SERISS  Synergies for Europe’s Research Infrastructures in the Social Sciences

Contract Number: 654221

Deliverable 5.8
Report on the online statistical training

Start date of project:  1-7-2015                    Duration: 48 months

Organisation name of lead contractor for this deliverable:   ESS (GESIS)

Due date of deliverable:  October 2018

Actual submission date:  October 2019
SERISS

WP5 Training and dissemination

Task 5.4 Statistical training for secondary users

1 Aims of the Task 5.4

The aim of this task is to equip secondary users of already collected survey data researchers with the skills and knowledge they need to analyse those data and improve the rigour of cross-national survey research in the European context and acquire a better understanding of the underlying methodology of comparative survey research. It consists of two parts:

1) Face-to-face training

A series of training courses with groups of about 25 participants per course were held over the period 2015-2019. The seminars were designed to enable intense, workshop-type courses, with participants interacting with trainers, receiving individualised feedback and doing practical exercises in small ad-hoc teams. Applicants were selected according to the published criteria by a panel consisting of senior scientists by ESS ERIC (UL and GESIS).

2) Online training

Using the same themes as in the workshops, short webinars enabling a wider train the trainers’ action will be made between April and June 2019. Webinars will allow for wider audiences and are especially of benefit for those researchers with low travel budget. The webinars will be offered through the ESS website and will be conducted and archived by ESS ERIC (GESIS) by the end of the SERISS project. The aim was to set up an online webinar training based on the four face-to-face training courses organised throughout the lifetime of the project led only to a reduced availability of training material for online access. While recording facilities and staff was lacking until late in the project, trainers were often not available on short notice towards the end of the project. In addition, courses with strong elements of hands-on training would have implied huge efforts to be converted into useful online seminars, which are difficult to deliver for casual trainers.

The alternative was therefore to make training materials available where possible.


Sampling and Estimation

Part 1: Introduction to Design Based Inference

Stefan Zins\textsuperscript{1} and Matthias Sand\textsuperscript{2}

April 24, 2017

\textsuperscript{1}Stefan.Zins@gesis.org
\textsuperscript{2}Matthias.Sand@gesis.org
Welcome and Introduction
Motivation

- What do you want to do?
**Motivation**

![Diagram of p-values and interpretation](http://xkcd.com/1478/)

Source: [http://xkcd.com/1478/](http://xkcd.com/1478/)
Motivation

- What do you want to do?
- How do you plan on doing it?
Motivation

- What do you want to do?
- How do you plan on doing it?
- What problems do you foresee?
What is a representative sample?
What is a representative sample?
The popular concept of a representative sample is that the sample is a *miniature* of the population.
However, what do we really want?
Representative Sample

However, what do we really want? We want to estimate a statistic of interest with a certain level of precision and if the level of precision is high enough, we say our estimation strategy is representative.
FRAMEWORK AND NOTATION
Finite population, Sample, and Sampling Design

\[ Y = \{ y_1, y_2, \ldots, y_k, \ldots, y_N \} \]  
finite population of size \( N \)

\[ U = \{ 1, 2, \ldots, k, \ldots, N \} \]  
sampling frame

\[ \delta \subset U \]  
sample of size \( n \)

\[ P(U) \]  
all possible subsets of \( U \)

The discrete probability distribution \( p(\cdot) \) over \( P(U) \) is called a sampling design and \( G = \{ \delta | \delta \in P(U), \ p(\delta) > 0 \} \) is called the support of \( p(\cdot) \) with

\[ \sum_{\delta \in G} p(\delta) = 1 \]

Hence, \( p : G \mapsto (0, 1] \).
Estimation

\[ \theta = f(Y) \]

\[ \hat{\theta} = f(Y, \lambda) \]

\[ \mathbb{E}\left(\hat{\theta}\right) = \sum_{\lambda \in G} p(\lambda) f(Y, \lambda) \]

\[ \mathbb{V}\left(\hat{\theta}\right) = \mathbb{E}\left(\hat{\theta}^2\right) - \mathbb{E}\left(\hat{\theta}\right)^2 \]

\[ \text{MSE}\left(\hat{\theta}\right) = \mathbb{E}\left((\hat{\theta} - \theta)^2\right) = \left(\mathbb{E}\left(\hat{\theta}\right) - \theta\right)^2 + \mathbb{V}\left(\hat{\theta}\right) \]

\[ \mathbb{E}\left(\theta\right), \mathbb{V}\left(\theta\right), \text{and MSE}\left(\theta\right) \text{ are always with respect to the sampling design } p() \text{ and an estimator is said to be unbiased if} \]

\[ \mathbb{E}\left(\hat{\theta}\right) = \theta . \]
**Expectation and Variance of a Random Sample**

$S_k$  
number of times element $k$ is selected

$l_k = \begin{cases} 
1 & \text{if } k \in \mathcal{S} \\
0 & \text{else} 
\end{cases}$  
sampling indicator element $k$

$E(S_k) = \nu_k$  
expected selection frequency of element $k$

$E(S_k S_l) = \nu_{kl}$  
joint expectation of $S_k$ and $S_l$

$E(l_k) = \pi_k$  
inclusion probability of element $k$

$E(l_k l_l) = \pi_{kl}$  
joint expectation of $l_k$ and $l_l$

$\sum_{k \in \mathcal{U}} \nu_k = E(n)$  
expected sample size
Simple Random Sampling
Simple random sampling without replacement (SRS): Drawing \( n \) elements out of a urn without putting them back (i.e. \( S_k \geq l_k \)) and without remembering the order of the selected element.

\[
g = \binom{N}{n} \quad \text{(1)}
\]

\[
p(\mathcal{S}) = \binom{N}{n}^{-1} \quad \text{(2)}
\]

\[
\pi_k = \nu_k = \frac{n}{N} \quad \text{(3)}
\]

\[
\pi_{kl} = \nu_{kl} = \frac{n(n - 1)}{N(N - 1)} \quad \text{for} \ k \neq l \quad \text{(4)}
\]
Sample Mean with SRS

\[ \theta = \mu = \frac{1}{N} \sum_{k \in U} y_k, \quad \hat{\theta} = \bar{y} = \frac{1}{n} \sum_{k \in S} y_k, \quad \sigma^2 = \frac{1}{N} \sum_{k \in U} (y_k - \mu)^2, \quad V^2 = \sigma^2 \frac{N}{N - 1} \]
Sample Mean with SRS

\[ \theta = \mu = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k, \quad \hat{\theta} = \bar{y} = \frac{1}{n} \sum_{k \in \mathcal{A}} y_k, \quad \sigma^2 = \frac{1}{N} \sum_{k \in \mathcal{U}} (y_k - \mu)^2, \quad V^2 = \sigma^2 \frac{N}{N - 1} \]

Expected value

\[
E(\bar{y}) = E \left( \sum_{k \in \mathcal{U}} S_k \frac{y_k}{n} \right) \\
= \frac{1}{n} \sum_{k \in \mathcal{U}} E(S_k) y_k \\
= \frac{1}{n} \sum_{k \in \mathcal{U}} \pi_k y_k \\
= \frac{1}{N} \sum_{k \in \mathcal{U}} y_k
\]
**Sample Mean with SRS**

\[
\theta = \mu = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k, \quad \hat{\theta} = \bar{y} = \frac{1}{n} \sum_{k \in \mathcal{S}} y_k, \quad \sigma^2 = \frac{1}{N} \sum_{k \in \mathcal{U}} (y_k - \mu)^2, \quad V^2 = \sigma^2 \frac{N}{N-1}
\]

**Expected value**

\[
\mathbb{E}(\bar{y}) = \mathbb{E}\left( \frac{1}{n} \sum_{k \in \mathcal{U}} S_k \frac{y_k}{n} \right) = \frac{1}{n} \sum_{k \in \mathcal{U}} \mathbb{E}(S_k) y_k = \frac{1}{n} \sum_{k \in \mathcal{U}} \pi_k y_k = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k
\]

**Variance**

\[
\text{Var}(\bar{y}) = \text{Var}\left( \frac{1}{n} \sum_{k \in \mathcal{U}} S_k \frac{y_k}{n} \right) = \frac{1}{n^2} \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \text{COV}(S_k, S_l) y_k y_l = -\frac{1}{2} \frac{1}{n^2} \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} (\pi_{kl} - \pi_k \pi_l) (y_k - y_l)^2 = \frac{N-n}{N-1} \frac{\sigma^2}{n} = \left(1 - \frac{n}{N}\right) \frac{V^2}{n}
\]
Complex Sampling Designs - Stratification
Features of Sampling Designs

- Stratification
- Cluster Sampling: Not elementary units are selected but clusters containing multiple elements.
- Multistage Sampling: The population is structured by hierarchically ordered clusters that are nested within each other. The sampling procedure has multiple selecting stages.
The universe $\mathcal{U}$ is decomposed into $H$ non-overlapping groups, $\mathcal{U}_1, \ldots, \mathcal{U}_H$, called strata.

- $\mathcal{U} = \bigcup_{h=1}^{H} \mathcal{U}_h$, where set $\mathcal{U}_h$ is the $h$-th strata.
- A sample $\mathcal{S}_h$ is selected from $\mathcal{U}_h$ according to a design $p_h(\cdot)$, for all $h = 1, \ldots, H$.
- The number of elements in $\mathcal{U}_h$ is called stratum size and denote with $N_h$.
- The number of elements in $\mathcal{S}_h$ is denoted with $n_h$. 
In stratified random sampling the sub-populations are called strata. For the \( h \)-th stratum we get:

\[
\mu_h = \frac{1}{N_h} \sum_{k=1}^{N_h} y_{kh}
\]

mean of stratum \( h \)

\[
\sigma^2_h = \frac{1}{N_h} \sum_{k=1}^{N_h} (y_{kh} - \mu_h)^2
\]

variance of stratum \( h \)

\[
V_h^2 = \sigma^2_h \frac{N_h}{N_h - 1}
\]

Where \( y_{kh} \) as the \( k \)-th element in the \( h \)-th stratum. Sampling from stratified populations is called stratified random sampling (StrRS).
A Population of 100 elements is stratified into $H = 6$ strata.

<table>
<thead>
<tr>
<th>h=1</th>
<th>h=2</th>
<th>h=3</th>
<th>h=4</th>
<th>h=5</th>
<th>h=6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

16 / 23
Stratification

A Population of 100 elements is stratified into $H = 6$ strata. 14 elements are selected from the population and their allocation is given by $n_1 = 2 \quad n_2 = 3 \quad n_3 = 2 \quad n_4 = 3 \quad n_5 = 3 \quad n_6 = 2$
Estimator for the mean:

\[ \bar{y}_{\text{str}} = \sum_{h=1}^{H} \gamma_h \bar{y}_h \]

where \( \gamma_h = \frac{N_h}{N} \) and \( E(\bar{y}_{\text{str}}) = \mu \) for SRS and SRSWR within each stratum.

Variance and variance estimator:

\[ V(\bar{y}_{\text{str}})_{\text{SRS}} = \sum_{h=1}^{H} \frac{N_h - n_h}{N_h} \gamma_h^2 \frac{V_h^2}{n_h} \]

\[ \hat{V}(\bar{y}_{\text{str}})_{\text{SRS}} = \sum_{h=1}^{H} \frac{N_h - n_h}{N_h} \gamma_h^2 \frac{s_h^2}{n_h} \]

\[ s_h^2 = \frac{1}{n_h - 1} \sum_{k \in \Delta_h} (y_k - \bar{y}_h)^2 \]
Why should stratification be used?

To reduce the sampling variance of estimators.

Sometimes it is necessary because of organizational reasons (e.g., no joint sampling frame).

A good set of variables needs to be found for stratification.

The number of strata has to be decided.

Achieve proportionality between sample and population (i.e., the frame).

Fulfill precision constraints for certain estimation domains.
Issues with Stratification

- Why should stratification be used?
  - To reduce the sampling variance of estimators.
  - Sometimes it is necessary because of organizational reasons (e.g. no joint sampling frame).
Issues with Stratification

- Why should stratification be used?
  - To reduce the sampling variance of estimators.
  - Sometimes it is necessary because of organizational reasons (e.g. no joint sampling frame).
- How should the population be stratified?
**Issues with Stratification**

- Why should stratification be used?
  - To reduce the sampling variance of estimators.
  - Sometimes it is necessary because of organizational reasons (e.g. no joint sampling frame).

- How should the population be stratified?
  - A *good* set of variables needs to be found for stratification.
  - The number of strata has to be decided.
Issues with Stratification

- Why should stratification be used?
  - To reduce the sampling variance of estimators.
  - Sometimes it is necessary because of organizational reasons (e.g. no joint sampling frame).

- How should the population be stratified?
  - A *good* set of variables needs to be found for stratification.
  - The number of strata has to be decided.
Issues with Stratification

- Why should stratification be used?
  - To reduce the sampling variance of estimators.
  - Sometimes it is necessary because of organizational reasons (e.g. no joint sampling frame).
- How should the population be stratified?
  - A good set of variables needs to be found for stratification.
  - The number of strata has to be decided.
- How should the overall sample size be allocated to the strata?
Issues with Stratification

- Why should stratification be used?
  - To reduce the sampling variance of estimators.
  - Sometimes it is necessary because of organizational reasons (e.g. no joint sampling frame).
- How should the population be stratified?
  - A good set of variables needs to be found for stratification.
  - The number of strata has to be decided.
- How should the overall sample size be allocated to the strata?
  - Achieve proportionality between sample and population (i.e. the frame)
  - Fulfill precision constraints for certain estimation domains
### Defining the Strata

#### Table: Population ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between strata</td>
<td>$H - 1$</td>
<td>$SSB = \sum_{h=1}^{H} N_h (\mu_h - \mu)^2$</td>
</tr>
<tr>
<td>Within strata</td>
<td>$N - H$</td>
<td>$SSW = \sum_{h=1}^{H} (N_h - 1) V_h^2$</td>
</tr>
<tr>
<td>Total, about $\mu_y$</td>
<td>$N - 1$</td>
<td>$SSTO = (N - 1) V^2$</td>
</tr>
</tbody>
</table>

The more homogeneous the strata are the higher is the gain in efficiency from using stratified simple random sample sampling (StrSRS) instead of SRS. Because then SSW (variance within) is considerably small in contrast to SSB (variance between). This is called the effect of stratification.
Allocation Methods

For all $h = 1, \ldots, H$

$$n_h = \begin{cases} \frac{n}{H} & \text{equal allocation} \\ \frac{N_h n}{N} & \text{proportional allocation} \\ \frac{N_h V_h}{\sum_{h=1}^{H} N_h V_h} n & \text{optimal allocation} \\ \frac{c \bar{c}_h}{\sum_{h=1}^{H} N_h V_h \sqrt{\bar{c}_h}} & \text{cost-optimal allocation} \end{cases}$$

where $\bar{c}_h$ are average cost of selecting a element from stratum $h$ and $c = \sum_{h=1}^{H} n_h \bar{c}_h$ are the total costs of the survey. For the cost-optimal allocation $c$ is given, not $n$. 
**On Proportional Allocation**

If \( n_h = \frac{N_h}{N} n \)

\[
V(\bar{y}_{str})_{StrSRS} = \left( \frac{N - n}{N} \right) \frac{1}{n} \sum_{h=1}^{H} N_h V^2_h \quad \text{and}
\]

\[
V(\bar{y})_{SRS} = \left( \frac{N - n}{N} \right) \frac{1}{n(n-1)} (SSW + SSB)
\]

\[
= V(\bar{y}_{str})_{StrSRS} + \left( \frac{N - n}{N} \right) \frac{1}{n(n-1)} \left[ SSB - \sum_{h=1}^{H} \frac{N - N_h}{N} V^2_h \right].
\]

Thus, StrSRS with prop. allocation will always result in an equal or smaller variance than SRS if

\[
SSB > \sum_{h=1}^{H} \frac{N - N_h}{N} V^2_h.
\]
L. Cox.
A Constructive Procedure for Unbiased Controlled Rounding. 

L. Gabler, A. Quatember.
Repräsentativität von Subgruppen bei geschichteten Zufallsstrichproben. 

S. Lohr.
Sampling: Design and Analysis. 

T. Lumley.
Complex Surveys: A Guide to Analysis Using R. 

C.-E. Särndal, B. Swensson, & J. Wretman.
Model Assisted Survey Sampling 
Practical Tools for Designing and Weighting Survey Samples.
Complex Sampling Designs - Cluster Sampling
A Population of 100 elements is clustered into $N_i = 6$ clusters.
A Population of 100 elements is clustered into $N_i = 6$ clusters 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 too much over the area and travel costs of interviewers would be too 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 too much over a certain area and travel costs of interviewers would be too 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 too much over a certain area and travel costs of interviewers would be too 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**

\[ y_{ki} = \text{value of the variable of interest for the } k\text{-th SSU in the } i\text{-th PSU} \]

\[ N_l = \text{number of PSU's in the population} \]

\[ N_i = \text{number of SSU's in the } i\text{-th PSU} \]

\[ N = \text{total number of SSU's in the Population} \]

\[ U = \text{set of SSU's in the population} \]

\[ U_l = \text{set of PSU's in the population} \]

\[ U_i = \text{set of SSU's in the } i\text{-th PSU} \]

\[ n_l = \text{number of PSU's in the sample} \]

\[ n_i = \text{number of SSU's in the sample from the } i\text{-th PSU} \]

\[ s_l = \text{sample of PSU's} \]

\[ s_i = \text{sample SSU's from the } i\text{-th PSU} \]

\[ p_l(\cdot) = \text{sampling design of the PSU's} \]
Estimation in Case of $p_I(\cdot) = \text{SRS}$

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

$$\bar{y}_{\text{SRCS}} = \frac{N_l}{N} \sum_{i \in \Delta_l} \frac{\tau_i}{n_l}$$

with variance

$$V(\bar{y})_{\text{SRCS}} = \frac{N_l^2}{N^2} \left(1 - \frac{n_l}{N_l}\right) \frac{V^2_\tau}{n_l},$$

where $V^2_\tau = \frac{1}{N_l - 1} \sum_{i \in \mathcal{U}_l} (\tau_i - \mu_\tau)^2$ and $\mu_\tau = \sum_{i \in \mathcal{U}_l} \frac{\tau_i}{N_l}$. 
Estimation in Case of \( p_I(.) = \text{SRS} \)

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

\[
\bar{y}_{\text{SRCS}} = \frac{N_I}{N} \sum_{i \in \delta_I} \frac{\tau_i}{n_i}
\]

An unbiased variance estimator is

\[
\hat{V}(\bar{y}_{\text{SRCS}})_{\text{SRS}} = \frac{N_I^2}{N^2} \left( 1 - \frac{n_I}{N_I} \right) \frac{s^2_\tau}{n_I},
\]

where

\[
s^2_\tau = \frac{1}{n_I - 1} \sum_{i \in \delta_I} (\tau_i - \bar{\tau})^2.
\]

with \( \bar{\tau} = \sum_{i \in \delta_I} \frac{\tau_i}{n_I} \).
**Table:** Two Variations of a Population Composed of 5 clusters of Size 5

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>$\mu_i.$</th>
<th>$V^2_i.$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>4.0</td>
<td>5.0</td>
<td>3.0</td>
<td>2.5</td>
</tr>
<tr>
<td>2</td>
<td>6.0</td>
<td>7.0</td>
<td>8.0</td>
<td>9.0</td>
<td>10.0</td>
<td>8.0</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>11.0</td>
<td>12.0</td>
<td>13.0</td>
<td>14.0</td>
<td>15.0</td>
<td>13.0</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>16.0</td>
<td>17.0</td>
<td>18.0</td>
<td>19.0</td>
<td>20.0</td>
<td>18.0</td>
<td>2.5</td>
</tr>
<tr>
<td>5</td>
<td>21.0</td>
<td>22.0</td>
<td>23.0</td>
<td>24.0</td>
<td>25.0</td>
<td>23.0</td>
<td>2.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mu_j.$</th>
<th>11.0</th>
<th>12.0</th>
<th>13.0</th>
<th>14.0</th>
<th>15.0</th>
<th>13.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V^2_j.$</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>54.2</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>$\mu_i.$</th>
<th>$V_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>4.0</td>
<td>5.0</td>
<td>3.0</td>
<td>2.5</td>
</tr>
<tr>
<td>2</td>
<td>6.0</td>
<td>7.0</td>
<td>8.0</td>
<td>9.0</td>
<td>10.0</td>
<td>8.0</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>11.0</td>
<td>12.0</td>
<td>13.0</td>
<td>14.0</td>
<td>15.0</td>
<td>13.0</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>16.0</td>
<td>17.0</td>
<td>18.0</td>
<td>19.0</td>
<td>20.0</td>
<td>18.0</td>
<td>2.5</td>
</tr>
<tr>
<td>5</td>
<td>21.0</td>
<td>22.0</td>
<td>23.0</td>
<td>24.0</td>
<td>25.0</td>
<td>23.0</td>
<td>2.5</td>
</tr>
<tr>
<td>$\mu_j.$</td>
<td>11.0</td>
<td>12.0</td>
<td>13.0</td>
<td>14.0</td>
<td>15.0</td>
<td>13.0</td>
<td></td>
</tr>
<tr>
<td>$V_j^2$</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>54.2</td>
<td></td>
</tr>
</tbody>
</table>

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 $n_i = 2$
   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

<table>
<thead>
<tr>
<th></th>
<th>SRS</th>
<th>SRCS a</th>
<th>SRCS b</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>13.0</td>
<td>13.0</td>
<td>13.0</td>
</tr>
<tr>
<td>var</td>
<td>3.2</td>
<td>0.8</td>
<td>18.9</td>
</tr>
</tbody>
</table>

True values: $\mu = 13$, $\sigma^2_y_{SRS} = 3.25$, $\sigma^2_y_{SRCS} = 0.75$.

Bias is not an issue, however variance is. If the cluster were strata, which stratification would you use, columns or rows? What good is for stratified sampling, i.e. low SSW, is bad for cluster sampling and vice versa.
Simple Cluster Sampling

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

**Table:** Results from the Simulation Study

<table>
<thead>
<tr>
<th></th>
<th>SRS</th>
<th>SRCS a</th>
<th>SRCS b</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>13.0</td>
<td>13.0</td>
<td>13.0</td>
</tr>
<tr>
<td>var</td>
<td>3.2</td>
<td>0.8</td>
<td>18.9</td>
</tr>
</tbody>
</table>

$\mu = 13$

$V(\bar{y})_{SRS} = 3.25$ SRS

$V(\bar{y})_{SRCS}^{a}_{SRS} = 0.75$ SRCS a

$V(\bar{y})_{SRCS}^{b}_{SRS} = 18.75$ SRCS b

Bias is not an issue, however variance is.
Comparing the mean and variance of $\bar{y}$ and $\bar{y}_{SRCS}$ after 100,000 samples:

Table: Results from the Simulation Study

<table>
<thead>
<tr>
<th></th>
<th>SRS</th>
<th>SRCS a</th>
<th>SRCS b</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>13.0</td>
<td>13.0</td>
<td>13.0</td>
</tr>
<tr>
<td>var</td>
<td>3.2</td>
<td>0.8</td>
<td>18.9</td>
</tr>
</tbody>
</table>

$\mu = 13$

$V(\bar{y})_{SRS} = 3.25$ SRS

$V(\bar{y}_{SRCS})_{SRS} = 0.75$ SRCS a

$V(\bar{y}_{SRCS})_{SRS} = 18.75$ SRCS b

True values:

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 $\bar{y}_{SRCS}$ after 100,000 samples:

<table>
<thead>
<tr>
<th>Table: Results from the Simulation Study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>mean</td>
</tr>
<tr>
<td>var</td>
</tr>
</tbody>
</table>

True values:

$\mu = 13$

$V(\bar{y})_{SRS} = 3.25$ SRS

$V(\bar{y}_{SRCS})_{SRS} = 0.75$ SRCS a

$V(\bar{y}_{SRCS})_{SRS} = 18.75$ SRCS b

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.
Complex Sampling Designs - Two Stage Sampling
**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 $N_i = 6$ 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.

<table>
<thead>
<tr>
<th>i=1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>i=2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>■</td>
<td>■</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>i=3</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>i=4</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>i=5</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>i=6</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Two Stage Sampling**

**First stage** A sample \( \mathcal{S}_1 \) of PSU’s is drawn from \( \mathcal{U}_1 \) according to some sampling design \( p_1(\cdot) \)

**Second stage** For every \( i \in \mathcal{S}_1 \) a sample \( \mathcal{S}_i \) of SSU’s is selected from \( \mathcal{U}_i \) according to some design \( p_i(\cdot|\mathcal{S}_1) \)

The resulting sample of SSU’s is denote \( \mathcal{S} = \bigcup_{i \in \mathcal{S}_1} \mathcal{S}_i \). 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

\[
\pi_k = \pi_{1i} \pi_{k|i},
\]

where \( \pi_{1i} \) 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.
Designs $p_i(\cdot)$ and $p_i(\cdot | \delta_i)$ are both SRS. Since not all SSU’s in the sampled PSU’s are surveyed $\tau_i$ has to be estimated by

$$\hat{\tau}_i = \frac{N_i}{n_i} \sum_{k \in \delta_i} y_{ki}.$$  

An unbiased estimator for the population mean is

$$\bar{y}_{2SRS} = \frac{N_i}{N} \sum_{i \in \delta_i} \frac{\hat{\tau}_i}{n_i}$$
Designs \( p_i(\cdot) \) and \( p_i(\cdot|\delta_i) \) are both SRS. Since not all SSU’s in the sampled PSU’s are surveyed \( \tau_i \) has to be estimated by 
\[
\hat{\tau}_i = \frac{N_i}{n_i} \sum_{k \in \delta_i} y_{ki}.
\]
An unbiased estimator for the population mean is 
\[
\bar{y}_{2\text{SRS}} = \frac{N_i}{N} \sum_{i \in \delta} \frac{\hat{\tau}_i}{n_i}
\]
with variance 
\[
V(\bar{y}_{2\text{SRS}})_{\text{SRS}} = \frac{1}{N^2} \left( N_i^2 \left( 1 - \frac{n_i}{N_i} \right) \frac{V_{\tau}^2}{n_i} + \frac{N_i}{n_i} \sum_{i \in U_i} N_i^2 \left( 1 - \frac{n_i}{N_i} \right) \frac{V_i^2}{n_i} \right),
\]
where 
\[
V_i^2 = \frac{1}{N_{i-1}} \sum_{k \in U_i} (y_{ki} - \mu_i)^2 \text{ with } \mu_i = \sum_{k \in U_i} \frac{y_{ki}}{N_i}.
\]
Designs \( p_i(.) \) and \( p_i(.|\Delta_i) \) are both SRS. Since not all SSU's in the sampled PSU's are surveyed \( \tau_i \) has to be estimated by

\[
\hat{\tau}_i = \frac{N_i}{n_i} \sum_{k \in \Delta_i} y_{ki}.
\]

An unbiased estimator for the population mean is

\[
\bar{y}_{2SRS} = \frac{N_i}{N} \sum_{i \in \Delta_i} \frac{\hat{\tau}_i}{n_i}
\]

An unbiased variance estimator is given by

\[
\hat{V} (\bar{y}_{2SRS})_{SRS} = \frac{1}{N^2} \left( N_i^2 \left( 1 - \frac{n_i}{N_i} \right) \frac{s^2_{\hat{\tau}}}{n_i} + \frac{N_i}{n_i} \sum_{i \in \Delta_i} N_i^2 \left( 1 - \frac{n_i}{N_i} \right) \frac{s^2_i}{n_i} \right)
\]

where \( s^2_{\hat{\tau}} = \frac{1}{n_i - 1} \sum_{i \in \Delta_i} (\hat{\tau}_i - \bar{\tau})^2 \) with \( \bar{\tau} = \sum_{i \in \Delta_i} \frac{\hat{\tau}_i}{n_i} \) and

\( s^2_i = \frac{1}{n_i - 1} \sum_{k \in \Delta_i} (y_{ki} - \bar{y}_i)^2 \) with \( \bar{y}_i = \sum_{k \in \Delta_i} \frac{y_{ki}}{n_i} \).
Complex Sampling Designs - Unequal Probability Sampling
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 $\tau = \sum_{k \in U} y_k$ is given by

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

which is also known as *Horvitz-Thompson* (HT) or $\pi$-estimator.
A Generic Design Based Estimator

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

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

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

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

which can be estimated by

$$\hat{V}(\hat{\tau}_\pi) = \sum_{k \in \varDelta} \sum_{l \in \varDelta} \left( \pi_{kl} - \pi_k \pi_l \right) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$
A design unbiased estimator for the total \( \tau = \sum_{k \in U} y_k \) is given by

\[
\hat{\tau}_\pi = \sum_{k \in \mathcal{D}} \frac{y_k}{\pi_k},
\]

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

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

which can be estimated by

\[
\hat{V}(\hat{\tau}_\pi)^2 = -\frac{1}{2} \sum_{k \in \mathcal{D}} \sum_{l \in \mathcal{D}} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2.
\]
A design unbiased estimator for the total \( \tau = \sum_{k \in \mathcal{U}} y_k \) is given by

\[
\hat{\tau}_\pi = \sum_{k \in \mathcal{D}} \frac{y_k}{\pi_k},
\]

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

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

which can be estimated by

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

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 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 $\nu_k$ 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 $\hat{V}$, $\hat{\tau}_\pi$ would even be zero.
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 \( \nu_i = \frac{N_i}{N} n_i > 1 \). The integer part of \( \nu_i \) indicates how many sampling points are \textit{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

\[
\frac{N_i}{N} n_i \times \frac{n_*}{N_i} = \frac{n_i n_*}{N}.
\]

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 \((\pi_k^{-1})\) do not imply the same sampling variance. Different designs can be used to select samples with same \(\pi_k\)'s, however their \(\pi_{kl}\)'s might be very different and so is their associated sampling variance.
Design Effect
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

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

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**

<table>
<thead>
<tr>
<th>Source</th>
<th>sf</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between cluster</td>
<td>$N_i - 1$</td>
<td>$SSB = \sum_{i=1}^{N_i} N_i(\mu_i - \mu)^2$</td>
</tr>
<tr>
<td>Within cluster</td>
<td>$N - N_i$</td>
<td>$SSW = \sum_{i=1}^{N_i} (N_i - 1) V_i^2$</td>
</tr>
<tr>
<td>Total</td>
<td>$N - 1$</td>
<td>$SSTO = (N - 1) V^2$</td>
</tr>
</tbody>
</table>
### Design Effect for a Cluster Sample

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>$\mu_i.$</th>
<th>$V_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>4.0</td>
<td>5.0</td>
<td>3.0</td>
<td>2.5</td>
</tr>
<tr>
<td>2</td>
<td>6.0</td>
<td>7.0</td>
<td>8.0</td>
<td>9.0</td>
<td>10.0</td>
<td>8.0</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>11.0</td>
<td>12.0</td>
<td>13.0</td>
<td>14.0</td>
<td>15.0</td>
<td>13.0</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>16.0</td>
<td>17.0</td>
<td>18.0</td>
<td>19.0</td>
<td>20.0</td>
<td>18.0</td>
<td>2.5</td>
</tr>
<tr>
<td>5</td>
<td>21.0</td>
<td>22.0</td>
<td>23.0</td>
<td>24.0</td>
<td>25.0</td>
<td>23.0</td>
<td>2.5</td>
</tr>
<tr>
<td>$\mu_j$</td>
<td>11.0</td>
<td>12.0</td>
<td>13.0</td>
<td>14.0</td>
<td>15.0</td>
<td>13.0</td>
<td></td>
</tr>
<tr>
<td>$V_j^2$</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>54.2</td>
<td></td>
</tr>
</tbody>
</table>

**Table**: Design Effects

<table>
<thead>
<tr>
<th></th>
<th>SRCS a</th>
<th>SRCS b</th>
</tr>
</thead>
<tbody>
<tr>
<td>deff</td>
<td>0.23077</td>
<td>5.76923</td>
</tr>
</tbody>
</table>
Issues associated with Estimation of $deff$. 

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

Find suitable estimators for both the denominator and the numerator.
Issues associated with estimation of $deff$. 

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

Find suitable estimators for both

the denominator and

the numerator

Treat the data as if it had arisen from SRS for estimation of the numerator (but using available weights).
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 $Y$ is assumed to follow the *common parameter model*,

\[
\begin{align*}
E(y_{ki})_M &= \mu \\
V(y_{ki})_M &= \sigma^2 \\
\text{COV}(y_{ki}, y_{k'i'})_M &= \begin{cases} 
\sigma^2 \rho & \text{for } k \neq k', i = i' \\
0 & \text{otherwise.}
\end{cases}
\]

$\rho$ 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.
Estimation of \(\text{deff}\) under Cluster Sampling

The considered estimator is

\[
\bar{y}_w = \frac{\sum_{k \in \Delta} w_k y_k}{\sum_{k \in \Delta} w_k},
\]

where \(w_k\) is the survey associated with \(k\)-th element, e.g.

\[
w_k = \begin{cases} 
\pi_k^{-1} & \text{for } k \in \Delta \\
0 & \text{else}
\end{cases}
\]

Under the model-based approach, \(\text{deff}\) can be displayed as the product of two factors:

- \(\text{deff}_c\) design effect due to clustering
- \(\text{deff}_\rho\) 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 \( k \in U \). Potential gains in precision by using proportional to size design are not considered, quite the contrary, unequal weight will increase the variance of \( \bar{y}_w \) in this setting.
Estimation of $deff_c$

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

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

with

$\bar{b}$ as the average cluster size $\frac{N}{N_i} \left( \frac{n}{n_i} \right)$ or an estimator for it and

$\hat{\rho}$ as an appropriate estimator of $\rho$.

When $w_k$’s vary we have

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

where $b^* = \frac{\sum_{i \in M} \left( \sum_{k \in \Delta_i} w_k \right)^2}{\sum_{k \in \Delta} w_k^2}$ is a kind of weighted average cluster size.
The design effect due to unequal weights is

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

Obviously, if \( w_k \)'s are constant, \( deff_p = 1 \)
Behavior of $\text{deff}_c$
There are many different ways to estimate $\rho$. The classical ANOVA estimator is:

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

where $\widehat{MSB} = (n_1 - 1)^{-1} \sum_{i \in \mathcal{P}} n_i (\bar{y}_i - \bar{y})^2$, $\widehat{MSW} = (n - n_1)^{-1} \sum_{i \in \mathcal{P}} (n_i - 1) s_i^2$, and

$$K = (n_1 - 1)^{-1} \left( n - \sum_{i \in \mathcal{P}} \frac{n_i^2}{n} \right).$$
An Estimators of $\rho$

There are many different ways to estimate $\rho$. The classical ANOVA estimator is:

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

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

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

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

<table>
<thead>
<tr>
<th></th>
<th>SRCS a</th>
<th>SRCS b</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>-0.01626</td>
<td>0.99136</td>
</tr>
<tr>
<td>$deff$</td>
<td>0.93496</td>
<td>4.96546</td>
</tr>
</tbody>
</table>

### Table: Design Effects

<table>
<thead>
<tr>
<th></th>
<th>SRCS a</th>
<th>SRCS b</th>
</tr>
</thead>
<tbody>
<tr>
<td>$deff$</td>
<td>0.23077</td>
<td>5.76923</td>
</tr>
</tbody>
</table>


Sampling and Estimation
Part 3: Calibration Weights

Stefan Zins$^1$ and Matthias Sand$^2$

April 24, 2017

$^1$Stefan.Zins@gesis.org
$^2$Matthias.Sand@gesis.org
Survey Errors
Methods to Handle Missing Data

Procedures based on the **available cases** only, i.e., only those cases that are completely recorded for the variables of interest.
Methods to Handle Missing Data

Procedures based on the available cases only, i.e., only those cases that are completely recorded for the variables of interest.

Weighting procedures that adjust design weights to compensate the bias that a MAR non-response might inflict on HT-type estimators.
Methods to Handle Missing Data

Procedures based on the available cases only, i.e., only those cases that are completely recorded for the variables of interest.

Weighting procedures that adjust design weights to compensate the bias that a MAR non-response might inflict on HT-type estimators.

Single imputation and correction of the variance estimates to account for imputation uncertainty.
Methods to Handle Missing Data

Procedures based on the available cases only, i.e., only those cases that are completely recorded for the variables of interest.

Weighting procedures that adjust design weights to compensate the bias that a MAR non-response might inflict on HT-type estimators.

Single imputation and correction of the variance estimates to account for imputation uncertainty.

Multiple imputation (MI).
Methods to Handle Missing Data

Procedures based on the **available cases** only, i.e., only those cases that are completely recorded for the variables of interest.

**Weighting procedures** that adjust design weights to compensate the bias that a MAR non-response might inflict on HT-type estimators.

**Single imputation** and correction of the variance estimates to account for imputation uncertainty.

**Multiple imputation** (MI).

Methods for handling coverage errors are not so widely spread, simply because there is often no reliable auxiliary information on just the target population. However if there is, it can receive a treatment similar to that of weighting by non-response.
Missing data is the norm, rather than the exception!
Missingness may be either

MCAR: missing completely at random, every unit has the same response propensity (RP); respondents are a random sample of the initial sample.

MAR: missing at random, or RP depends on auxiliary variables X; can be modeled, if X is known for both respondents & non-respondents.

MNAR: missing not at random, RP depends on variables of interest Y; cannot be modeled, because Y is not known for non-respondents.

In multivariate analysis, 30% to 40% of the data are often lost with case deletion, assuming MCAR.
Missing data is the norm, rather than the exception!
Missingness may be either

**MCAR** missing completely at random,
  every unit has same response propensity (RP);
  respondents are a random sample of the initial sample

**MAR** missing at random,
  response propensity (RP) depends on auxiliary variables
  can be modeled, if variables known for both respondents & non-respondents

**MNAR** missing not at random
  response propensity (RP) depends on variables of interest
  cannot be modeled, because variables not known for non-respondents

In multivariate analysis 30% to 40% of the data are often lost
with case deletion, assuming MCAR!
Missing data is the norm, rather than the exception!

Missingness may be either

**MCAR**  missing completely at random,
- every unit has same response propensity (RP);
- respondents are a random sample of the initial sample

**MAR**  missing at random, or
- RP depends on auxiliary variables $X$;
  can be modeled, if $X$ is known for both respondents & non-respondents

**MNAR**  missing not at random
- RP depends on variables of interest $Y$;
  cannot be modeled, because $Y$ not known for non-respondents

In multivariate analysis 30% to 40% of the data are often lost with case deletion, assuming MCAR!
Missing data is the norm, rather than the exception!

Missingness may be either

**MCAR** missing completely at random,
- every unit has same response propensity (RP);
- respondents are a random sample of the initial sample

**MAR** missing at random, or
- RP depends on auxiliary variables $\mathcal{X}$;
- can be modeled, if $\mathcal{X}$ is known for both respondents & non-respondents

**MNAR** missing not at random
- RP depends on variables of interest $\mathcal{Y}$;
- cannot be modeled, because $\mathcal{Y}$ not known for non-respondents

$\rightarrow$ In multivariate analysis 30% to 40% of the data are often lost with case deletion, assuming MCAR!
Weighting Methods
**Calibration approach**  The design weighs are calibrated to the totals of some auxiliary variables $\mathcal{X}$.

Sample estimates using the calibrated weights will exactly replicated those totals.
If the used auxiliary variables help to explain the response process the calibrated weight can reduce the non-response error.
Two-phase approach  The response process is modeled to obtain the response propensities $\psi_k$ for all $k \in \mathcal{S}$. The new weight of element $k$ is $\frac{d_k}{\psi_k}$. (Two phases: 1. Sampling $\rightarrow$ 2. Responding).

In addition the new weights $\frac{d_k}{\psi_k}$ might then also be calibrated.

Often used models are:

- Response homogeneity classes, every element in a class has the same probability to respond.
- Generalized liner models (probit, logit, log-log), treating response as a latent variable.
The calibration approach is more direct, as the design weights are directly calibrated without considering the response propensities. Also, if the same models are used for both the modeling of the response propensities and the calibration, the two approaches can be equivalent.
Weights

Generic estimators for a total and a mean

\[ \hat{\tau}_w = \sum_{k \in \Delta} w_k y_k \]  
and \[ \bar{y}_w = \frac{\sum_{k \in \Delta} w_k y_k}{\sum_{k \in \Delta} w_k}, \]

where \( w_k \) is the survey weight of element \( k \), with
Weights

Generic estimators for a total and a mean

\[ \hat{t}_w = \sum_{k \in \Delta} w_k y_k \quad \text{and} \quad \bar{y}_w = \frac{\sum_{k \in \Delta} w_k y_k}{\sum_{k \in \Delta} w_k}, \]

where \( w_k \) is the survey weight of element \( k \), with

\[ w_k = \begin{cases} d_k g_k & \text{for } k \in \Delta \\ 0 & \text{else} \end{cases} . \]
Weights

\[ w_k = \begin{cases} 
  d_k g_k & \text{for } k \in \mathcal{D} \\
  0 & \text{else}
\end{cases} . \]

Sometimes called base weights or design weights, the inverse of inclusion probabilities \( d_k = \pi^{-1} \) is usually the first step in weighting. If we have \( g_k = 1 \), the \( \hat{\tau}_w \) would be the HT estimator or \( \pi \)-estimator. The factor \( g_k \) adjusts the design weights to reduce

- the sampling error (i.e. variance),
- the non-response error, and
- the coverage error

of estimator \( \hat{\tau}_w \) or \( \bar{y}_w \). Thereby the \( w_k \)'s should not deviate too much from the \( d_k \)'s as these weights ensure an unbiased estimation.
The general idea is to exploit the relationship between auxiliary variables and the variable of interest to improve the efficiency of estimators.
The following problem is solved with weight calibration:
For a given design $p(.)$ and a sample $\bar{s}$ weights $w_k$ for all $k \in \Delta$ have to be found that minimize

$$\sum_{k \in \Delta} G_k(w_k, d_k, c_k),$$

subject to constraints

$$\sum_{k \in \Delta} w_k x_k = \sum_{k \in U} x_k = \tau_x$$

where $x_k = (x_{k1}, x_{k2}, \ldots, x_{kQ})^\top$ is a vector of $q$ auxiliary variables for element $k$. $G_k$ is a measure of distance between $w_k$ and $d_k$ and $c_k$ is a factor that can be freely chosen for additional flexibility.
To calculate the weights, the $x_k$'s are only needed for the elements in the net sample (i.e. typically only for the respondents), but $\tau_x$, their population totals, need to be known.

The auxiliary variables can be metric (e.g. income or age) or categorical (e.g. gender or age groups).

Depending on the choice of $G_k$, different calibration estimators can be obtained, some of the most common are:

- Post-stratification Estimator
- Raking Estimator
- Generalized Regression Estimator

Note that the $w_k$'s typically depend on the sample $\mathcal{S}$, in contrast to the $d_k$, which are given by the sampling design.
Post-stratification is typically used if only categorical auxiliary variables are available. It is implemented by forming weighting cells by crossing all categories of the auxiliary variables. These weighting cells are the post-strata $\mathcal{U}_q$ with $q = 1, \ldots, Q$. The weights are then adjusted to replicate the counts in these cells. For $k \in \mathcal{U}_q$ we have

$$g_k = \frac{\tau_{x_q}}{\hat{\tau}_{x_q}},$$

where $\tau_{x_q} = \sum_{k \in \mathcal{U}} x_{kq}$ and

$$x_{kq} = \begin{cases} 1 & \text{if } k \in \mathcal{U}_q \\ 0 & \text{else} \end{cases}.$$

$\hat{\tau}_{x_q}$ its estimator for $\tau_{x_q}$ based on the design weights. The auxiliary variables are the post-stratum indicators, i.e. $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kQ})^\top$. An adjustment to the totals of a metric variable within the post-strata would also be possible.
## Post-stratification

**Table:** Population Counts $\tau_{xq}$ for Hair and Eye Colour

<table>
<thead>
<tr>
<th></th>
<th>Brown</th>
<th>Blue</th>
<th>Hazel</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>68</td>
<td>20</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Brown</td>
<td>119</td>
<td>84</td>
<td>54</td>
<td>29</td>
</tr>
<tr>
<td>Red</td>
<td>26</td>
<td>17</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Blond</td>
<td>7</td>
<td>94</td>
<td>10</td>
<td>16</td>
</tr>
</tbody>
</table>

Beware, there must be at least one element in the sample from each post-stratum, otherwise we divide by zero!
**Post-stratification**

**Table:** Population Counts $\tau_{xq}$ for Hair and Eye Colour

<table>
<thead>
<tr>
<th></th>
<th>Brown</th>
<th>Blue</th>
<th>Hazel</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>68</td>
<td>20</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Brown</td>
<td>119</td>
<td>84</td>
<td>54</td>
<td>29</td>
</tr>
<tr>
<td>Red</td>
<td>26</td>
<td>17</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Blond</td>
<td>7</td>
<td>94</td>
<td>10</td>
<td>16</td>
</tr>
</tbody>
</table>

**Table:** Sample counts $\sum_{k \in \mathcal{S}} x_{kq}$ in a SRS with $n = 150$

<table>
<thead>
<tr>
<th></th>
<th>Brown</th>
<th>Blue</th>
<th>Hazel</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>14</td>
<td>7</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Brown</td>
<td>36</td>
<td>22</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>Red</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Blond</td>
<td>1</td>
<td>23</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Beware, there must be at least one element in the sample from each post-stratum, otherwise we divide by zero!
Table: Population Counts $\tau_{xq}$ for Hair and Eye Colour

<table>
<thead>
<tr>
<th></th>
<th>Brown</th>
<th>Blue</th>
<th>Hazel</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>68</td>
<td>20</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Brown</td>
<td>119</td>
<td>84</td>
<td>54</td>
<td>29</td>
</tr>
<tr>
<td>Red</td>
<td>26</td>
<td>17</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Blond</td>
<td>7</td>
<td>94</td>
<td>10</td>
<td>16</td>
</tr>
</tbody>
</table>

Table: Estimated totals $\hat{\tau}_{xq \pi} = \sum_{k \in \delta} x_{kq} d_k$

<table>
<thead>
<tr>
<th></th>
<th>Brown</th>
<th>Blue</th>
<th>Hazel</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>55.2533</td>
<td>27.6267</td>
<td>7.8933</td>
<td>7.8933</td>
</tr>
<tr>
<td>Brown</td>
<td>142.0800</td>
<td>86.8267</td>
<td>67.0933</td>
<td>19.7333</td>
</tr>
<tr>
<td>Red</td>
<td>27.6267</td>
<td>11.8400</td>
<td>3.9467</td>
<td>15.7867</td>
</tr>
<tr>
<td>Blond</td>
<td>3.9467</td>
<td>90.7733</td>
<td>3.9467</td>
<td>19.7333</td>
</tr>
</tbody>
</table>

Beware, there must be at least one element in the sample from each post-stratum, otherwise we divide by zero!
**Post-stratification**

**Table:** Population Counts $\tau_{xq}$ for Hair and Eye Colour

<table>
<thead>
<tr>
<th></th>
<th>Brown</th>
<th>Blue</th>
<th>Hazel</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>68</td>
<td>20</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Brown</td>
<td>119</td>
<td>84</td>
<td>54</td>
<td>29</td>
</tr>
<tr>
<td>Red</td>
<td>26</td>
<td>17</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Blond</td>
<td>7</td>
<td>94</td>
<td>10</td>
<td>16</td>
</tr>
</tbody>
</table>

**Table:** Post-stratification $g_k = \frac{\tau_{xq}}{\hat{\tau}_{xq}}$

<table>
<thead>
<tr>
<th></th>
<th>Brown</th>
<th>Blue</th>
<th>Hazel</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>1.2307</td>
<td>0.7239</td>
<td>1.9003</td>
<td>0.6334</td>
</tr>
<tr>
<td>Brown</td>
<td>0.8376</td>
<td>0.9674</td>
<td>0.8048</td>
<td>1.4696</td>
</tr>
<tr>
<td>Red</td>
<td>0.9411</td>
<td>1.4358</td>
<td>3.5473</td>
<td>0.8868</td>
</tr>
<tr>
<td>Blond</td>
<td>1.7736</td>
<td>1.0355</td>
<td>2.5338</td>
<td>0.8108</td>
</tr>
</tbody>
</table>

Beware, there must be at least one element in the sample from each post-stratum, otherwise we divide by zero!
In raking, only the marginal totals are needed, *not* the totals for all the cross-categories. Raking can be implemented as iterative post-stratification to adjust the design weights to the margins of the different auxiliary variables.
The design weights of a SRSC cluster sample of school districts are raked to variables school type (sttpe) and the accomplishment of the growth target (sch.wide).

### Table: Population Counts $\tau_{xq}$ for School Type (stype) and School Target (sch.wide)

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>472</td>
<td>3949</td>
<td>4421</td>
</tr>
<tr>
<td>H</td>
<td>334</td>
<td>421</td>
<td>755</td>
</tr>
<tr>
<td>M</td>
<td>266</td>
<td>752</td>
<td>1018</td>
</tr>
<tr>
<td>SUM</td>
<td>1072</td>
<td>5122</td>
<td>6194</td>
</tr>
</tbody>
</table>
Raking

```r
set.seed(-57844)
# selection the SRCS
apiclus <- apipop[apipop$dnum %in% sample(unique(apipop$dnum), 10),]
apiclus$fpc <- length(unique(apipop$dnum))

dclus1 <- svydesign(id=~dnum, data=apiclus, fpc=~fpc)
# initial weight
w1 <- weights(dclus1)
# convergence is declared if the maximum change in a table entry is less than 'eps'...
eps <- 1
# ... otherwise the process stops after 'maxit' iterations
maxit <- 100

tau_stype <- table(apipop$stype)
tau_sch.wide <- table(apipop$sch.wide)

# Raking (i.e. iterative post-stratification) for two variables
tab_x <- tab_y <- list()
```
for (i in 1:maxit) {
    ## Post-stratification to the first variable
    w1 <- split(w1, apiclus$stype)
    adj1 <- tau_stype/sapply(w1, sum)
    # new weight
    w1. <- w1 <- mapply(function(x, y) x * y, w1, adj1)
    # return to original order
    w1 <- unlist(w1.)
    names(w1) <- unlist(sapply(w1., names))
    w1 <- w1[as.character(sort(as.numeric(names(w1))))]
    tab_x[[i]] <- tapply(w1, list(apiclus$stype, apiclus$sch.wide), sum)

    ## Post-stratification to the second variable
    w2 <- split(w1, apiclus$sch.wide)
    adj2 <- tau_sch.wide/sapply(w2, sum)
    # new weight
    w2. <- w2 <- mapply(function(x, y) x * y, w2, adj2)
    # return to original order
    w2 <- unlist(w2.)
    names(w2) <- unlist(sapply(w2., names))
    w2 <- w2[as.character(sort(as.numeric(names(w2))))]
    tab_y[[i]] <- tapply(w2, list(apiclus$stype, apiclus$sch.wide), sum)

    if (i > 1) {
        tab.diff <- abs(tab_y[[i - 1]] - tab_y[[i]])

        if (max(tab.diff) < eps)
            break
    }
    w1 <- w2
}
**Table:** Estimated Totals $\hat{r}_{xq} = \sum_{k \in \delta} x_k q_k d_k$ from a SRCS of Districts (dname) with $n_i = 10$

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>984.1</td>
<td>4087.8</td>
<td>5071.9</td>
</tr>
<tr>
<td>H</td>
<td>378.5</td>
<td>302.8</td>
<td>681.3</td>
</tr>
<tr>
<td>M</td>
<td>378.5</td>
<td>832.7</td>
<td>1211.2</td>
</tr>
<tr>
<td>SUM</td>
<td>1741.1</td>
<td>5223.3</td>
<td>6964.4</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>SUM</td>
</tr>
<tr>
<td>-----</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>E</td>
<td>857.8</td>
<td>3563.2</td>
<td>4421.0</td>
</tr>
<tr>
<td>H</td>
<td>419.4</td>
<td>335.6</td>
<td>755.0</td>
</tr>
<tr>
<td>M</td>
<td>318.1</td>
<td>699.9</td>
<td>1018.0</td>
</tr>
<tr>
<td>SUM</td>
<td>1595.4</td>
<td>4598.6</td>
<td>6194.0</td>
</tr>
</tbody>
</table>
**Table:** Estimated Totals after Adjustment to 'sch.wide' in the 1 Iteration

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>576.4</td>
<td>3968.7</td>
<td>4545.1</td>
</tr>
<tr>
<td>H</td>
<td>281.8</td>
<td>373.7</td>
<td>655.6</td>
</tr>
<tr>
<td>M</td>
<td>213.8</td>
<td>779.5</td>
<td>993.3</td>
</tr>
<tr>
<td>SUM</td>
<td>1072.0</td>
<td>5122.0</td>
<td>6194.0</td>
</tr>
</tbody>
</table>
**Table:** Estimated Totals after Adjustment to 'stype' in the 2 Iteration

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>560.7</td>
<td>3860.3</td>
<td>4421.0</td>
</tr>
<tr>
<td>H</td>
<td>324.6</td>
<td>430.4</td>
<td>755.0</td>
</tr>
<tr>
<td>M</td>
<td>219.1</td>
<td>798.9</td>
<td>1018.0</td>
</tr>
<tr>
<td>SUM</td>
<td>1104.3</td>
<td>5089.7</td>
<td>6194.0</td>
</tr>
</tbody>
</table>
### Table: Estimated Totals after Adjustment to 'sch.wide' in the 2 Iteration

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>544.2</td>
<td>3884.9</td>
<td>4429.1</td>
</tr>
<tr>
<td>H</td>
<td>315.1</td>
<td>433.2</td>
<td>748.2</td>
</tr>
<tr>
<td>M</td>
<td>212.7</td>
<td>804.0</td>
<td>1016.7</td>
</tr>
<tr>
<td>SUM</td>
<td>1072.0</td>
<td>5122.0</td>
<td>6194.0</td>
</tr>
</tbody>
</table>
**Table:** Estimated Totals after Adjustment to ‘stype’ in the 3 Interation

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>543.3</td>
<td>3877.7</td>
<td>4421.0</td>
</tr>
<tr>
<td>H</td>
<td>317.9</td>
<td>437.1</td>
<td>755.0</td>
</tr>
<tr>
<td>M</td>
<td>212.9</td>
<td>805.1</td>
<td>1018.0</td>
</tr>
<tr>
<td>SUM</td>
<td>1074.1</td>
<td>5119.9</td>
<td>6194.0</td>
</tr>
</tbody>
</table>
**Table:** Estimated Totals after Adjustment to ‘sch.wide’ in the 3 Interation

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>542.2</td>
<td>3879.4</td>
<td>4421.5</td>
</tr>
<tr>
<td>H</td>
<td>317.3</td>
<td>437.3</td>
<td>754.6</td>
</tr>
<tr>
<td>M</td>
<td>212.5</td>
<td>805.4</td>
<td>1017.9</td>
</tr>
<tr>
<td>SUM</td>
<td>1072.0</td>
<td>5122.0</td>
<td>6194.0</td>
</tr>
</tbody>
</table>
**Table:** Estimated Totals after Adjustment to 'stype' in the 4 Iteration

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>542.1</td>
<td>3878.9</td>
<td>4421.0</td>
</tr>
<tr>
<td>H</td>
<td>317.5</td>
<td>437.5</td>
<td>755.0</td>
</tr>
<tr>
<td>M</td>
<td>212.5</td>
<td>805.5</td>
<td>1018.0</td>
</tr>
<tr>
<td>SUM</td>
<td>1072.1</td>
<td>5121.9</td>
<td>6194.0</td>
</tr>
</tbody>
</table>
**Table**: Estimated Totals after Adjustment to 'sch.wide' in the 4 Iteration

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>542.0</td>
<td>3879.0</td>
<td>4421.0</td>
</tr>
<tr>
<td>H</td>
<td>317.4</td>
<td>437.5</td>
<td>755.0</td>
</tr>
<tr>
<td>M</td>
<td>212.5</td>
<td>805.5</td>
<td>1018.0</td>
</tr>
<tr>
<td>SUM</td>
<td>1072.0</td>
<td>5122.0</td>
<td>6194.0</td>
</tr>
</tbody>
</table>
Raking with the survey Package

dclus1r <- rake(dclus1, list(~stype, ~sch.wide),
                list(table(stype=apipop$stype),
                     table(sch.wide=apipop$sch.wide)))

svytable(~stype+sch.wide, dclus1r, round=TRUE)

### sch.wide
### stype  No  Yes
###   E  542 3879
###   H  317  438
###   M  213  805

(w1/weights(dclus1r))[1:10]

###  863  1138  1139  1140  1141  1142  1143
### 0.9999724 1.0001319 0.9999724 0.9999724 0.9999724 0.9999724 0.9999724
### 1144  1145  1146
### 0.9999724 0.9999724 0.9999724

summary(w1/weights(dclus1r))

### Min. 1st Qu.  Median  Mean  3rd Qu.  Max.
### 1 1 1 1 1 1
For the linear generalized regression estimator (GREG), the measure of distance $G_k$ is

$$G_k(w, \pi, c) = G(w_k, d_k, c_k) = \frac{(w_k - d_k)^2}{2d_k c_k},$$

and we have

$$\hat{\tau}_{\text{GREG}} = \hat{\tau}_\pi + (\tau_x - \hat{\tau}_x \pi)^\top \hat{\beta},$$

where

$$\hat{\beta} = \left( \sum_{k \in \delta} d_k c_k x_k (x_k)^\top \right)^{-1} \sum_{k \in \delta} d_k c_k x_k y_k,$$

and $\hat{\tau}_x \pi = (\hat{\tau}_{x_1} \pi, \ldots, \hat{\tau}_{x_Q} \pi)^\top$.

The adjustment to the design weight $g_k$ can be written as:

$$g_k = 1 + \left( \left( \sum_{k \in \delta} x_k - \sum_{k \in \delta} d_k x_k \right)^\top \left( \sum_{k \in \delta} d_k c_k x_k (x_k)^\top \right)^{-1} \right) c_k x_k$$
Graphical presentation of $\pi$- and GREG Estimator
We want to estimate total expenditures of hospitals. To improve a possible estimate we use data from survey in 1998 to explore, if there are any useful predictors for our variable of interest.

```r
library(PracTools) #load the package
data(smho.N874) #load the data set
head(smho.N874)
```

```
## EXPTOTAL BEDS SEENCNT EOYCNT FINDIRCT hosp.type
## 1 9066430 81 1791 184 2 1
## 2 9853392 80 1870 244 2 1
## 3 3906074 26 1273 0 2 1
## 4 9853392 90 1781 154 2 1
## 5 9853392 71 1839 206 2 1
## 6 9853392 81 1823 196 2 1
```

```
?smho.N874 #for a description of the variables

#only hospitals other than 'type 4' are considered
```
Generalized Regression Estimator

Fitting a linear model for `EXPTOTAL` with common slopes for `SEENCNT` and `EOYCNT` but a different slope for `BEDS` in each hospital type.

**Table:** Model Summary

|                  | Estimate  | Std. Error  | t value | Pr(>|t|) |
|------------------|-----------|-------------|---------|----------|
| (Intercept)      | 1318589.11| 912432.21   | 1.45    | 0.15     |
| SEENCNT          | 1033.94   | 310.63      | 3.33    | 0.00     |
| EOYCNT           | 2036.15   | 603.58      | 3.37    | 0.00     |
| FINDIRCT2        | 78026.06  | 965237.62   | 0.08    | 0.94     |
| hosp.type1:BEDS  | 98139.28  | 3318.84     | 29.57   | 0.00     |
| hosp.type2:BEDS  | 39489.35  | 5644.51     | 7.00    | 0.00     |
| hosp.type3:BEDS  | 77578.37  | 15082.20    | 5.14    | 0.00     |
| hosp.type5:BEDS  | 36855.78  | 8650.48     | 4.26    | 0.00     |
Generalized Regression Estimator

Residuals vs Fitted

Fitted values

lm(EXPTOTAL ~ SEENCNT + EOYCNT + FINDIRCT + hosp.type:BEDS)
We select a sample of hospitals with probability proportional to the square root of $\text{BEDS}$ using a systematic sample.

```r
### Select a pps to sqrt(BEDS) sample
library(sampling)  #load the 'sample' package
                 #for the 'UPsystematic' function
smho. <-  # before sampling order the data set by hospital type
  smho.[order(smho.$hosp.type),]

x <- smho.[,"BEDS"]
x[x <= 5] <- 5  # recode small hospitals to have a minimum size
x <- sqrt(x)

n <- 80  #sample size
IP <- n*x/sum(x)

set.seed(428274453)
sam <- UPsystematic(IP)

sam.dat <- smho.[sam==1,]
sam.dat$IP <- IP[sam==1]  #the design weight
sam.dat$d <- 1/IP[sam==1]  #the design weight
```
Now we use the `survey` package to calibrate the weights.

```r
library(survey)  # load the 'survey' package

# 1. build a 'design' object
sam.dsgn <-
  svydesign(  
    ids = ~1,  # no clusters
    data = sam.dat  # the sample data
    ,fpc = ~IP  # incl. prob
    ,pps= "brewer")  # variance approx. method

# the model we use for the GREG
lmod2 <- lm(EXPTOTAL ~ SEENCNT + EOYCNT + hosp.type:BEDS, data=smho.)

# 2. compute pop totals of auxiliaries
pop.tots <- colSums(model.matrix(lmod2))  # Inefficient but convenient!

# 3. use 'calibrate' to compute the new weights
sam.cal <-
  calibrate(design = sam.dsgn,
            formula = ~ SEENCNT + EOYCNT + hosp.type:BEDS,
            population = pop.tots,
            calfun='linear' )

Setting argument calfun='linear' in 'calibrate' results in the GREG weights, other calibration function are possible, already built-in are 'raking' and 'logit'.
```
Now we check if the calibration constrains are satisfied:

**#BEDS by hospital type**

```r
svyby(~BEDS, by=~hosp.type, design=sam.cal, FUN=svytotal)
```

```
## hosp.type BEDS  se  
## 1 1 37978 4.100694e-12  
## 2 2 13066 1.122603e-12  
## 3 3 9573 3.028628e-13  
## 5 5 10077 7.135892e-13
```

**#SEENCNT and EOYCNT**

```r
svytotal(~SEENCNT+EOYCNT, sam.cal)
```

```
## SEENCNT  EOYCNT  
## total 1349241 505345 0  
```

**pop.tots**

```
## (Intercept) SEENCNT EOYCNT hosp.type1:BEDS  
## 725 1349241 505345 37978  
## hosp.type2:BEDS hosp.type3:BEDS hosp.type5:BEDS  
## 13066 9573 10077
```
Nothing prevents the GREG weights from becoming negative, which is theoretically not a problem, as long as we infer to the population (or sub-populations) to which we calibrated.
Nothing prevents the GREG weights from becoming negative, which is theoretically not a problem, as long as we infer to the population (or sub-populations) to which we calibrated. However the effects might be catastrophic of domain estimation, in case of estimation domains that where not considered in the calibration.
Nothing prevents the GREG weights from becoming negative, which is theoretically not a problem, as long as we infer to the population (or sub-populations) to which we calibrated. However the effects might be catastrophic of domain estimation, in case of estimation domains that where not considered in the calibration.

In general, it is advisable to only use calibrated weights to infer to the whole population or sub-populations that are found in the marginal totals used for the calibration!
Nothing prevents the GREG weights from becoming negative, which is theoretically not a problem, as long as we infer to the population (or sub-populations) to which we calibrated. However the effects might be catastrophic of domain estimation, in case of estimation domains that were not considered in the calibration.

In general, it is advisable to only use calibrated weights to infer to the whole population or sub-populations that are found in the marginal totals used for the calibration! Design weights can always be used to do unbiased domain estimation, although the precision of these estimates can be very poor.
VARIANCE ESTIMATION FOR COMPLEX STATISTICS
In practice design based variance estimation will *not* be an option for most surveys.
In practice design based variance estimation will \textit{not} be an option for most surveys. Why?
In practice design based variance estimation will *not* be an option for most surveys. Why? We do not know the $\pi_{kl}$ and in some cases not even the $\pi_k$ and even if we would, the response process and frame imperfections remain still unknown.
In practice design based variance estimation will not be an option for most surveys. Why? We do not know the $\pi_{kl}$ and in some cases not even the $\pi_k$ and even if we would, the response process and frame imperfections remain still unknown. Thus we have to resort to second best solutions:

- Model based estimation (as shown for the estimation of the deff)
- Estimation of approximated variances
What exactly has variance estimation to fulfill?
Deliver adequate quality measures
Consider practical issues
What exactly has variance estimation to fulfill?

Deliver adequate quality measures

Consider practical issues

But

Can we apply general methods like for point estimation?
In variance estimation we face often following obstacles:
Non-linear Statistics
In variance estimation we face often following obstacles:

Non-linear Statistics
Complex Designs (incl. non-response)
In variance estimation we face often following obstacles:

- Non-linear Statistics
- Complex Designs (incl. non-response)

Two of the commonly used methods to overcome these are

- **Linearization**: An approximation to the variance of the estimator is sought that is far easier to estimate.
- **Resampling Methods**: The distribution of the estimator is simulated in order to estimate its variance.

The line between model and design based approaches is not so clear in some cases. There are resampling methods and linearization techniques that assume infinite population, i.e. they rely on models. We try to stay with the finite population model.
In variance estimation we face often following obstacles:

- Non-linear Statistics
- Complex Designs (incl. non-response)

Two of the commonly used methods to overcome these are

**Linearization**; An approximation to the variance of the estimator is sought that is far easier to estimate.
In variance estimation we face often following obstacles:

- Non-linear Statistics
- Complex Designs (incl. non-response)

Two of the commonly used methods to overcome these are

**Linearization** ; An approximation to the variance of the estimator is sought that is far easier to estimate.

**Resampling Methods** ; The distribution of the estimator is simulated in order to estimate its variance.
In variance estimation we face often following obstacles:

Non-linear Statistics
Complex Designs (incl. non-response)

Two of the commonly used methods to overcome these are

Linearization ; An approximation to the variance of the estimator is sought that is far easier to estimate.

Resampling Methods ; The distribution of the estimator is simulated in order to estimate its variance.

The line between model and design based approaches is not so clear in some cases. There are resampling methods and linearization techniques that assume infinite population, i.e. they rely on models. We try to stay with the finite population model.
The estimator we use are often non-linear, like $\bar{y}_w$. Which is a problem, in particular for variance estimation. Mind that:

$$
\begin{align*}
E \left( \frac{\hat{\tau}_{y_1 w}}{\hat{\tau}_{y_2 w}} \right) & \approx \frac{E(\hat{\tau}_{y_1 w})}{E(\hat{\tau}_{y_2 w})} \\
V \left( \frac{\hat{\tau}_{y_1 w}}{\hat{\tau}_{y_2 w}} \right) & \neq \frac{V(\hat{\tau}_{y_1 w})}{V(\hat{\tau}_{y_2 w})}
\end{align*}
$$
In case our statistic of interest $\theta$ can be displayed as a function $f$ of $Q$ totals

$$\theta = f (\tau) ,$$

with $\tau = (\tau_{x_1}, \ldots, \tau_{x_q}, \ldots, \tau_{x_Q})^\top$ and $\tau_{x_1} = \sum_{k \in U} x_{kq}$. We estimate $\theta$ by

$$\hat{\theta} = f (\hat{\tau}) ,$$

with $\hat{\tau} = (\hat{\tau}_{x_1}, \ldots, \hat{\tau}_{x_q}, \ldots, \hat{\tau}_{x_Q})^\top$. 
Taylor-linearization

If $f$ is continuously differentiable up to second order between $\tau$ and $\hat{\tau}$ we can use the Taylor series of estimator $\hat{\theta}$ to obtain a linearized version of it

$$\hat{\theta} = \theta + \sum_{q=1}^{Q} \left[ \frac{\partial f(t_1, \ldots, t_Q)}{\partial t_q} \right]_{t=\tau} (\hat{x}_q - \tau_{x_q}) + R(\hat{\tau}, \tau),$$

where

$$R(\hat{\tau}, \tau) = \frac{1}{2} \sum_{q=1}^{Q} \sum_{p=1}^{Q} \left[ \frac{\partial^2 f(t_1, \ldots, t_Q)}{\partial t_q \partial t_p} \right]_{t=\bar{\tau}} (\hat{x}_q - \tau_{x_q})(\hat{x}_p - \tau_{x_p})$$

and $\bar{\tau}$ is between $\hat{\tau}$ and $\tau$. In most application the remainder term $R$ is ignored for large enough sample sizes.
Thus we can approximate the variance of \( \hat{\theta} \) by

\[
V \left( \hat{\theta} \right) \approx V \left( \sum_{q=1}^{Q} \left[ \frac{\partial f(t_1, \ldots, t_Q)}{\partial t_q} \right]_{t=\tau} \hat{\tau}_{x_q} \right)
\]

\[
= \sum_{q=1}^{Q} a_q^2 V \left( \hat{\tau}_{x_q} \right) + 2 \sum_{q=1}^{Q} \sum_{p=1}^{Q} a_q a_p \text{COV} \left( \hat{\tau}_{x_q}, \hat{\tau}_{x_p} \right),
\]

with \( a_q = \left[ \frac{\partial f(t_1, \ldots, t_Q)}{\partial t_q} \right]_{t=\tau} \).
For $\hat{\mathbf{r}} = (\hat{r}_{x_1w}, \ldots, \hat{r}_{x_qw}, \ldots, \hat{r}_{x_Qw})^T$ Woodruff (1971) proposes the transformation of $x_{kq}$

$$z_k = \sum_{q=1}^{Q} a_q x_{kq}$$

and use the following expression as an approximate variance of $\hat{\theta}$

$$V\left(\hat{\theta}\right) \approx V\left(\sum_{k \in \delta} w_k z_k\right).$$

This approximation is far more convenient to estimate then to estimate all the different variances and covariance in the above formula separately. $z_k$ is also sometimes called the linearized variable.
Example Taylor-linearization

Suppose we want to estimate $\theta = \frac{\bar{\gamma}}{N}$ the population mean of variable $\gamma$. To do this we use estimator $\bar{y}_w = \bar{y}_\pi$, i.e. we set $w_k = d_k$ for all $k \in \Delta$. Now we have

$$\hat{\theta} = f(\tau) = f((\hat{\gamma}_{y,\pi}, \hat{\gamma}_{x,\pi})) = \frac{\hat{\gamma}_{y,\pi}}{\hat{\gamma}_{x,\pi}},$$

were $\hat{\gamma}_{y,\pi} = \sum_{k \in \Delta} d_k y_k$ and $\hat{\gamma}_{x,\pi} = \sum_{k \in \Delta} d_k$, because $x_k = 1$ for all $k \in \mathcal{U}$. Our estimator is a function of $Q=2$ totals, and we have

$$a_1 = \frac{1}{\tau_x}, \quad a_2 = -\frac{\tau_y}{\tau_x^2} = -\frac{\theta}{\tau_x},$$

$$z_k = a_1 y_k + a_2 x_k = \frac{1}{\tau_x} (y_k - \theta)$$
To estimate the approximate variance $V(\sum_{k \in \Delta} d_k z_k)$ we need estimates for the $z_k$'s, because they involve unknown statistics $\tau_x$ and $\theta$.

$$\hat{z}_k = \frac{1}{\hat{\tau}_{x,\pi}} (y_k - \bar{y}_\pi)$$

Our variance estimator would be $\hat{V}(\sum_{k \in \Delta} d_k \hat{z}_k)$ for which $\hat{V}(.)_1$ or $\hat{V}(.)_2$ could be used or an estimator for an approximate design variance. Note that the variances of the $\hat{z}_k$'s (and the covariances between them) is often thought to be negligible and is therefore usually not considered.
Calibrated weights are not independent of the selected sample, i.e. they are random variables. Thus, the variance of the GREG estimator cannot be estimated as straightforwardly as for the $\pi$-estimator. We can write its approximate variance as

$$\text{AV} \left( \hat{\tau}_{\text{GREG}} \right) = \sum_{k \in U} \sum_{l \in U} \left( \pi_{kl} - \pi_k \pi_l \right) \frac{E_k}{\pi_k} \frac{E_l}{\pi_l},$$

where $E_k = y_k - y_k^0$, $y_k^0 = x_k^T \beta$ and

$$\beta = \left( \sum_{k \in U} c_k x_k(x_k)\T \right)^{-1} \sum_{k \in U} c_k x_k y_k.$$
Calibrated weights are not independent of the selected sample, i.e. they are random variables. Thus, the variance of the GREG estimator cannot be estimated as straightforwardly as for the $\pi$-estimator. A variance estimator for $\hat{T}_{\text{GREG}}$ is given by

$$
\hat{V}(\hat{T}_{\text{GREG}}) = \sum_{k \in \mathcal{S}} \sum_{l \in \mathcal{S}} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} g_k \frac{e_k}{\pi_k} g_l \frac{e_l}{\pi_l},
$$

where $e_k = y_k - \hat{y}_k$ and $\hat{y}_k = \mathbf{x}_k^T \hat{\beta}$. 
Calibrated weights are not independent of the selected sample, i.e. they are random variables. Thus, the variance of the GREG estimator cannot be estimated as straightforwardly as for the $\pi$-estimator. Compare the estimates for the total expenditure with calibrated and design weights.

\begin{verbatim}
svytotal(~EXPTOTAL, sam.dsgn)
## total SE
## EXPTOTAL 9.55e+09 7.72e+08

svytotal(~EXPTOTAL, sam.cal)
## total SE
## EXPTOTAL 9.03e+09 5.92e+08
\end{verbatim}

We find that $\hat{\tau}_{GREG}/\hat{\tau}_\pi = 0.9451$, but $\hat{V}(\hat{\tau}_{GREG})/\hat{V}(\hat{\tau}_\pi) = 0.588$. 
R.J.A. Little, D.B. Rubin.
Statistical Analysis with Missing Data.

C.-E. Särndal, S. Lundström.
Estimation in Surveys with Nonresponse.

R.S. Woodruff.
A Simple Method for Approximating the Variance of a Complicated Estimate.