4. GESIS Summer School in Survey Methodology Cologne. Course 16: Sampling, Weighting and Estimation

Day 1: Sampling Designs

Stefan Zins[1] and Matthias Sand[2]

Kingdom, come, 1

[1]Stefan.Zins@gesis.org
[2]Matthias.Sand@gesis.org

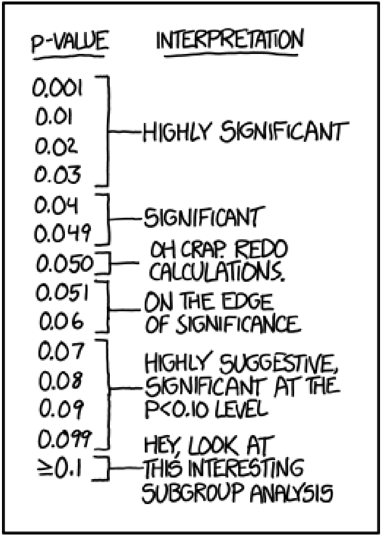
- Understanding the basic principles of design-based inference,

- Understanding the basic principles of design-based inference,
- apply them to estimated form complex survey samples (or the planning of surveys), and

Objectives of the Course

- Understanding the basic principles of design-based inference,
- apply them to estimated form complex survey samples (or the planning of surveys), and
- learn how to do this with R!

- What do you want to do?



- Source: http://xkcd.com/1478/

- 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?

$\mathcal{Y} = \{y_1, y_2, \ldots, y_k, \ldots, y_N\}$	finite population of size N
$\mathcal{U} = \{1, 2, \ldots, k, \ldots, N\}$	sampling frame
$\mathscr{s} \subset \mathcal{U}$	sample of size n
$\mathcal{P}(\mathcal{U})$	all possible subsets of $\mathcal{U}$

The discrete probability distribution $p(.)$ over $\mathcal{P}(\mathcal{U})$ is called a *sampling design* and $\mathcal{G} = \{\mathscr{s} | \mathscr{s} \in \mathcal{P}(\mathcal{U}),\ p(\mathscr{s}) > 0\}$ is called the support of $p(.)$ with

$$\sum_{\mathscr{s} \in \mathcal{G}} p(\mathscr{s}) = 1$$

Hence, $p: \ \mathcal{G} \mapsto (0, 1]$.

$$\theta = f(\mathcal{Y}) \qquad \text{statistic of interest}$$

$$\hat{\theta} = f(\mathcal{Y}, \scriptstyle\mathcal{S}) \qquad \text{estimator for } \theta$$

$$\mathrm{E}\left(\hat{\theta}\right) = \sum_{\mathcal{S} \in \mathcal{G}} p(\mathcal{S}) f(\mathcal{Y}, \mathcal{S}) \qquad \text{expected value of } \hat{\theta}$$

$$\mathrm{V}\left(\hat{\theta}\right) = \mathrm{E}\left(\hat{\theta}^2\right) - \mathrm{E}\left(\hat{\theta}\right)^2 \qquad \text{variance of } \hat{\theta}$$

$$\mathrm{MSE}\left(\hat{\theta}\right) = \mathrm{E}\left((\hat{\theta} - \theta)^2\right) = \left(\mathrm{E}\left(\hat{\theta}\right) - \theta\right)^2 + \mathrm{V}\left(\hat{\theta}\right) \qquad \text{mean square error of } \hat{\theta}$$

$\mathrm{E}(.)$, $\mathrm{V}(.)$, and $\mathrm{MSE}(.)$ are always with respect to the sampling design $p()$ and an estimator is said to be unbiased if

$$\mathrm{E}\left(\hat{\theta}\right) = \theta \, .$$

What is a representative sample?

What is a representative sample?
The popular concept of a representative sample it that the sample is a *miniature* of the population.

However, what do we really want?

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.

A *strategy* is the combination of a sampling design and an estimator.
For the statistic of interest the aim is to find the best possible strategy, that is, one that estimates the statistic as accurately as possible.
A measure of accuracy can be the mean square error.

Expectation and Variance of a Random Sample

S_k	number of times element k is selected
$I_k = \begin{cases} 1 & \text{if } k \in \mathcal{s} \\ 0 & \text{else} \end{cases}$	sampling indicator element k
$\mathrm{E}(S_k) = \nu_k$	expected selection frequency of element k
$\mathrm{E}(S_k S_l) = \nu_{kl}$	joint expectation of S_k and S_l
$\mathrm{E}(I_k) = \pi_k$	inclusion probability of element k
$\mathrm{E}(I_k I_l) = \pi_{kl}$	joint expectation of I_k and I_l
$\sum_{k \in \mathcal{U}} \nu_k = \mathrm{E}(n)$	expected sample size

Simple random sampling without replacement (SRS): Drawing n elements out of a urn without putting them back (i.e. $S_k = I_k$) and without remembering the order of the selected element.

$$\mathcal{G} = \binom{N}{n} \tag{1}$$

$$p(s) = \binom{N}{n}^{-1} \tag{2}$$

$$\pi_k = \nu_k = \frac{n}{N} \tag{3}$$

$$\pi_{kl} = \nu_{kl} = \frac{n(n-1)}{N(N-1)} \text{ for } k \neq l \tag{4}$$

Leibniz Institute for the Social Sciences

$$\theta = \mu = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k, \quad \hat{\theta} = \overline{y} = \sum_{k \in \mathcal{s}} \frac{y_k}{n}, \quad \sigma^2 = \frac{1}{N} \sum_{k \in \mathcal{U}} (y_k - \mu)^2, \quad V^2 = \sigma^2 \frac{N}{N-1}$$

gesis Leibniz Institute for the Social Sciences

$$\theta = \mu = \frac{1}{N}\sum_{k\in\mathcal{U}} y_k, \quad \hat{\theta} = \overline{y} = \sum_{k\in\mathcal{s}} \frac{y_k}{n}, \quad \sigma^2 = \frac{1}{N}\sum_{k\in\mathcal{U}} (y_k - \mu)^2, \quad V^2 = \sigma^2 \frac{N}{N-1}$$

Expected value

$$\begin{aligned}
\mathrm{E}\left(\overline{y}\right) &= \mathrm{E}\left(\sum_{k\in\mathcal{U}} S_k \frac{y_k}{n}\right) \\
&= \frac{1}{n}\sum_{k\in\mathcal{U}} \mathrm{E}\left(S_k\right) y_k \\
&= \frac{1}{n}\sum_{k\in\mathcal{U}} \pi_k y_k \\
&= \frac{1}{N}\sum_{k\in\mathcal{U}} y_k
\end{aligned}$$

gesis Leibniz Institute for the Social Sciences

$$\theta = \mu = \frac{1}{N}\sum_{k\in\mathcal{U}} y_k, \quad \hat{\theta} = \overline{y} = \sum_{k\in\mathcal{s}} \frac{y_k}{n}, \quad \sigma^2 = \frac{1}{N}\sum_{k\in\mathcal{U}} (y_k-\mu)^2, \quad V^2 = \sigma^2\frac{N}{N-1}$$

Expected value

$$\begin{aligned} \mathrm{E}\left(\overline{y}\right) &= \mathrm{E}\left(\sum_{k\in\mathcal{U}} S_k \frac{y_k}{n}\right) \\ &= \frac{1}{n}\sum_{k\in\mathcal{U}} \mathrm{E}\left(S_k\right) y_k \\ &= \frac{1}{n}\sum_{k\in\mathcal{U}} \pi_k y_k \\ &= \frac{1}{N}\sum_{k\in\mathcal{U}} y_k \end{aligned}$$

Variance

$$\begin{aligned} \mathrm{V}\left(\overline{y}\right) &= \mathrm{V}\left(\sum_{k\in\mathcal{U}} S_k \frac{y_k}{n}\right) \\ &= \frac{1}{n^2}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}} \mathrm{COV}\left(S_k,\, S_l\right) y_k y_l \\ &= -\frac{1}{2}\frac{1}{n^2}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}} \left(\pi_{kl} - \pi_k\pi_l\right)\left(y_k - y_l\right)^2 \\ &= \frac{N-n}{N-1}\frac{\sigma^2}{n} = \left(1-\frac{n}{N}\right)\frac{V^2}{n} \end{aligned}$$

Simple random sampling with replacement (SRSWR): Drawing n elements out of a urn by making n successive draws and putting after each draw the element back (i.e. $S_k \neq I_k$). The order of the selected elements is also not remembered.

$$\mathcal{G} = \binom{N+n-1}{n} \quad p(s) = \binom{N+n-1}{n}^{-1}$$

$$\nu_k = \frac{n}{N} \qquad \pi_k = 1 - \left(\frac{N-1}{N}\right)^n$$

$$\nu_{kl} = \frac{n(n-1)}{N^2} \qquad \pi_{kl} = 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n \qquad \text{for } k \neq l$$

Leibniz Institute for the Social Sciences

Expected value

$$
\begin{aligned}
\mathrm{E}\left(\bar{y}\right) &= \mathrm{E}\left(\sum_{k\in\mathcal{U}} S_k \frac{y_k}{n}\right) \\
&= \frac{1}{n}\sum_{k\in\mathcal{U}} \mathrm{E}\left(S_k\right) y_k \\
&= \frac{1}{n}\sum_{k\in\mathcal{U}} \nu_k y_k \\
&= \frac{1}{N}\sum_{k\in\mathcal{U}} y_k
\end{aligned}
$$

gesis
Leibniz Institute for the Social Sciences

Expected value

$$\begin{aligned}
\mathrm{E}\left(\overline{y}\right) &= \mathrm{E}\left(\sum_{k\in\mathcal{U}} S_k \frac{y_k}{n}\right) \\
&= \frac{1}{n}\sum_{k\in\mathcal{U}} \mathrm{E}\left(S_k\right) y_k \\
&= \frac{1}{n}\sum_{k\in\mathcal{U}} \nu_k y_k \\
&= \frac{1}{N}\sum_{k\in\mathcal{U}} y_k
\end{aligned}$$

Variance

$$\begin{aligned}
\mathrm{V}\left(\overline{y}\right) &= \mathrm{V}\left(\sum_{k\in\mathcal{U}} S_k \frac{y_k}{n}\right) \\
&= \frac{1}{n^2}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}} \mathrm{COV}\left(S_k, S_l\right) y_k y_l \\
&= -\frac{1}{2}\frac{1}{n^2}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}} \left(\nu_{kl} - \nu_k\nu_l\right)\left(y_k - y_l\right)^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

gesis Leibniz Institute for the Social Sciences

Expected value

$$
\begin{aligned}
\mathrm{E}(\bar{y}) &= \mathrm{E}\left(\sum_{k\in\mathcal{U}} S_k \frac{y_k}{n}\right) \\
&= \frac{1}{n}\sum_{k\in\mathcal{U}} \mathrm{E}(S_k)\, y_k \\
&= \frac{1}{n}\sum_{k\in\mathcal{U}} \nu_k y_k \\
&= \frac{1}{N}\sum_{k\in\mathcal{U}} y_k
\end{aligned}
$$

Variance

$$
\begin{aligned}
\mathrm{V}(\bar{y}) &= \mathrm{V}\left(\sum_{k\in\mathcal{U}} S_k \frac{y_k}{n}\right) \\
&= \frac{1}{n^2}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}} \mathrm{COV}(S_k, S_l)\, y_k y_l \\
&= -\frac{1}{2}\frac{1}{n^2}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}} (\nu_{kl} - \nu_k\nu_l)(y_k - y_l)^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

Note: $\frac{\sigma^2}{n} > \left(1 - \frac{n}{N}\right)\frac{V^2}{n}$, if $n > 1$.

Variance Estimation SRS/WR

$$\hat{\mathrm{V}}\left(\overline{y}\right)_{\mathrm{SRS}} = \frac{N-n}{N}\frac{s^2}{n}$$

$$\hat{\mathrm{V}}\left(\overline{y}\right)_{\mathrm{SRSWR}} = \frac{s^2}{n}$$

Sample variance

$$s^2 = \frac{1}{n-1}\sum_{k\in s}(y_k - \overline{y})^2 = \frac{1}{n(n-1)}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}}(y_k - y_l)^2 S_k S_l$$

$$\mathrm{E}\left(s^2\right)_{\mathrm{SRS}} = \frac{N}{N-1}\sigma^2 = V^2$$

$$\mathrm{E}\left(s^2\right)_{\mathrm{SRSWR}} = \sigma^2$$

The sample data: $\mathcal{y} = \{y_1, \ldots, y_k, \ldots, y_n\}$. All $y_k \in \mathcal{y}$ are independent identical distributed (iid) random variables, with

$$y_k \sim NV(\mu, \sigma) \; .$$

The sample data: $\mathcal{y} = \{y_1, \ldots, y_k, \ldots, y_n\}$. All $y_k \in \mathcal{y}$ are independent identical distributed (iid) random variables, with

$$y_k \sim NV(\mu, \sigma) \; .$$

Expected value

$$\begin{aligned} \mathrm{E}\left(\overline{y}\right)_M &= \mathrm{E}\left(\sum_{k \in \mathcal{s}} \frac{y_k}{n}\right) \\ &= \frac{1}{n} \sum_{k \in \mathcal{s}} \mu \\ &= \mu \end{aligned}$$

The sample data: $\mathcal{y} = \{y_1, \ldots, y_k, \ldots, y_n\}$. All $y_k \in \mathcal{y}$ are independent identical distributed (iid) random variables, with

$$y_k \sim NV(\mu, \sigma) \; .$$

Expected value

$$\begin{aligned} \mathrm{E}\left(\bar{y}\right)_M &= \mathrm{E}\left(\sum_{k \in \mathcal{s}} \frac{y_k}{n}\right) \\ &= \frac{1}{n} \sum_{k \in \mathcal{s}} \mu \\ &= \mu \end{aligned}$$

Variance

$$\begin{aligned} \mathrm{V}\left(\bar{y}\right)_M &= \mathrm{V}\left(\sum_{k \in \mathcal{s}} \frac{y_k}{n}\right)_M \\ &= \frac{1}{n^2} \sum_{k \in \mathcal{s}} \sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

The sample data: $\mathcal{y} = \{y_1, \ldots, y_k, \ldots, y_n\}$. All $y_k \in \mathcal{y}$ are independent identical distributed (iid) random variables, with

$$y_k \sim NV(\mu, \sigma) .$$

Expected value

$$\begin{aligned} \mathrm{E}\left(\bar{y}\right)_M &= \mathrm{E}\left(\sum_{k \in s} \frac{y_k}{n}\right) \\ &= \frac{1}{n} \sum_{k \in s} \mu \\ &= \mu \end{aligned}$$

Variance

$$\begin{aligned} \mathrm{V}\left(\bar{y}\right)_M &= \mathrm{V}\left(\sum_{k \in s} \frac{y_k}{n}\right)_M \\ &= \frac{1}{n^2} \sum_{k \in s} \sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Note that the variance of $\bar{y}$ is the same under the model based approach and SRSWR, i.e. there is no finite population correction.

- 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 s_h is selected from $\mathcal{U}_h$ according to a design $p_h(.)$, 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 s_h is denoted with n_h.

In stratified random sampling the sub-populations are called strata. For the *h*-ht stratum we get:

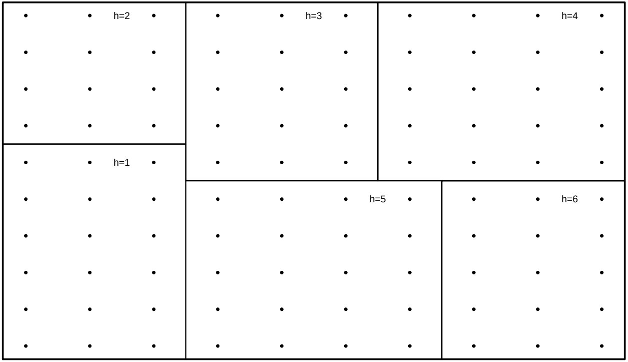
$$\mu_h = \frac{1}{N_h} \sum_{k=1}^{N_h} y_{kh}$$ mean of stratum h

$$\sigma_h^2 = \frac{1}{N_h} \sum_{k=1}^{N_h} (y_{kh} - \mu_h)^2$$ variance of stratum *h*

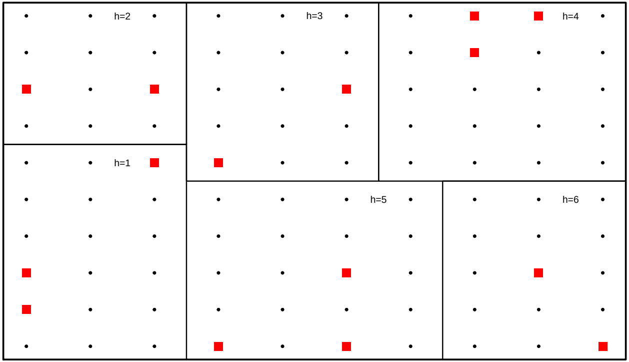
$$V_h^2 = \sigma_h^2 \frac{N_h}{N_h - 1}$$

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.



A Population of 100 elements is stratified into $H = 6$ strata.
14 elements are selected 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 $\gamma_h = \dfrac{N_h}{N}$ and $\mathrm{E}\left(\bar{y}_{\text{str}}\right) = \mu$ for SRS and SRSWR within each stratum.
Variance and variance estimator:

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

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

$$s_h^2 = \frac{1}{n_h - 1} \sum_{k \in s_h} \left(y_k - \bar{y}_h\right)^2$$

- Why should stratification be used?

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).

- 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?

- 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.
  - (You might check out the `SamplingStrata` package, it implements a method to determine of the best stratification of a sampling frame, of minimal cost under the condition to satisfy precision constraints in a multivariate and multi-domain case.)

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.
  - (You might check out the `SamplingStrata` package, it implements a method to determine of the best stratification of a sampling frame, of minimal cost under the condition to satisfy precision constraints in a multivariate and multi-domain case.)
- 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.
    - (You might check out the `SamplingStrata` package, it implements a method to determine of the best stratification of a sampling frame, of minimal cost under the condition to satisfy precision constraints in a multivariate and multi-domain case.)
- 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

Table: Population ANOVA

Source	df	Sum of Squares
Between strata	$H-1$	$\text{SSB} = \sum_{h=1}^{H} N_h(\mu_h - \mu)^2$
Within strata	$N-H$	$\text{SSW} = \sum_{h=1}^{H}(N_h - 1)V_h^2$
Total, about μ_y	$N-1$	$\text{SSTO} = (N-1)V^2$

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.

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

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

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

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

$$
\begin{aligned}
\mathrm{V}\left(\bar{y}_{\text{str}}\right)_{\text{StrSRS}} &= \left(\frac{N-n}{N}\right) \frac{1}{n} \sum_{h=1}^{H} N_h V_h^2 \qquad \text{and} \\
\mathrm{V}\left(\bar{y}\right)_{\text{SRS}} &= \left(\frac{N-n}{N}\right) \frac{1}{n(N-1)} \left(\mathrm{SSW} + \mathrm{SSB}\right) \\
&= \mathrm{V}\left(\bar{y}_{\text{str}}\right)_{\text{StrSRS}} + \left(\frac{N-n}{N}\right) \frac{1}{n(N-1)} \left[\mathrm{SSB} - \sum_{h=1}^{H} \frac{N-N_h}{N} V_h^2\right] .
\end{aligned}
$$

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

$$
SSB > \sum_{h=1}^{H} \frac{N-N_h}{N} V_h^2 .
$$

- It is not assured that $\gamma_h n$ is an integer. If $n_h^* = [n_h]$ is used instead, the allocation is no longer strictly proportional. Furthermore $\sum_{h=1}^{H} n_h^* = n$ is also not assured.
- However stochastic techniques can be used that ensure that $\mathrm{E}\left(n_h^*\right) = n_h$ and thus $\mathrm{E}\left(\sum_{h=1}^{H} n_h^*\right) = n$.

$$n_h^* = \begin{cases} \lfloor n_h \rfloor & \text{with prob. } 1 - (n_h \mod 1) \\ \lceil n_h \rceil & \text{with prob. } (n_h \mod 1) \end{cases}$$

There are stochastic rounding procedures that are controlled and unbiased, i.e. $\sum_{h=1}^{H} n_h = n$ and $\mathrm{E}\left(n_h^*\right) = \frac{N_h}{N} n$ and $|n_h^* - n_h| \leqslant 1$ [Cox, 1987].

The elements of a population of size $N = nH$ are ordered in a specific way (every unit having a unique rank). Starting form a random number k, with $1 \leqslant k \leqslant H$ and $k \in \{1,2, \ldots, H\}$ the sample is defined as the elements with ranks

$$k, k + H, k + 2H, k + 3H, \ldots, k + (n - 1)H .$$

$$\bar{y}_k = \frac{1}{n} \sum_{i=0}^{n-1} y_{(k+iH)} \qquad \text{sample mean}$$

$$\mathrm{E}\left(\bar{y}_k\right)_{\mathrm{SyS}} = \mu \qquad \text{expected value of } \bar{y}_k$$

$$\mathrm{V}\left(\bar{y}_k\right)_{\mathrm{SyS}} = \frac{1}{H} \sum_{h=1}^{H} \left(\bar{y}_k - \mu\right)^2 \qquad \text{variance of } \bar{y}_k$$

There is no unbiased variance estimator for systematic sampling. Ordering the population with respect to certain variables has a similar effects as stratification by the same variable with proportional allocation.

The sample size can be set to achieve a desired level of precision in terms of the variance $\mathrm{V}\left(\hat{\theta}\right)$ or the variation coefficient $\mathrm{CV}(\hat{\theta}) = \frac{\sqrt{\mathrm{V}_0(\hat{\theta})}}{\hat{\theta}}$.
Set $\mathrm{CV}(\overline{y}) = \mathrm{CV}_0$ as a precision requirement (representative!).

$$n = \frac{V^2\mu^{-2}}{\mathrm{CV}_0^2 + V^2 N^{-1}\mu^{-2}} \qquad \text{SRS}$$

$$n = \frac{\sigma^2\mu^{-2}}{\mathrm{CV}_0^2} \qquad \text{SRSWR}$$

If the variable of interest is binary we have $\mathrm{V}\left(\overline{y}\right)_{\mathrm{SRS}} = \frac{\mu(1-\mu)}{n}\frac{N-n}{N-1}$

and $\mathrm{CV}^2(\overline{y})_{\mathrm{SRS}} = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{N}{N-1}\frac{(1-\mu)}{\mu}$ However

$$\lim_{\mu \to 0} \mathrm{CV}^2(\overline{y})_{\mathrm{SRS}} = \infty \ ,$$

thus for rare observation to meet a CV target the sample size can become very large.

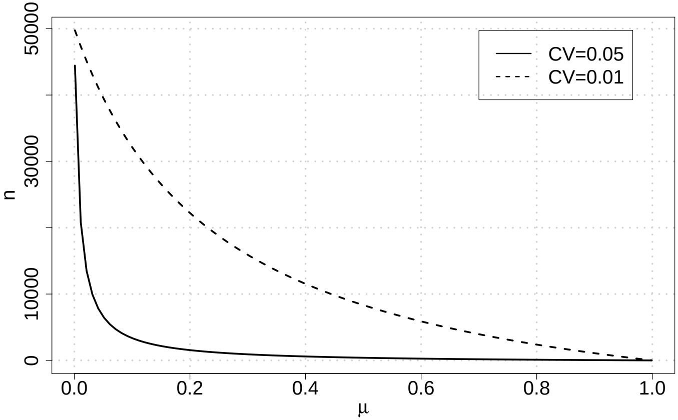


Figure: Sample Sizes to Achieve CV's of 0.05 and 0.01 for N=50000

If the variable of interest is binary we have $\mathrm{V}\left(\overline{y}\right)_{\mathrm{SRS}} = \dfrac{\mu(1-\mu)}{n}\dfrac{N-n}{N-1}$ and $\mathrm{CV}^2(\overline{y})_{\mathrm{SRS}} = \left(\dfrac{1}{n} - \dfrac{1}{N}\right)\dfrac{N}{N-1}\dfrac{(1-\mu)}{\mu}$ However

$$\lim_{\mu \to 0} \mathrm{CV}^2(\overline{y})_{\mathrm{SRS}} = \infty \ ,$$

thus for rare observation to meet a CV target the sample size can become very large. Target values for $\mathrm{V}\left(\overline{y}\right)_{\mathrm{SRS}}$ can instead be set to achieve a CI's with a maximal length of 2ϵ and we select the sample size the following way

$$\begin{aligned} \epsilon \geqslant & z_{1-\alpha/2}\sqrt{\frac{\mu(1-\mu)}{n}\frac{N-n}{N-1}} \\ n \geqslant & \frac{z^2_{1-\alpha/2}\frac{N}{N-1}\mu(1-\mu)}{\epsilon^2 + \frac{1}{N}z^2_{1-\alpha/2}\frac{N}{N-1}\mu(1-\mu)} \end{aligned}$$

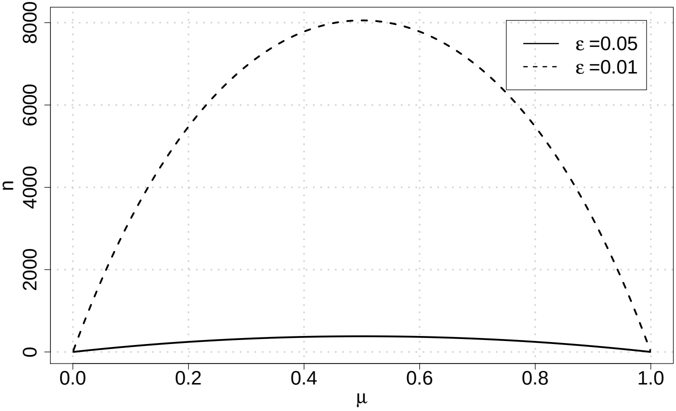


Figure: Sample Sizes to Achieve Absolute Errors of 0.05 and 0.01 for N=50000

Our one-sided t-test: $H_0 : \mu \geqslant \mu_0$ against $H_1 : \mu < \mu_0$. We reject the Null if $t < t_{1-\alpha}(df)$. If the true mean is $\mu = \mu_0 + \delta$ then the probability of *not* rejecting H_0 is

$$P(t \geqslant t_{1-\alpha}(df) | \mu = \mu_0 + \delta) \approx 1 - \Phi \left(z_{1-\alpha} - \delta \left[\mathrm{V} \left(\hat{\theta} \right) \right]^{-\frac{1}{2}} \right)$$

where Φ is the distribution function of the standard normal distribution.

Suppose $\mathcal{Y}$ is binary and our strategy to estimate μ us $\overline{y}$ in combination with SRS, thus $V\left(\hat{\theta}\right) = (\frac{1}{n} - \frac{1}{N})\frac{N}{N-1}\mu(1-\mu)$.

Suppose $\mathcal{Y}$ is binary and our strategy to estimate μ us $\overline{y}$ in combination with SRS, thus $\mathrm{V}\left(\hat{\theta}\right) = (\frac{1}{n} - \frac{1}{N})\frac{N}{N-1}\mu(1-\mu)$.
We want our test to have a power of $1-\beta$, i.e. our planned probability of an error of type II is β.

Suppose $\mathcal{Y}$ is binary and our strategy to estimate μ us $\overline{y}$ in combination with SRS, thus $\mathrm{V}\left(\hat{\theta}\right) = (\frac{1}{n} - \frac{1}{N})\frac{N}{N-1}\mu(1-\mu)$.
We want our test to have a power of $1-\beta$, i.e. our planned probability of an error of type II is β.To select the appropriate sample size we set $z_{1-\alpha} - \delta\left[(\frac{1}{n} - \frac{1}{N})\frac{N}{N-1}\mu(1-\mu)\right]^{-\frac{1}{2}} = z_{\beta}$. Solving this for n give us

$$n = \frac{\frac{N}{N-1}\mu(1-\mu)}{N\left(\frac{\delta}{z_{1-\alpha}-z_{\beta}}\right)^2 + \frac{\mu(1-\mu)}{N-1}} .$$

($z_{1-\alpha}$ and z_{β} are the $1-\alpha$ and β quantiles of the standard normal, respectively.)

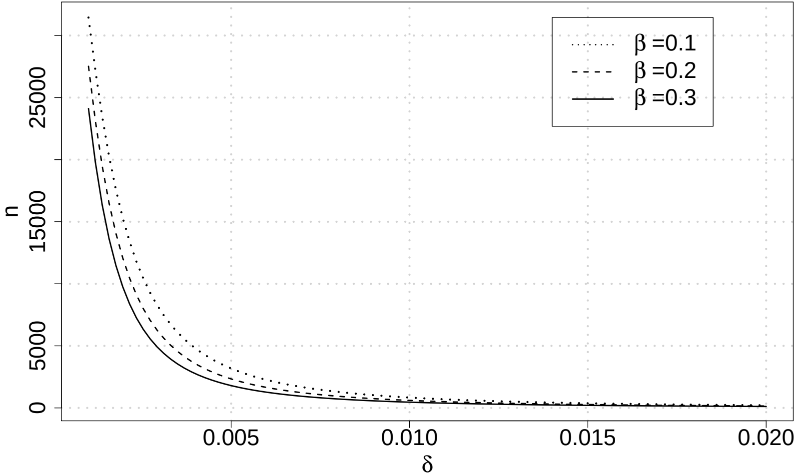


FIGURE: Sample Sizes for Type II Errors 0.1, 0.2, and 0.3 for N=50000

A multidimensional problem sample size problem:

TABLE: Stratfication Table

	P1	P2	SUM
R1	100	1000	1100
R2	200	1000	1200
R3	300	2000	2300
R4.1	50	3000	3050
R4.2	100	3000	3100
R4.3	250	5000	5250
SUM	1000	15000	16000

The domains of interest are rows R1, R2, R3, R4.1, R4.2, R4 and columns P1 and P2. The variable of interest is binary with $\mu = 0.5$. Minimize $n = \sum_{h=1}^{H} n_h$, subject to the following constraints

- $0.01 N_h \leqslant n_h \leqslant N_h$ (size constraint)
- $0.04 \geqslant 1.96 \sum_{h \in \mathcal{D}_i} \dfrac{N_h - n_h}{N_h} \gamma_h^2 \dfrac{V_h^2}{n_h}$ (precision requirement)

where $\mathcal{D}_i$ is the set of strata that constitute the i-th domain of interest [Gabler and Quatember, 2013].

```
fct  <- function(x){sum(1/x)}
#constraints
hin0 <- function(x){c(   x - 1/(Nh+0.01)        #upper size
                       , 1/(0.01*Nh+0.01) - x   #lower size
                       , c-A%*%x)}              #precision

ans <- constrOptim.nl(
  par = 1/(0.42*Nh),        # starting value
  fn = fct,                 # objective function
  hin = hin0,
  control.outer = list(eps     = 1.e-09,
                       mu0     = 1e-01,
                       method = "BFGS",
                       trace   = FALSE

 ))
```

Solution of the Allocation Problem

Table: Allocation Table

	P1	P2	SUM
R1	47	345	392
R2	91	317	408
R3	112	388	500
R4.1	17	489	506
R4.2	35	478	513
R4.3	80	132	212
SUM	382	2149	2531

L. Cox.
A Constructive Procedure for Unbiased Controlled Rounding.
Journal of the American Statistical Association, 1987.

L. Gabler, A. Quatember.
Repräsentativität von Subgruppen bei geschichteten Zufalssstichproben.
Wirtschafts- und Sozialstatistisches Archiv, 2013.

S. Lohr.
Sampling: Design and Analysis.
Duxbury Press, 1999.

T. Lumley.
Complex Surveys: A Guide to Analysis Using R.
Wiley, 2010.

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

R. Valliant, J.A. Dever, & F. Kreuter.
Practical Tools for Designing and Weighting Survey Samples.
Statistics for Social and Behavioral Sciences: Springer, 2013.