling Designs

Stefan Zins[1], Matthias Sand[2], and Jan-Philipp Kolb[3]

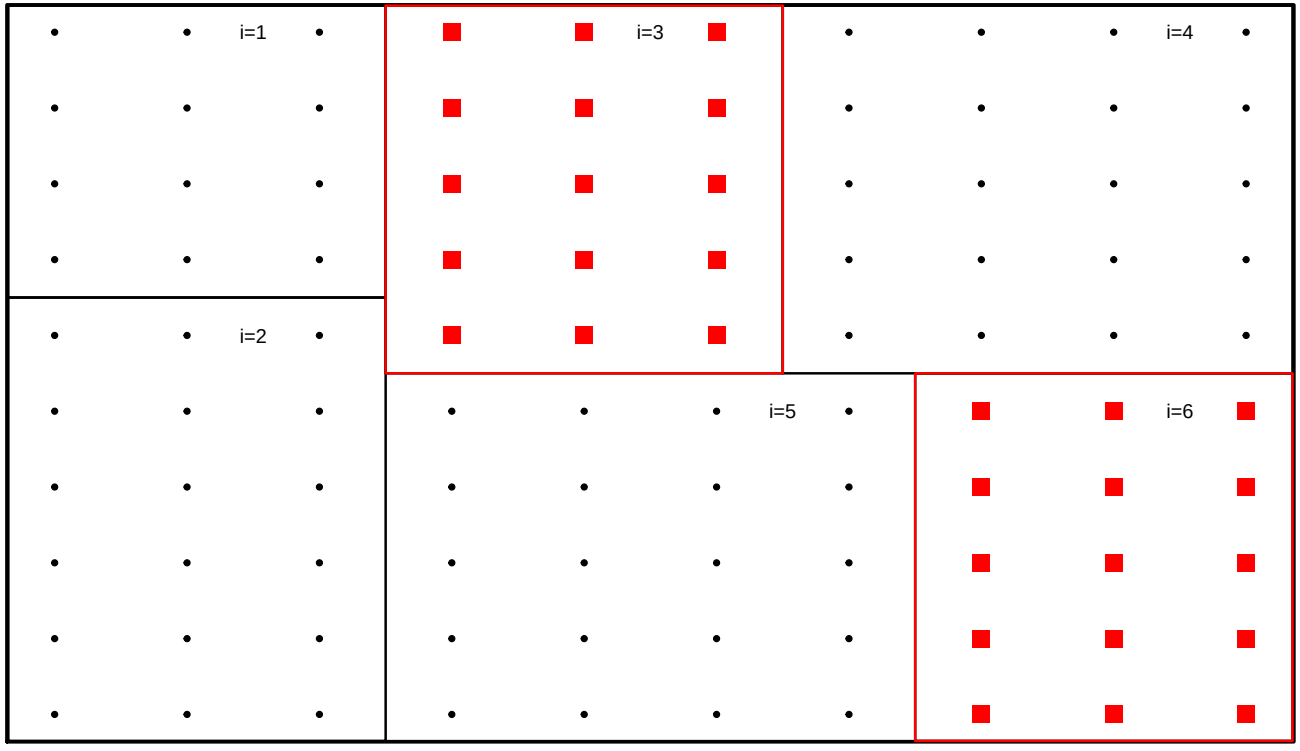
February 5, 2016

[1] Stefan.Zins@gesis.org
[2] Matthias.Sand@gesis.org
[3] Jan-Philipp.Kolb@gesis.org

A Population of 100 elements is clustered into $N_l = 6$ cluster



A Population of 100 elements is clustered into $N_l = 6$ cluster and $n_l = 2$ clusters are selected from the population.



Sampling elementary units is often not feasible (e.g. persons or businesses). Maybe there is no uniform sampling frame available to select them from, or it would be costly to do, because the selected elements would scatter to much over the a certain area and travel costs of interviewers would be to high.

Sampling elementary units is often not feasible (e.g. persons or businesses). Maybe there is no uniform sampling frame available to select them from, or it would be costly to do, because the selected elements would scatter to much over the a certain area and travel costs of interviewers would be to high.

Thus, it is very common to select clusters, so called *primary sampling units* (PSU's) that are populated by *secondary sampling units* (SSU's).

Sampling elementary units is often not feasible (e.g. persons or businesses). Maybe there is no uniform sampling frame available to select them from, or it would be costly to do, because the selected elements would scatter to much over the a certain area and travel costs of interviewers would be to high.

Thus, it is very common to select clusters, so called *primary sampling units* (PSU's) that are populated by *secondary sampling units* (SSU's).

Cluster sampling makes it still possible to obtain unbiased estimates but it can have a big influence on the variance.

Notation

y_{ki}	=	value of the variable of interest for the k-th SSU in the i-th PSU
N_{I}	=	number of PSU's in the population
N_i	=	number of SSU's in the i-th PSU
N	=	total number of SSU's in the Population
$\mathcal{U}$	=	set of SSU's in the population
$\mathcal{U}_{\mathrm{I}}$	=	set of PSU's in the population
$\mathcal{U}_i$	=	set of SSU's in the i-th PSU
n_{I}	=	number of PSU's in the sample
n_i	=	number of SSU's in the sample from the i-th PSU
$\mathcal{s}_{\mathrm{I}}$	=	sample of PSU's
$\mathcal{s}_i$	=	sample SSU's from the i-th PSU
$p_{\mathrm{I}}(.)$	=	sampling design of the PSU's

All SSU's in the sampled PSU's are surveyed, thus

$$\tau_i = \sum_{k \in \mathcal{U}_i} y_{ki}$$

is known of all selected PSU's. An unbiased estimator for the population mean is

$$\overline{y}_{\text{SRCS}} = \frac{N_I}{N} \sum_{i \in \mathcal{S}_I} \frac{\tau_i}{n_I}$$

with variance

$$\text{V}\left(\overline{y}\right)_{\text{SRCS}} = \frac{N_I^2}{N^2} \left(1 - \frac{n_I}{N_I}\right) \frac{V_\tau^2}{n_I} \ ,$$

where $V_\tau^2 = \dfrac{1}{N_I - 1} \sum_{i \in \mathcal{U}_I} \left(\tau_i - \mu_\tau\right)^2$ and $\mu_\tau = \sum_{i \in \mathcal{U}_I} \dfrac{\tau_i}{N_I}$.

All SSU's in the sampled PSU's are surveyed, thus

$$\tau_i = \sum_{k \in \mathcal{U}_i} y_{ki}$$

is known of all selected PSU's. An unbiased estimator for the population mean is

$$\overline{y}_{\text{SRCS}} = \frac{N_I}{N} \sum_{i \in s_I} \frac{\tau_i}{n_I}$$

An unbiased variance estimator is

$$\widehat{\text{V}}\left(\overline{y}_{\text{SRCS}}\right)_{\text{SRS}} = \frac{N_I^2}{N^2} \left(1 - \frac{n_I}{N_I}\right) \frac{s_\tau^2}{n_I} \ ,$$

where

$$s_\tau^2 = \frac{1}{n_I - 1} \sum_{i \in s_I} \left(\tau_i - \overline{\tau}\right)^2 \ .$$

with $\overline{\tau} = \sum_{i \in s_I} \frac{\tau_i}{n_I}$.

Table: Two Variations of a Population Composed of 5 clusters of Size 5

	1	2	3	4	5	$\mu_{i.}$	$V_{i.}^2$
1	1.0	2.0	3.0	4.0	5.0	3.0	2.5
2	6.0	7.0	8.0	9.0	10.0	8.0	2.5
3	11.0	12.0	13.0	14.0	15.0	13.0	2.5
4	16.0	17.0	18.0	19.0	20.0	18.0	2.5
5	21.0	22.0	23.0	24.0	25.0	23.0	2.5
$\mu_{.j}$	11.0	12.0	13.0	14.0	15.0	13.0	
$V_{.j}^2$	62.5	62.5	62.5	62.5	62.5		54.2

We have homogeneity of means between columns and heterogeneity of means between rows.

Table: Two Variations of a Population Composed of 5 clusters of Size 5

	1	2	3	4	5	$\mu_{i.}$	$V_{i.}^2$
1	1.0	2.0	3.0	4.0	5.0	3.0	2.5
2	6.0	7.0	8.0	9.0	10.0	8.0	2.5
3	11.0	12.0	13.0	14.0	15.0	13.0	2.5
4	16.0	17.0	18.0	19.0	20.0	18.0	2.5
5	21.0	22.0	23.0	24.0	25.0	23.0	2.5
$\mu_{.j}$	11.0	12.0	13.0	14.0	15.0	13.0	
$V_{.j}^2$	62.5	62.5	62.5	62.5	62.5		54.2

We have homogeneity of means between columns and heterogeneity of means between rows. Sample $n = 10$ SSU's by

1) SRS n=10
2) Simple random cluster sampling $n_{\text{I}} = 2$
   a) columns
   b) rows

Comparing the mean and variance of $\overline{y}$ and $\overline{y}_{SRCS}$ after 100,000 samples:

Table: Results from the Simulation Study

	SRS	SRCS a	SRCS b
mean	13.0	13.0	13.0
var	3.2	0.8	18.9

Comparing the mean and variance of $\overline{y}$ and $\overline{y}_{SRCS}$ after 100,000 samples:

Table: Results from the Simulation Study

	SRS	SRCS a	SRCS b
mean	13.0	13.0	13.0
var	3.2	0.8	18.9

True values:

$$\begin{aligned} \mu &= 13 \\ V(\overline{y})_{SRS} &= 3.25 \quad \text{SRS} \\ V(\overline{y}_{SRCS})_{SRS} &= 0.75 \quad \text{SRCS a} \\ V(\overline{y}_{SRCS})_{SRS} &= 18.75 \quad \text{SRCS b} \end{aligned}$$

Bias is not an issue, however variance is.

Comparing the mean and variance of $\overline{y}$ and $\overline{y}_{SRCS}$ after 100,000 samples:

Table: Results from the Simulation Study

	SRS	SRCS a	SRCS b
mean	13.0	13.0	13.0
var	3.2	0.8	18.9

True values:

$$\begin{aligned} \mu &= 13 \\ V(\overline{y})_{SRS} &= 3.25 \quad \text{SRS} \\ V(\overline{y}_{SRCS})_{SRS} &= 0.75 \quad \text{SRCS a} \\ V(\overline{y}_{SRCS})_{SRS} &= 18.75 \quad \text{SRCS b} \end{aligned}$$

Bias is not an issue, however variance is.

If the cluster where strata, which stratification would you use, columns or rows?

Comparing the mean and variance of $\overline{y}$ and $\overline{y}_{SRCS}$ after 100,000 samples:

TABLE: Results from the Simulation Study

	SRS	SRCS a	SRCS b
mean	13.0	13.0	13.0
var	3.2	0.8	18.9

True values:

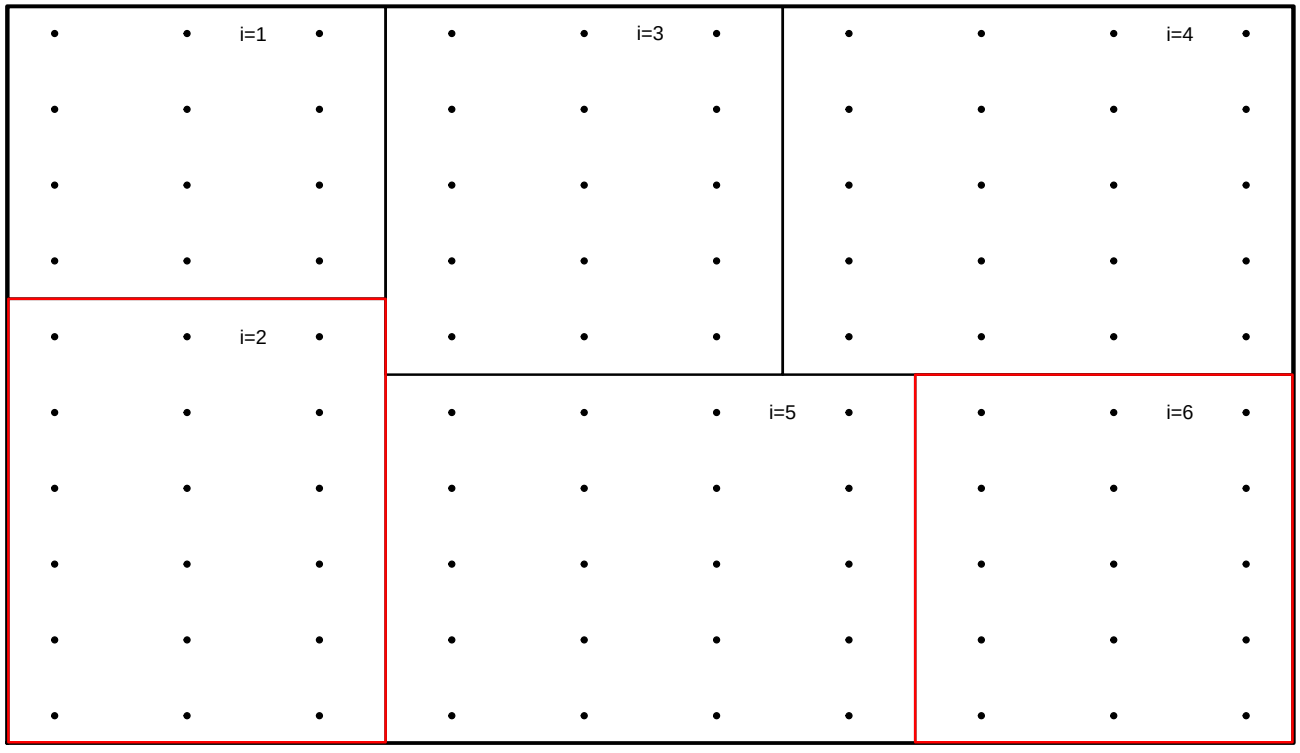
$$\begin{aligned} \mu &= 13 \\ V(\overline{y})_{SRS} &= 3.25 \quad \text{SRS} \\ V(\overline{y}_{SRCS})_{SRS} &= 0.75 \quad \text{SRCS a} \\ V(\overline{y}_{SRCS})_{SRS} &= 18.75 \quad \text{SRCS b} \end{aligned}$$

Bias is not an issue, however variance is.

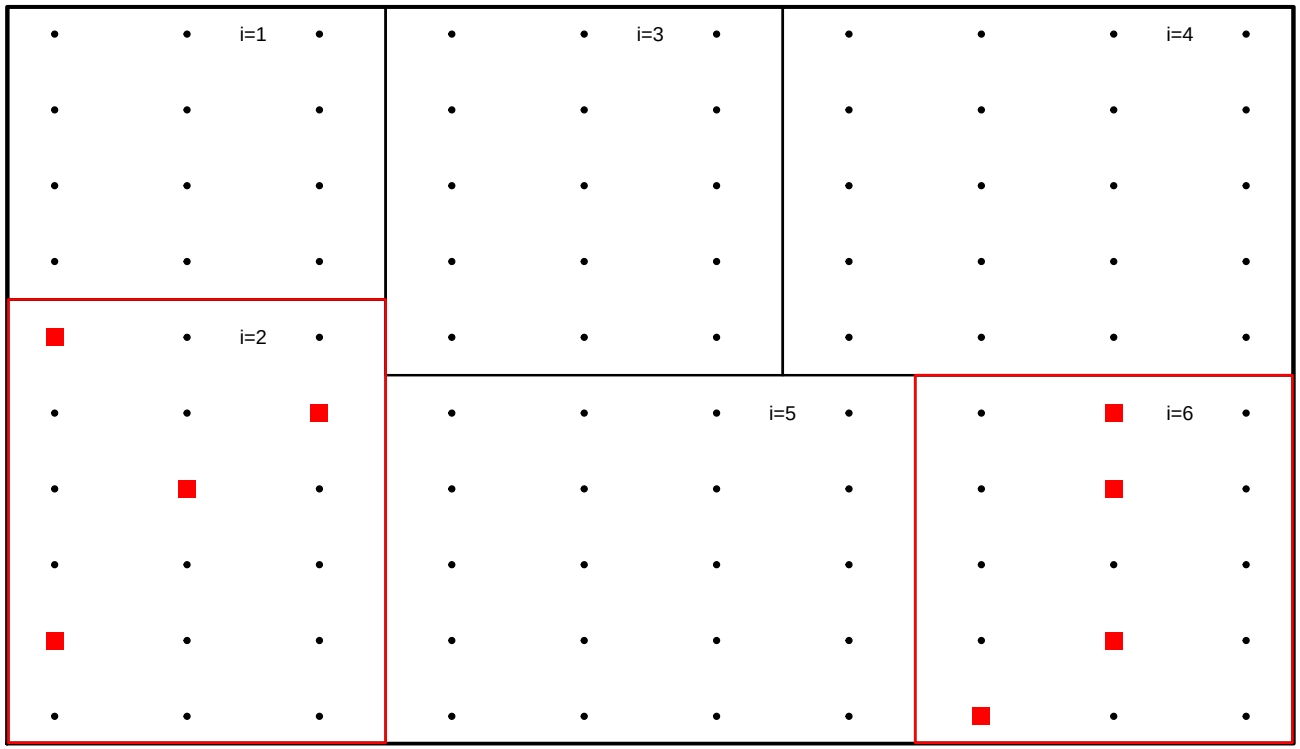
If the cluster where strata, which stratification would you use, columns or rows?

What good is for stratified sampling, i.e. low SSW, is bad of cluster sampling and vice versa.

A Population of 100 elements is clustered into $N_I = 6$ clusters and $n_I = 2$ clusters (PSU) are selected at the first sampling stage



A Population of 100 elements is clustered into $N_I = 6$ clusters and $n_I = 2$ clusters (PSU) are selected at the first sampling stage and $n_i = 4$ elements are selected from each sampled cluster.



First stage A sample $\mathcal{s}_I$ of PSU's is drawn from $\mathcal{U}_I$ according to some sampling design $p_I(.)$

Second stage For every $i \in \mathcal{s}_I$ a sample $\mathcal{s}_i$ of SSU's is selected from $\mathcal{U}_i$ according to some design $p_i(.|\mathcal{s}_I)$

The resulting sample of SSU's is denote $\mathcal{s} = \bigcup_{i \in \mathcal{s}_I} \mathcal{s}_i$. In general, samples $\mathcal{s}_i$ are selected independently of each other, thus, the inclusion probability of a element $k \in \mathcal{U}_i$ is

$$\pi_k = \pi_{I i} \pi_{k|i} \ ,$$

where $\pi_{I i}$ is the probability of selecting the i-th PSU and $\pi_{k|i}$ the probability of selecting the k-th SSU in the i-th PSU.

Estimation Simple Random Two Stage Sampling

Designs $p_I(.)$ and $p_i(.|\mathcal{s}_I)$ are both SRS. Since not all SSU's in the sampled PSU's are surveyed τ_i has to be estimated by $\hat{\tau}_i = \frac{N_i}{n_i} \sum_{k \in \mathcal{s}_i} y_{ki}$. An unbiased estimator for the population mean is

$$\overline{y}_{2SRS} = \frac{N_I}{N} \sum_{i \in \mathcal{s}_I} \frac{\hat{\tau}_i}{n_I}$$

Estimation Simple Random Two Stage Sampling

Designs $p_{\mathrm{I}}(.)$ and $p_i(.|s_{\mathrm{I}})$ are both SRS. Since not all SSU's in the sampled PSU's are surveyed τ_i has to be estimated by $\hat{\tau}_i = \frac{N_i}{n_i} \sum_{k \in s_i} y_{ki}$. An unbiased estimator for the population mean is

$$\overline{y}_{2\mathrm{SRS}} = \frac{N_{\mathrm{I}}}{N} \sum_{i \in s_{\mathrm{I}}} \frac{\hat{\tau}_i}{n_{\mathrm{I}}}$$

with variance

$$\mathrm{V}\left(\overline{y}_{2\mathrm{SRS}}\right)_{\mathrm{SRS}} = \frac{1}{N^2} \left(N_{\mathrm{I}}^2 \left(1 - \frac{n_{\mathrm{I}}}{N_{\mathrm{I}}}\right) \frac{V_\tau^2}{n_{\mathrm{I}}} + \frac{N_{\mathrm{I}}}{n_{\mathrm{I}}} \sum_{i \in \mathcal{U}_{\mathrm{I}}} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{V_i^2}{n_i} \right) ,$$

where $V_i^2 = \frac{1}{N_i - 1} \sum_{k \in \mathcal{U}_i} (y_{ki} - \mu_i)^2$ with $\mu_i = \sum_{k \in \mathcal{U}_i} \frac{y_{ki}}{N_i}$.

Estimation Simple Random Two Stage Sampling

Designs $p_{\mathrm{I}}(.)$ and $p_i(.|\mathcal{S}_{\mathrm{I}})$ are both SRS. Since not all SSU's in the sampled PSU's are surveyed τ_i has to be estimated by $\hat{\tau}_i = \frac{N_i}{n_i} \sum_{k \in \mathcal{S}_i} y_{ki}$. An unbiased estimator for the population mean is

$$\overline{y}_{2\mathrm{SRS}} = \frac{N_{\mathrm{I}}}{N} \sum_{i \in \mathcal{S}_{\mathrm{I}}} \frac{\hat{\tau}_i}{n_{\mathrm{I}}}$$

An unbiased variance estimator is given by

$$\hat{\mathrm{V}}\left(\overline{y}_{2\mathrm{SRS}}\right)_{\mathrm{SRS}} = \frac{1}{N^2}\left(N_{\mathrm{I}}^2\left(1-\frac{n_{\mathrm{I}}}{N_{\mathrm{I}}}\right)\frac{s_{\hat{\tau}}^2}{n_{\mathrm{I}}} + \frac{N_{\mathrm{I}}}{n_{\mathrm{I}}}\sum_{i \in \mathcal{S}_i} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{s_i^2}{n_I}\right) ,$$

where $s_{\hat{\tau}}^2 = \frac{1}{n_{\mathrm{I}}-1}\sum_{i \in \mathcal{S}_{\mathrm{I}}}(\hat{\tau}_i - \overline{\hat{\tau}})^2$ with $\overline{\hat{\tau}} = \sum_{i \in \mathcal{S}_{\mathrm{I}}} \frac{\hat{\tau}_i}{n_{\mathrm{I}}}$ and $s_i^2 = \frac{1}{n_i-1}\sum_{k \in \mathcal{S}_i}(y_{ki} - \overline{y}_i)^2$ with $\overline{y}_i = \sum_{k \in \mathcal{S}_i} \frac{y_{ki}}{n_i}$.

There are good reasons to deviate from the simple selection procedure that gives every unit the same inclusion probability. If good prior information is available, its incorporation into the sampling design can dramatically improve the efficiency of an estimator.

- An optimal allocation would be favorable to a proportional allocation.
- Selecting the elements proportional to a variable that is correlated to the variable of interest can greatly improve the quality of estimates.

There are many techniques (i.e. sampling algorithms) to select elements with unequal probabilities [Tillé, 2006].

A design unbiased estimator for the total $\tau = \sum_{k \in \mathcal{U}} y_k$ is given by

$$\hat{\tau}_\pi = \sum_{k \in \mathcal{s}} \frac{y_k}{\pi_k} \ ,$$

which is also known as *Horvitz-Thompson* (HT) or π-estimator.

A design unbiased estimator for the total $\tau = \sum_{k \in \mathcal{U}} y_k$ is given by

$$\hat{\tau}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} \ ,$$

which is also known as *Horvitz-Thompson* (HT) or π-estimator. The variance of $\hat{\tau}_\pi$ is

$$\mathsf{V}\left(\hat{\tau}_\pi\right) = \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \left(\pi_{kl} - \pi_k \pi_l\right) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \ ,$$

which can be estimated by

$$\widehat{\mathsf{V}}\left(\hat{\tau}_\pi\right)_1 = \sum_{k \in s} \sum_{l \in s} \frac{\left(\pi_{kl} - \pi_k \pi_l\right)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \ .$$

A design unbiased estimator for the total $\tau = \sum_{k \in \mathcal{U}} y_k$ is given by

$$\hat{\tau}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} \ ,$$

which is also known as *Horvitz-Thompson* (HT) or π-estimator. For a fixed size design, we may write the variance of $\hat{\tau}_\pi$ as

$$\mathrm{V}\left(\hat{\tau}_\pi\right) = -\frac{1}{2} \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \left(\pi_{kl} - \pi_k \pi_l\right) \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \ ,$$

which can be estimated by

$$\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_2 = -\frac{1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\left(\pi_{kl} - \pi_k \pi_l\right)}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \ .$$

A design unbiased estimator for the total $\tau = \sum_{k \in \mathcal{U}} y_k$ is given by

$$\hat{\tau}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} \, ,$$

which is also known as *Horvitz-Thompson* (HT) or π-estimator. For a fixed size design, we may write the variance of $\hat{\tau}_\pi$ as

$$\mathrm{V}\left(\hat{\tau}_\pi\right) = -\frac{1}{2} \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \left(\pi_{kl} - \pi_k \pi_l\right) \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \, ,$$

which can be estimated by

$$\hat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_2 = -\frac{1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\left(\pi_{kl} - \pi_k \pi_l\right)}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \, .$$

Provided that $\pi_{kl} > 0$ for all $k \neq l \in \mathcal{U}$ both variance estimators are unbiased. Nevertheless both variance estimators can become negative!

If there is some prior information available in the form of a variable $\mathcal{X} = \{x_1, x_2, \ldots, x_k, \ldots, x_N\}$, which is correlated to our variable of interest $\mathcal{Y}$, we can select elements proportional to it

$$\nu_k = \frac{x_k}{\sum_{k \in \mathcal{U}} x_k} n \, .$$

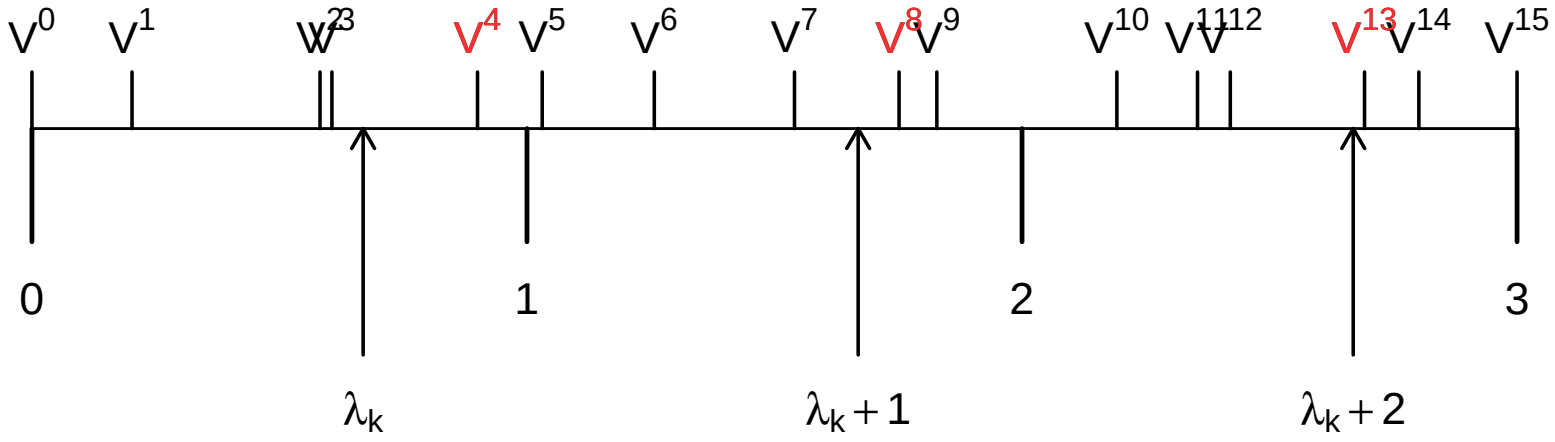
Unequal inclusion probabilities can reduce the variance of an estimator, *if* they are related to the variable of interest. We may have $\nu_k = \pi_k$ but in general ν_k can be greater than 1.

For instance, to estimate the sales in an industry, sampling companies or business with equal probabilities might be bad idea. It would be better to sample companies proportional to a variable that is related to their sales, say their number of employee. That way there is a much higher chance to included the biggest companies into the sample that accumulate the major share of the sales.

Note: In the extreme case if $\pi_k = \alpha y_k$, with $\alpha \in \mathbb{R}$, for all $k \in \mathcal{U}$ and we have a fixed size design $\mathrm{V}(\hat{\tau}_\pi)$ would even be zero.

Systematic Sampling with Unequal Inclusion Probabilities

Again the elements of the population are brought into a specific ordered and $V^i = \sum_{k=1}^{i} \pi_k$. Then λ_k is drawn from a uniform distribution between 0 and 1.



Systematic selection remains popular because of its simplicity. Also it can easily be applied to the case where an element can be selected more than one time, i.e. $\pi_k \neq \nu_k > 1$. Then we would use $V^i = \sum_{k=1}^{i} \nu_k$.

A two-stage Sampling Design:

First Stage Municipalities are the PSU's. The sampling design for the PSU's is a stratified design with an allocation proportional to the population within each stratum (not number of PSU's). Within the strata PSU's are sampled proportional to their population size.

Second Stage Persons are the SSU's. The SSU's are selected form the population register of the municipalities by a simple systematic sample.

Very large municipalities, (e.g. Berlin), are selected with certainty, this happens if $\nu_i = \frac{N_i}{N} n_| > 1$. The integer part of ν_i indicates how many sampling points are *at least* associated with a municipalities. A sampling point is here a multiplier, indicating how many times n_i SSU's are selected from the i-th PSU, where n_i is usually fix for all PSU's.

For instance, $\nu_i = 3.4$, means that the i-th PSU will always be in the sample with at least 3 sampling points, but with a probability of 0.4 it can be in the sample with 4 sampling points.

Has this design equal inclusion probabilities?

Has this design equal inclusion probabilities?
Yes, if for each sampling point the same number of SSU's n_* is sampled. Because

$$\frac{N_i}{N} n_{\mathrm{I}} \times \frac{n_*}{N_i} = \frac{n_{\mathrm{I}} n_*}{N} \ .$$

Note, that n_{I} is not the size of the PSU sample, but the number of sampling points, which can be higher.

Selecting PSU's or clusters proportional to some size measure is very common. This however does not mean that the inclusion probabilities of elementary units are unequal.

The concept of two-stage designs can also be extended to three, four, or more stages. The principle of such multi-stage design remains the same, select clusters then select again within clusters.

There are many ways to optimize the sampling design with respect to one particular goal, i.e. the estimation of a specific statistic. However, it becomes difficult to optimize a design and at the same time retain a balance for a maximum of possible applications, which is a problem when planning a multipurpose survey that has a multitude of variables and covers different topics. Thus, simple designs, such as SRS or StrSRS, are justifiable, as these designs are robust towards any possible analysis of the sample data.

Multi-stage sampling is usually not a matter of choose, but done out of necessity.

Most importantly, the same design weights (π_k^{-1}) do *not* imply the same sampling variance. Different designs can be used to select samples with same π_k's, however their π_{kl}'s might be very different and so is their associated sampling variance.

The design effect compares strategies, i.e. a combination of a sampling design and an estimator.
If $p(.)$ is some other design than SRS, however with $\sum_{i=1}^{N} \pi_k$ equal to the sample size n of the SRS design, then the *design effect* for strategy $(p(.), \hat{\tau}_\pi)$ can be defined as

$$deff(p, \hat{\tau}_\pi) = \frac{V(\hat{\tau}_\pi)_p}{V(\hat{\tau}_\pi)_{SRS}} = \frac{\sum_{k=1}^{N} \sum_{l=1}^{N} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}}{N^2 \left(1 - \frac{1}{N}\right) \frac{V^2}{n}} .$$

The design effect $deff(p, \hat{\tau}_\pi)$ expresses how well a design $p(.)$ fares in comparison to reference design SRS.

- $deff(p, \hat{\tau}_\pi) > 1$ precision is lost by not using SRS
- $deff(p, \hat{\tau}_\pi) < 1$ precision is gained by not using SRS

Recall from stratified sampling:

Table: Population ANOVA

Source	sf	Sum of Squares
Between cluster	$N_{\text{l}} - 1$	$\text{SSB} = \sum_{i=1}^{N_{\text{l}}} N_i(\mu_i - \mu)^2$
Within cluster	$N - N_{\text{l}}$	$\text{SSW} = \sum_{i=1}^{N_i} (N_i - 1) V_i^2$
Total	$N - 1$	$\text{SSTO} = (N - 1) V^2$

Table: Two Variations of a Population Composed of 5 clusters of Size 5

	1	2	3	4	5	$\mu_{i.}$	$V_{i.}^2$
1	1.0	2.0	3.0	4.0	5.0	3.0	2.5
2	6.0	7.0	8.0	9.0	10.0	8.0	2.5
3	11.0	12.0	13.0	14.0	15.0	13.0	2.5
4	16.0	17.0	18.0	19.0	20.0	18.0	2.5
5	21.0	22.0	23.0	24.0	25.0	23.0	2.5
$\mu_{.j}$	11.0	12.0	13.0	14.0	15.0	13.0	
$V_{.j}^2$	62.5	62.5	62.5	62.5	62.5		54.2

Table: Design Effects

	SRCS a	SRCS b
deff	0.23077	5.76923

$$deff(p(.), \hat{\theta}) = \frac{V\left(\hat{\theta}\right)_p}{V\left(\hat{\theta}\right)_{SRS}}$$

Find suitable estimators for both

- the denominator and
- the numerator

$$deff(p(.), \hat{\theta}) = \frac{V\left(\hat{\theta}\right)_p}{V\left(\hat{\theta}\right)_{SRS}}$$

Find suitable estimators for both

- the denominator and
- the numerator

Treat the data as if it had arisen from SRS for estimation of the enumerator (but using available weights).

$$deff(p(.), \hat{\theta}) = \frac{V\left(\hat{\theta}\right)_p}{V\left(\hat{\theta}\right)_{SRS}}$$

Find suitable estimators for both

- the denominator and
- the numerator

Treat the data as if it had arisen from SRS for estimation of the enumerator (but using available weights).

It is also common to use SRSWR as the reference design, which can simplify the design effect estimation.

Since it is so difficult to estimate the design effect, models are used to describe it. Effectively an alternative version of the design effect is defined which is easier to estimate.

Again we have N_I clusters in the population of size N_i, $i = 1, \ldots, N_I$

Variable $\mathcal{Y}$ is assumed to follow the *common parameter model*,

$$\begin{aligned} \mathrm{E}\left(y_{ki}\right)_M &= \mu \\ \mathrm{V}\left(y_{ki}\right)_M &= \sigma^2 \\ \mathrm{COV}\left(y_{ki},\, y_{k'i'}\right)_M &= \begin{cases} \sigma^2\rho & \text{for } k \neq k', i = i' \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

ρ the so called *Intra-Class Correlation Coefficient* is a central parameter to the model. It determine how similar to each other the elements from the same cluster are.

The considered estimator is $\overline{y}_w = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$, where w_k is the survey associated with k-th element, e.g.

$$w_k = \begin{cases} \pi_k^{-1} & \text{for } k \in s \\ 0 & \text{else} \end{cases}.$$

Under the model-based approach, *deff* can be displayed as the product of two factors:

***deff*$_c$** design effect due to clustering
***deff*$_p$** design effect due to unequal survey weights

They indicate loss in precision due to cluster sampling and unequal weights, respectively.

Note that under the model-based approach, w_k is treated as independent of y_k for all $k \in \mathcal{U}$. Potential gains in precision by using proportional to size design are not considered, quite the contrary, unequal weight will increase the variance of $\overline{y}_w$ in this setting.

Under cluster sampling and two-stage sampling we can use:

$$\widehat{deff}_c = 1 + \left(\bar{b} - 1\right) \hat{\rho}$$

with

$\bar{b}$ as the average cluster size $\frac{N}{N_l}$ ($\frac{n}{n_l}$) or an estimator for it and $\hat{\rho}$ as an appropriate estimator of ρ.

When w_k's vary we have

$$\widehat{deff}_c = 1 + \left(b^* - 1\right) \hat{\rho}$$

where $b^* = \frac{\sum_{i \in s_l} \left(\sum_{k \in s_i} w_k\right)^2}{\sum_{k \in s} w_k^2}$ is a kind of weighted average cluster size [**?**].

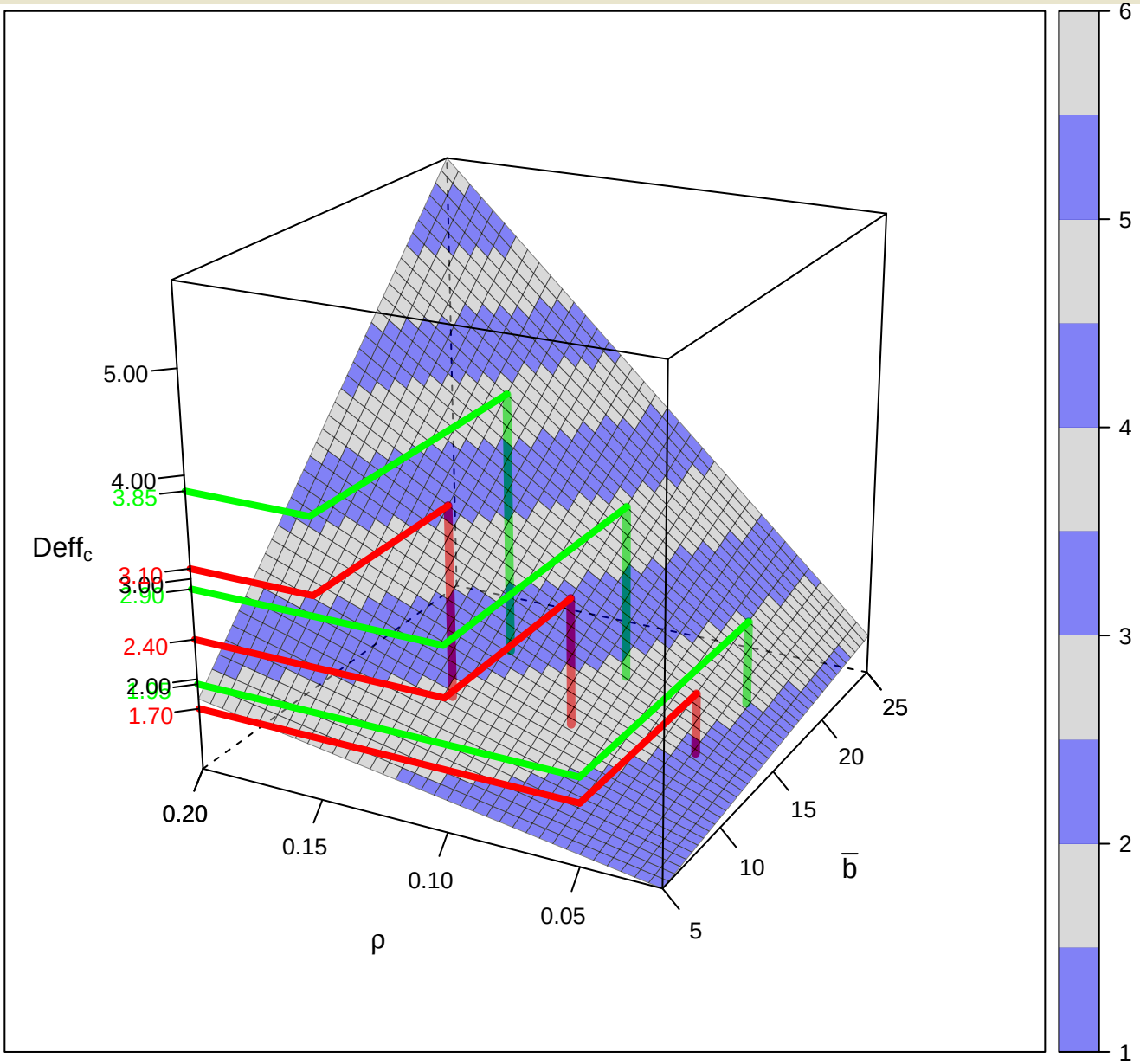
Calculation of $deff_p$

The design effect due to unequal weights is

$$deff_p = n \frac{\sum_{k \in s} w_k^2}{\left(\sum_{k \in s} w_k\right)^2}$$

Obviously, if w_k's are constant, $deff_p = 1$

Behavior of $deff_c$



An Estimators of ρ

There are many different ways to estimate ρ [**?**]. The classical ANOVA estimator is:

$$\hat{\rho} = \frac{\widehat{MSB} - \widehat{MSW}}{\widehat{MSB} + (K-1)\widehat{MSW}} ,$$

where $\widehat{MSB} = (n_l - 1)^{-1} \sum_{i \in s_l} n_i (\overline{y}_i - \overline{y})^2$,
$\widehat{MSW} = (n - n_l)^{-1} \sum_{i \in s_l} (n_i - 1) s_i^2$, and

$$K = (n_l - 1)^{-1} \left(n - \sum_{i \in s_l} \frac{n_i^2}{n} \right) .$$

There are many different ways to estimate ρ [**?**]. The classical ANOVA estimator is:

$$\hat{\rho} = \frac{\widehat{MSB} - \widehat{MSW}}{\widehat{MSB} + (K-1)\widehat{MSW}} ,$$

where $\widehat{MSB} = (n_| - 1)^{-1} \sum_{i \in s_|} n_i (\overline{y}_i - \overline{y})^2$,
$\widehat{MSW} = (n - n_|)^{-1} \sum_{i \in s_|} (n_i - 1) s_i^2$, and

$$K = (n_| - 1)^{-1} \left(n - \sum_{i \in s_|} \frac{n_i^2}{n} \right) .$$

The model-based approach is widely used, because it often presents the only option for data users to estimate a design effect.

For multi-stage design, the used model considers only the cluster effect of the PSU's and neglects any subsequent sampling stages.

Comparison of Design Based and Model Based Design Effects

gesis Leibniz Institute for the Social Sciences

Table: Intra-Class Correlation Coefficients and (Model) Design Effects

	SRCS a	SRCS b
ρ	-0.01626	0.99136
deff	0.93496	4.96546

Table: Design Effects

	SRCS a	SRCS b
deff	0.23077	5.76923

Literature I


S. Gabler, S. Häder, & P. Lahiri.
A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering.
Survey Methodology, 1999.


M. Ganninger.
Design Effects: Model-based versus Design-based Approach
PhD Thesis, *GESIS-Schriftenreihe Band 3*, 2009


C.-E. Särndal, B. Swensson, & J. Wretman.
Model Assisted Survey Sampling
Springer, 1992.


Y. Tillé.
Sampling Algorithms
Springer Series in Statistics: Springer, 2006.