Sampling and Estimation

Part 2: Complex Sampling Designs

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

January 27, 2016

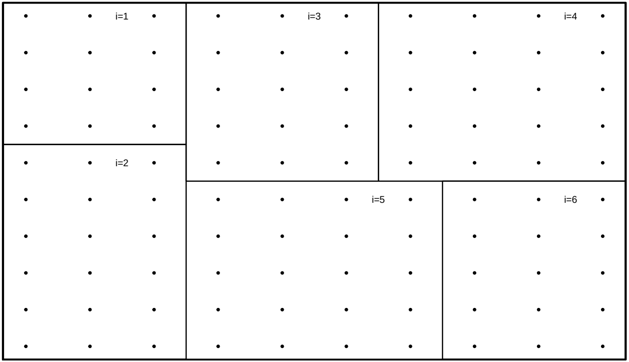
[1] Stefan.Zins@gesis.org

[2] Matthias.Sand@gesis.org

[3] Jan-Philipp.Kolb@gesis.org

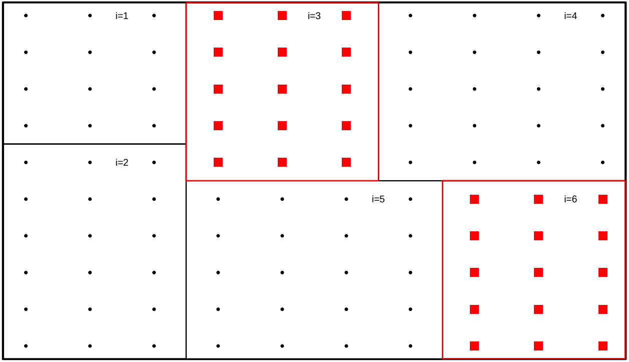
CLUSTERING

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



Clustering

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



Cluster Sampling

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.

Cluster Sampling

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

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

$\mathscr{s}_{\mathrm{I}}$ = sample of PSU's

$\mathscr{s}_i$ = sample SSU's from the *i*-th PSU

$p_{\mathrm{I}}(.)$ = sampling design of the PSU's

Estimation in Case of $p_{\mathrm{I}}(.) = \text{SRS}$

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

$$\overline{y}_{\text{SRCS}} = \frac{N_{\text{I}}}{N} \sum_{i \in \mathcal{S}_{\text{I}}} \frac{\tau_i}{n_{\text{I}}}$$

with variance

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

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

Estimation in Case of $p_{\mathrm{I}}(.) = \mathrm{SRS}$

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

$$\overline{y}_{\mathrm{SRCS}} = \frac{N_{\mathrm{I}}}{N} \sum_{i \in \mathcal{S}_{\mathrm{I}}} \frac{\tau_i}{n_{\mathrm{I}}}$$

An unbiased variance estimator is

$$\widehat{\mathrm{V}}\left(\overline{y}_{\mathrm{SRCS}}\right)_{\mathrm{SRS}} = \frac{N_{\mathrm{I}}^2}{N^2} \left(1 - \frac{n_{\mathrm{I}}}{N_{\mathrm{I}}}\right) \frac{s_\tau^2}{n_{\mathrm{I}}} \ ,$$

where

$$s_\tau^2 = \frac{1}{n_{\mathrm{I}} - 1} \sum_{i \in \mathcal{S}_{\mathrm{I}}} \left(\tau_i - \overline{\tau}\right)^2 \ .$$

with $\overline{\tau} = \sum_{i \in \mathcal{S}_{\mathrm{I}}} \frac{\tau_i}{n_{\mathrm{I}}}$.

Simple Cluster Sampling

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

	1	2	3	4	5	$\mu_{i.}$	$V^2_{i.}$
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	
$V^2_{.j}$	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.

Simple Cluster Sampling

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

	1	2	3	4	5	$\mu_{i.}$	$V^2_{i.}$
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	
$V^2_{.j}$	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

Simple Cluster Sampling

Comparing the mean and variance of $\overline{y}$ and $\overline{y}_{\text{SRCS}}$ 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

Simple Cluster Sampling

Comparing the mean and variance of $\bar{y}$ and $\bar{y}_{\text{SRCS}}$ 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:

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

Bias is not an issue, however variance is.

Simple Cluster Sampling

Comparing the mean and variance of $\overline{y}$ and $\overline{y}_{\text{SRCS}}$ 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:
$$\begin{array}{lcrl} \mu & = & 13 & \\ V\left(\overline{y}\right)_{\text{SRS}} & = & 3.25 & \text{SRS} \\ V\left(\overline{y}_{\text{SRCS}}\right)_{\text{SRS}} & = & 0.75 & \text{SRCS a} \\ V\left(\overline{y}_{\text{SRCS}}\right)_{\text{SRS}} & = & 18.75 & \text{SRCS b} \end{array}$$

Bias is not an issue, however variance is.

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

Simple Cluster Sampling

Comparing the mean and variance of $\bar{y}$ and $\bar{y}_{\text{SRCS}}$ 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:

$$\begin{array}{lcrl} \mu & = & 13 & \\ \text{V}\left(\bar{y}\right)_{\text{SRS}} & = & 3.25 & \text{SRS} \\ \text{V}\left(\bar{y}_{\text{SRCS}}\right)_{\text{SRS}} & = & 0.75 & \text{SRCS a} \\ \text{V}\left(\bar{y}_{\text{SRCS}}\right)_{\text{SRS}} & = & 18.75 & \text{SRCS b} \end{array}$$

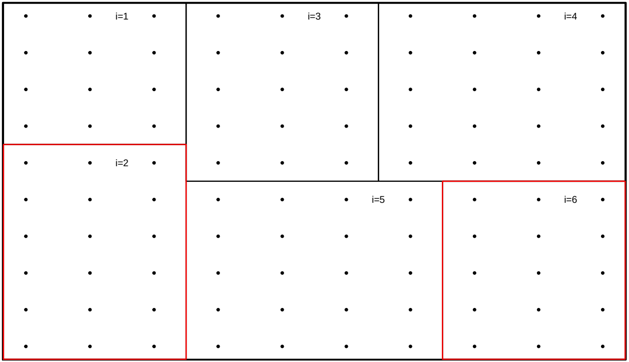
Bias is not an issue, however variance is.

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.

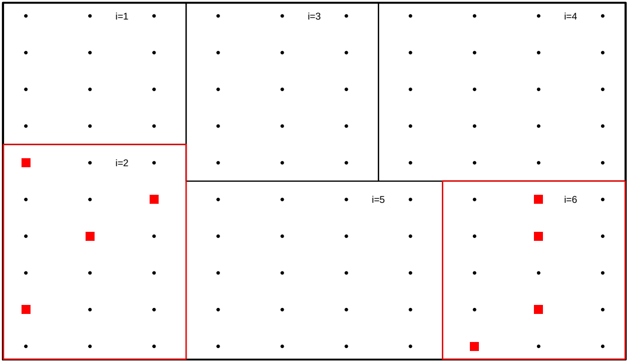
Two Stage Sampling

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



Two Stage Sampling

A Population of 100 elements is clustered into $N_{\mathrm{I}} = 6$ clusters and $n_{\mathrm{I}} = 2$ clusters (PSU) are selected at the first sampling stage and $n_i = 4$ elements are selected from each sampled cluster.



Two Stage Sampling

First stage A sample $\mathscr{s}_{\mathrm{I}}$ of PSU's is drawn from $\mathcal{U}_{\mathrm{I}}$ according to some sampling design $p_{\mathrm{I}}(.)$

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

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

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

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

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

Estimation Simple Random Two Stage Sampling

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

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

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 $\mu_i = \sum_{k\in\mathcal{U}_i}\dfrac{y_{ki}}{N_i}$.

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

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

An unbiased variance estimator is given by

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

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

Unequal Probability Sampling 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 Generic Design Based Estimator

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

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

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

A Generic Design Based Estimator

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

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

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

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

which can be estimated by

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

A Generic Design Based Estimator

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

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

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

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

A Generic Design Based 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. For a fixed size design we may write the variance of $\hat{\tau}_\pi$ as

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

which can be estimated by

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

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!

Unequal Probability Sampling II

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 interest $\mathcal{Y}$ we can select elements proportional to it

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

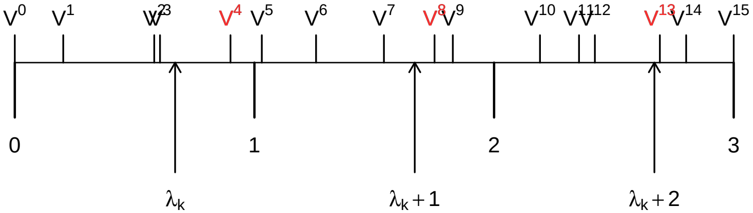
Unequal inclusion probabilities can reduce the variance of an estimator, *if* they are related to the variable of interest. We may have $\nu_k = \pi_k$ but in general ν_k can be greater than 1.

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 $\mathrm{V}\left(\hat{\tau}_\pi\right)$ 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 Typical Sample of Persons in Germany

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.

A Typical Sample of Persons

in Germany

Very large municipalities, (e.g. Berlin), are selected with certainty, this happens if $\nu_i = \dfrac{N_i}{N} n_{\text{I}} > 1$. 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.

A Typical Sample of Persons

in Germany

Has this design equal inclusion probabilities?

A Typical Sample of Persons

in Germany

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.

On Complex Sampling Design

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.

On Complex Sampling Design

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.

On Complex Sampling Design

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 design, such as SRS or StrSRS, are justifiable, as these designs are robust towards any possible analysis of the sample data.

On Complex Sampling Design

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

On Complex Sampling Design

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

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

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

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

Design Effect for a Cluster Sample

Recall from stratified sampling:

Table: Population ANOVA

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

Design Effect for a Cluster Sample

The homogeneity coefficient

$$\delta = 1 - \frac{SSW(N - N_{\mathrm{I}})^{-1}}{SSTO(N - 1)^{-1}}$$

is a measure for the similarity of elements within the same cluster.

Design Effect for a Cluster Sample

The homogeneity coefficient

$$\delta = 1 - \frac{SSW(N - N_{\mathrm{I}})^{-1}}{SSTO(N-1)^{-1}}$$

is a measure for the similarity of elements within the same cluster. It can be shown that

$$\mathrm{V}\left(\overline{y}_{\mathrm{SRCS}}\right)_{\mathrm{SRS}} = \left(1 + \frac{N - N_{\mathrm{I}}}{N_{\mathrm{I}} - 1}\delta\right)\mathrm{V}\left(\overline{y}\right)_{\mathrm{SRS}} + N_{\mathrm{I}}^2\left(1 - \frac{n_{\mathrm{I}}}{N_{\mathrm{I}}}\right)\frac{\mathrm{COV}}{n_{\mathrm{I}}}\ ,$$

where $\mathrm{COV} = \frac{1}{N_{\mathrm{I}}-1}\sum_{i\in\mathcal{U}_{\mathrm{I}}}(N_i - \frac{N}{N_{\mathrm{I}}})N_i\mu_i^2$ [Sändal, 1992, p. 131f.] and the design of SRCS is given by

$$deff(SRCS, \hat{\tau}_\pi/N) = 1 + \frac{N - N_{\mathrm{I}}}{N_{\mathrm{I}} - 1}\delta + \frac{N\,\mathrm{COV}}{N_{\mathrm{I}}V^2}\ .$$

Design Effect for a Cluster Sample

It can be shown that

$$\mathrm{V}\left(\bar{y}_{\mathrm{SRCS}}\right)_{\mathrm{SRS}} = \left(1 + \frac{N - N_{\mathrm{I}}}{N_{\mathrm{I}} - 1}\delta\right)\mathrm{V}\left(\bar{y}\right)_{\mathrm{SRS}} + N_{\mathrm{I}}^2\left(1 - \frac{n_{\mathrm{I}}}{N_{\mathrm{I}}}\right)\frac{\mathrm{COV}}{n_{\mathrm{I}}},$$

where $\mathrm{COV} = \frac{1}{N_{\mathrm{I}}-1}\sum_{i\in\mathcal{U}_{\mathrm{I}}}(N_i - \frac{N}{N_{\mathrm{I}}})N_i\mu_i^2$ [Sändal, 1992, p. 131f.] and the design of SRCS is given by

$$\mathit{deff}(SRCS, \hat{\tau}_\pi/N) = 1 + \frac{N - N_{\mathrm{I}}}{N_{\mathrm{I}} - 1}\delta + \frac{N\,\mathrm{COV}}{N_{\mathrm{I}}V^2}\ .$$

In case N_i is constant for all cluster COV = 0 and we have

$$\mathit{deff}(SRCS, \hat{\tau}_\pi/N) = 1 + \frac{N - N_{\mathrm{I}}}{N_{\mathrm{I}} - 1}\delta \approx 1 + \left(\frac{N}{N_{\mathrm{I}}} - 1\right)\delta$$

Note that δ is the adjusted measure of fit for fitting the linear regression of $\mathcal{Y}$ on $N_{\mathrm{I}} - 1$ dummy variables, indicating cluster membership.

Design Effect for a Cluster Sample

Table: Intra-Cluster Homogeneity and Design Effects

	SRCS a	SRCS b
δ	-0.15385	0.95385
deff	0.23077	5.76923

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

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.