gesis Leibniz Institute for the Social Sciences

Sample Theory

Epidemiological Study Design and Statistical Methods

Stefan Zins - GESIS

12.12.2017

Mitglied der Leibniz Gemeinschaft

Content

- Design Based Inference
- Sampling Designs
- Sample Size Planning

SamplignAndEstimation on GitHub, branch sampling_short:
`https://github.com/BernStZi/SamplingAndEstimation/tree/sampling_short`

Motivation

- Did you ever work with sample data?

Motivation

- What kind of analysis have you done with sample data?

gesis Leibniz Institute for the Social Sciences

Motivation

- What concerns did you have while applying your analytically methods?

Section 1

Design Based Inference

Finite Population, Sample, and Sampling Design

$$\mathcal{Y} = \{y_1, y_2, \ldots, y_k, \ldots, y_N\}$$ finite population of size N

$$\mathcal{U} = \{1, 2, \ldots, k, \ldots, N\}$$ sampling frame

$$s \subset \mathcal{U}$$ 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 $\mathcal{G} = \{s | s \in \mathcal{P}(\mathcal{U}), p(s) > 0\}$ is called the support of $p(.)$ with

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

Estimation

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

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

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

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

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

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

Law of Large Numbers (LLN)

Weak law of large numbers:
Suppose $\{y_1, y_2, \ldots, y_k, \ldots, y_N\}$ is a sequence of i.i.d. random variables with mean μ and $\mu \neq \infty$ and $\mu \neq -\infty$. Then for $n \to \infty$ then we have:

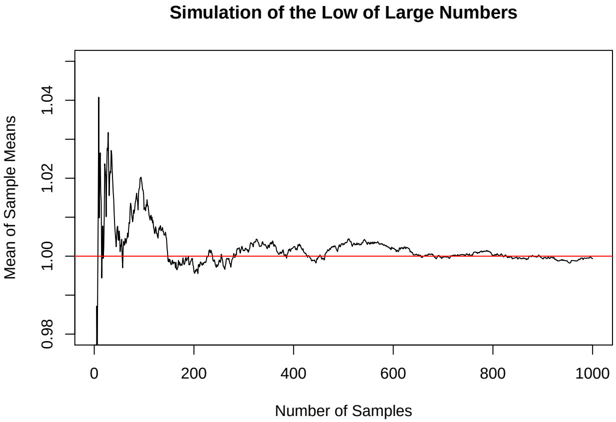
$$\bar{y} \xrightarrow{P} \mu$$

If the LLN holds we can have unbiased estimates, as our estimates will converge in probability to their expected (true) value. That is, it assures $\mathrm{E}(I_k) = \pi_k$.

LLN Demonstration

Suppose all y_i follow an exponential distribution with mean and variance equal to one. We take repeatedly a sample of size 50. For each sample the sample mean is calculated. The mean of the sample means should converge towards the true mean of the distribution with increasing number of samples.

LLN Demonstration



Central Limit Theorem (CLT)

CLT of *Lindeberg–Lévy*:
Suppose $\{y_1, y_2, \ldots, y_k, \ldots, y_N\}$ is a sequence of i.i.d. random variables with $V(y_i) < \infty \; \forall \; i = 1, \ldots, N$. Then for $n \to \infty$ then we have:

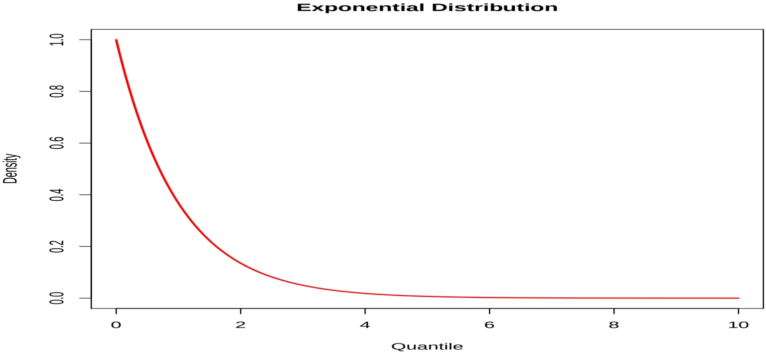
$$\frac{\bar{y} - \mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1)$$

If the CLT holds, symmetric confidence intervals can be constructed with quantiles from the standard normal distribution $\Phi(z)$

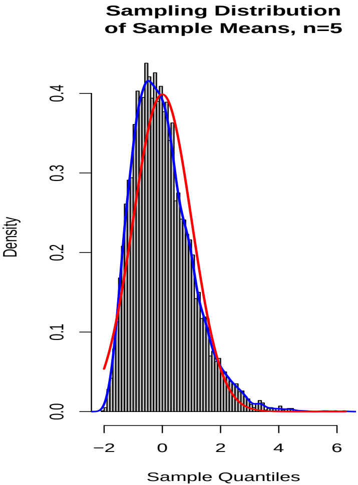
$$\left[\bar{y} + \Phi(\alpha/2)\sigma\sqrt{n}\,;\; \bar{y} + \Phi(1 - \alpha/2)\sigma\sqrt{n}\right]$$

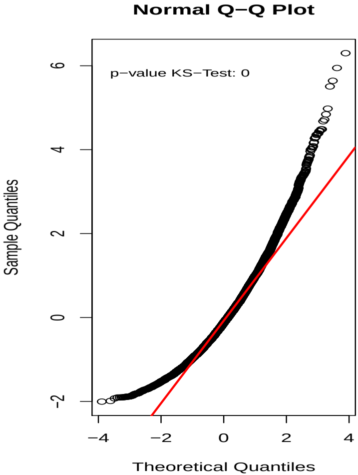
CLT Demonstration

Suppose all y_i follow an exponential distribution with mean and variance equal to one.



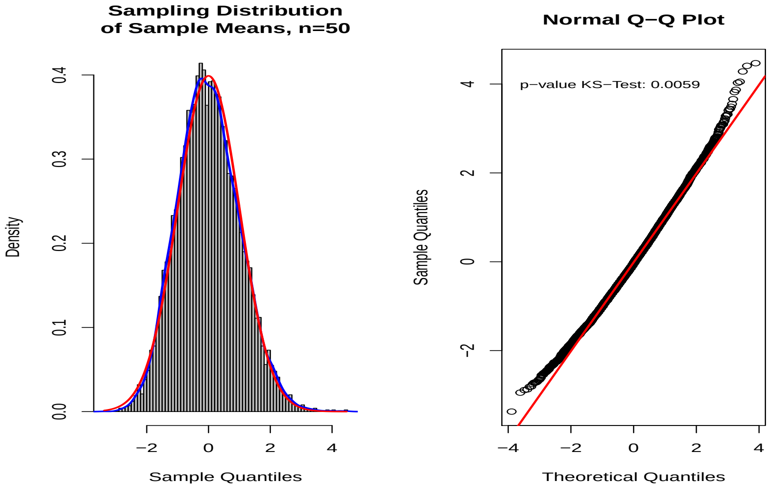
CLT Demonstration



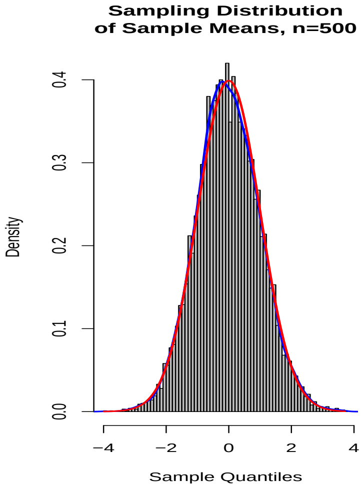


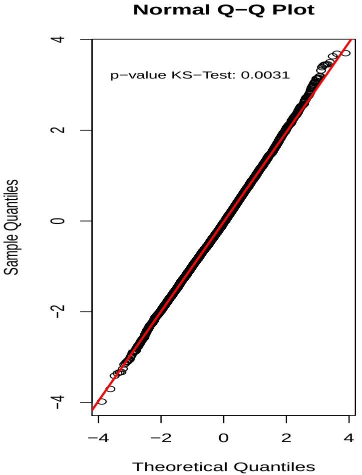
gesis Leibniz Institute for the Social Sciences

CLT Demonstration

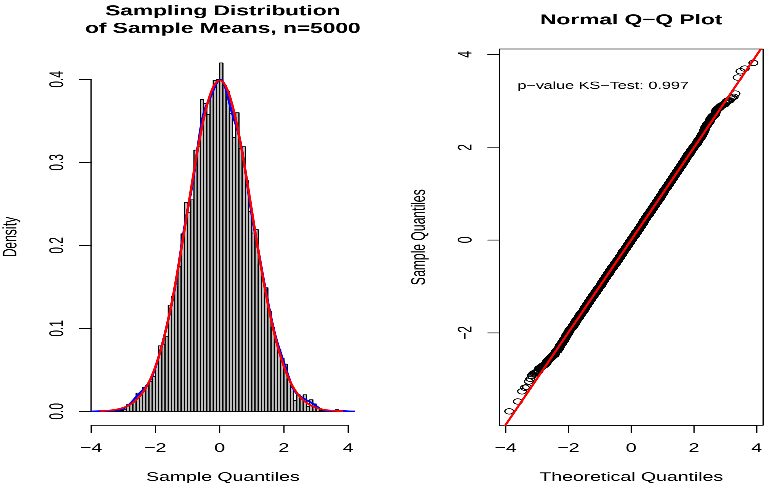


CLT Demonstration





CLT Demonstration



Inclusion Probabilities

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

$$\mathrm{E}(I_k) = \pi_k$$ inclusion probability of element k

$$\mathrm{E}(I_k I_l) = \pi_{kl}$$ joint expectation of I_k and I_l

$$\sum_{k \in \mathcal{U}} \pi_k = \mathrm{E}(n)$$ expected sample size

The I_k are the *only* random variables in the design based frame work and they follow a theoretical distribution. E.g. a Hypergeometric distribution for SRS.

Inclusion Probabilities

With the inclusion probabilities design unbiased estimators can be constructed. For example an estimator for a total $\tau = \sum_{k \in \mathcal{U}} y_k$.

$$\hat{\tau} = \sum_{k \in \mathcal{S}} \frac{y_k}{\pi_k} \qquad \mathrm{E}\left(\hat{\tau}\right) = \sum_{k \in \mathcal{U}} E(I_k) \frac{y_k}{\pi_k} = \tau$$

π_k^{-1} is also called the design weight of element k.
$V(\hat{\theta}) = f(\mathcal{Y}, \Sigma)$, with $\Sigma = \left(\mathrm{E}\left(I_k I_l\right) - \mathrm{E}\left(I_k\right) \mathrm{E}\left(I_l\right)\right)_{k,l=1,\dots,N}$. For complex sampling designs Σ can be very complex too and difficult to compute. In practice it is thus often unknown to data users. However there are approximations to $V(\hat{\theta})$ that only require the π_k's and are much simpler to estimate than $V(\hat{\theta})$.

gesis Leibniz Institute for the Social Sciences

Sample Mean with SRS

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

Leibniz Gemeinschaft

Sample Mean with SRS

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

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

gesis Leibniz Institute for the Social Sciences

Sample Mean with SRS

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

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

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

Leibniz Gemeinschaft

Model-based Approach

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

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

Model-based Approach

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

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

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

Model-based Approach

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

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

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

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

Model-based Approach

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

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

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

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

Note that there is no finite population correction.

Section 2

Sampling Designs

Representative Sample

What is a representative sample?

gesis Leibniz Institute for the Social Sciences

Representative Sample

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

Representative Sample

However, what do we actually want?

Representative Sample

However, what do we actually 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.

Sampling Frames

Access to the target population is of major importance for the selection of any sample. This is often done with the help of a sampling frame, a register that links observational units to a identifier units. Then the units of the register can be sampled.

Sampling Frames

Access to the target population is of major importance for the selection of any sample. This is often done with the help of a sampling frame, a register that links observational units to a identifier units. Then the units of the register can be sampled.

For samples of persons popular sampling frame are:

- Address Registers
  - Address of buildings
  - Address of dwellings
  - Address of persons
  - Address for post delivery points
- Telephone number
  - Set of possible landline numbers
  - Set of possible mobile numbers
  - Union of possible landline and mobile numbers (Multi-Frame)

Sampling Frames

Access to the target population is of major importance for the selection of any sample. This is often done with the help of a sampling frame, a register that links observational units to a identifier units. Then the units of the register can be sampled.

Ideally the sampling frame should have one and one entry only for each observational unit of the target population. In practice it is often difficult to find such a *perfect* sampling frame, i.e. without any over or under coverage.
And Some sampling designs do not use a sampling frame at all.

Probability Based Samples

Probability Samples - A finite set of possible samples each having a certain probability of being selected, given by the sampling design. The sampling design should be measurable, that is:

- $\pi_k > 0 \ \forall \ k \in \mathcal{U}$
- $\pi_{kl} > 0 \ \forall \ k \neq l \in \mathcal{U}$

Probability Based Samples

Probability Samples - A finite set of possible samples each having a certain probability of being selected, given by the sampling design. The sampling design should be measurable, that is:

- $\pi_k > 0 \; \forall \; k \in \mathcal{U}$
- $\pi_{kl} > 0 \; \forall \; k \neq l \in \mathcal{U}$

Probability based samples do not require any (parametric) assumptions about the data to analyse them, in that respect they can be considered a robust strategies.

Non-Probability Based Samples

Non-Probability Samples - Literally speaking the sampling method should select always the same sample if repeated. Often used for selection processes that are less controlled and often to complex to be modeled.

Non-Probability Based Samples

Non-Probability Samples - Literally speaking the sampling method should select always the same sample if repeated. Often used for selection processes that are less controlled and often to complex to be modeled. Examples:

- Convenience Samples
- Purposive Samples
- Opt-in Samples
  - (Online) Access Panels
  - Invitations to a survey on webpages
- Quota Samples

Non-Probability Based Samples

Non-Probability Samples - Literally speaking the sampling method should select always the same sample if repeated. Often used for selection processes that are less controlled and often to complex to be modeled. Examples:

- Convenience Samples
- Purposive Samples
- Opt-in Samples
  - (Online) Access Panels
  - Invitations to a survey on webpages
- Quota Samples

Non-probability based samples require (often unverifiable) assumptions about the observed data to analyse it (model-based inference). They often lack a theoretical frame work that could be used to construct unbiased or consistent estimates.

Sampling Algorithms

A sampling algorithm is a set of rules used to select a sample from a population. Two distinctions can be made:

Sampling Algorithms

A sampling algorithm is a set of rules used to select a sample from a population. Two distinctions can be made:

- Algorithms for simple random sampling. All samples have the same probability of being selected, i.e. $p(\mathscr{s})$ is constant for all possible samples.

Sampling Algorithms

A sampling algorithm is a set of rules used to select a sample from a population. Two distinctions can be made:

- Algorithms for simple random sampling. All samples have the same probability of being selected, i.e. $p(\mathcal{s})$ is constant for all possible samples.
- Algorithms for unequal probability sampling. There can be difficulties with non-random samples sizes, (other constraint such as balancing on auxiliary variables) and large sampling frames.

Sampling Algorithms

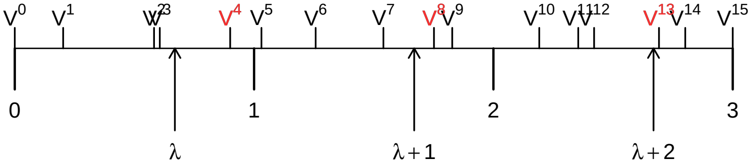
A sampling algorithm is a set of rules used to select a sample from a population. Two distinctions can be made:

- Algorithms for simple random sampling. All samples have the same probability of being selected, i.e. $p(\mathcal{s})$ is constant for all possible samples.
- Algorithms for unequal probability sampling. There can be difficulties with non-random samples sizes, (other constraint such as balancing on auxiliary variables) and large sampling frames.

A sequential sampling algorithm can be applied to a sampling frame. That is, it is not necessary to enumerate all samples in a support of sampling design to select one of them.

Systematic Sampling

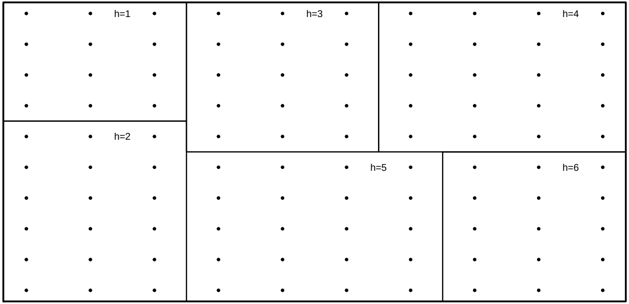
The elements of the population are brought into a specific ordered and $V^i = \sum_{k=1}^{i} \pi_k$. A value λ is selected from a uniform distribution between 0 and 1.



Systematic selection remains popular because of its simplicity. Although unbiased variance estimation is in general not possible.

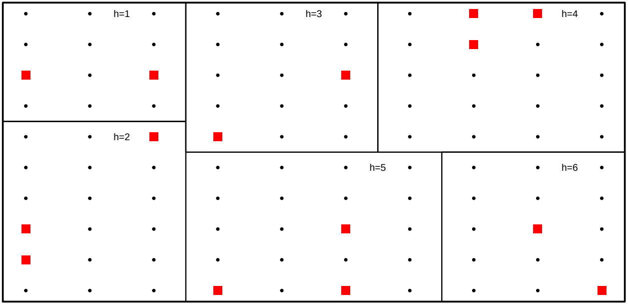
Stratification

A Population of 100 elements is stratified into $H = 6$ strata.



Stratification

A Population of 100 elements is stratified into $H = 6$ strata.
14 elements are selected population and their allocation is given by $n_1 = 2$ $n_2 = 3$ $n_3 = 2$ $n_4 = 3$ $n_5 = 3$ $n_6 = 2$



Defining the Strata

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	$N-1$	$\text{SSTO} = (N-1)V^2$

Stratification can reduce the sampling variance of estimators. The more homogeneous the strata are the higher is the gain in efficiency from using a stratified sample sample instead of SRS. That is if the SSW (variance within) is considerably smaller that than the SSB (variance between).

Allocation Methods

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

$$n_h = \begin{cases} \dfrac{n}{H} & \text{equal allocation} \\ \dfrac{N_h}{N} n & \text{proportional allocation} \\ \dfrac{N_h V_h}{\sum_{h=1}^{H} N_h V_h} n & \text{optimal allocation} \end{cases},$$

Proportional allocation can also be done with respect to another variable, e.g. $\frac{\tau_h}{\tau} n$

Example Stratification

We would like to estimate the difference in the mean Academic Performance Index (API) of all Californian schools between year 1999 and 2000 (32.8). To do that we select from all Californian schools two samples. One sample in 1999 and one in 2000. Both samples are selected by a stratified (simple random) sample, where the Counties of California are used as the strata. The samples size for both samples is 205. From each County at least 2 schools are selected. The rest of the sampled size is allocated proportionally to the number of schools in the strata. The inclusion probability of a school in a particular stratum is the number of schools selected from that stratum divided by the total number of schools in that stratum.

Example Stratification

We use two estimator for variance estimation. One is design unbiased and the other is a naive estimator that uses no other design information than the design weights ($\hat{\sigma}^2/n$).

	Est	Vest	CI.lb	CI.ub
Design	24.07	231.501	-5.754	53.888
Naive	24.07	190.810	-3.007	51.141

The stratification seems to be not very effective. So we construct 10 strata that are more homogeneous with regard to API_{99} and API_{00} (using *k-means*) and select two new samples.

	Est	Vest	CI.lb	CI.ub
Design	33.24	4.879	28.916	37.574
Naive	33.24	165.890	8.001	58.489

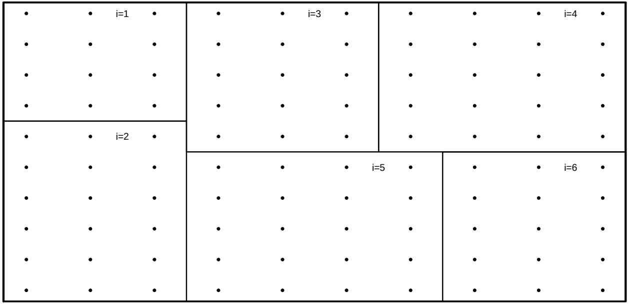
Example Stratification

We repeat the sampling with the better stratification 1000 times and compute the coverage rates for our confidence intervals.

	Design	Naive
Coverage Rate	0.965	1.000

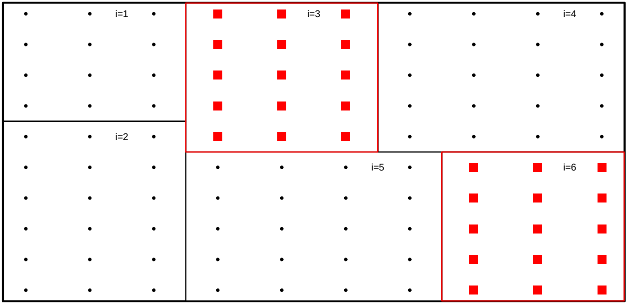
Clustering

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



Clustering

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



Cluster Sampling

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

gesis Leibniz Institute for the Social Sciences

Cluster Sampling

Sampling elementary units is often not feasible (e.g. persons or businesses). Maybe there is no uniform sampling frame available to select them from, or it would be costly to do, because the selected elements would scatter to much over the a certain area and travel costs of interviewers would be to high. Thus, it is very common to select clusters, so called *primary sampling units* (PSU's) that are populated by *secondary sampling units* (SSU's).

Cluster Sampling

Sampling elementary units is often not feasible (e.g. persons or businesses). Maybe there is no uniform sampling frame available to select them from, or it would be costly to do, because the selected elements would scatter to much over the a certain area and travel costs of interviewers would be to high. Thus, it is very common to select clusters, so called *primary sampling units* (PSU's) that are populated by *secondary sampling units* (SSU's). Cluster sampling makes it still possible to obtain unbiased estimates but it can have a big influence on the variance.

Cluster Sampling

Sampling elementary units is often not feasible (e.g. persons or businesses). Maybe there is no uniform sampling frame available to select them from, or it would be costly to do, because the selected elements would scatter to much over the a certain area and travel costs of interviewers would be to high. Thus, it is very common to select clusters, so called *primary sampling units* (PSU's) that are populated by *secondary sampling units* (SSU's). Cluster sampling makes it still possible to obtain unbiased estimates but it can have a big influence on the variance. Compared to stratification cluster sampling tends to increase the sampling variance. What makes stratification efficient, a small within variance, has the opposite effect on cluster sampling.

Example Clustering

Now we use for our Californian school survey cluster sampling. Both samples are selected by a (simple) cluster sample, where the clusters are the School Districts of California. 25 clusters are selected for both samples and the expected number of schools in each sample is 205. Each cluster has the same inclusion probability, 0.0330251 (25 divided by 757, the number of clusters).

gesis Leibniz Institute for the Social Sciences

Example Clustering

We use two estimator for variance estimation. One is design unbiased and the other is a naive estimator that uses no other design information than the design weights ($\hat{\sigma}^2/n$).

	Est	Vest	CI.lb	CI.ub
Design	110.35	1755.376	28.229	192.463
Naive	110.35	168.378	84.914	135.779

Example Clustering

We repeat the sampling 1000 times and compute the coverage rates for our confidence intervals.

	Design	Naive
Coverage Rate	0.863	0.415

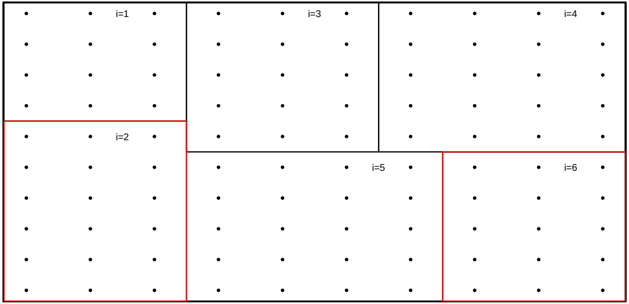
Because of the under estimation by the naive variance estimator the naive approach results in a severe under coverage. The design based approach does not under estimate the variance but their is a problem with the application of the CLT for building the confidence intervals.

Example Clustering

We repeat the simulation, but only with 100 replications and this time we sample the clusters proportional to their number of schools. Thus the inclusion probability of each cluster is $\frac{N_i}{N} * 25$, where N_i is the number of schools in the i-th cluster and N the total number of schools (6194).

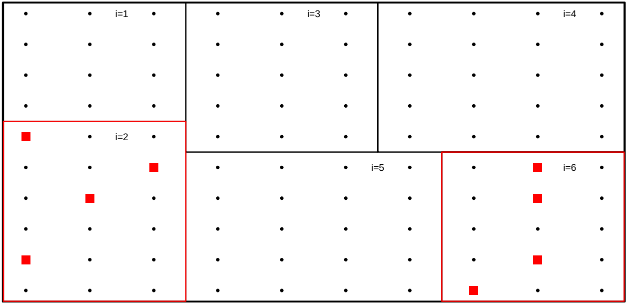
	Design	Naive
Coverage Rate	0.980	0.330

Two Stage Sampling



Two Stage Sampling

and $n_i = 4$ elements are selected from each sampled cluster.



Two Stage Sampling

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

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

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

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

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

Example Two Stage Sampling

For our Californian schools would like to estimate the following model $\text{API}_{00} = \text{ell} + \text{meals} + \text{mobility} + \text{stype}$, where

ell = English Language Learners (percent)

meals = Percentage of students eligible for subsidized meals,

mobility = percentage of students for whom this is the first year at the school,

stype = Elementary/Middle/High School

Now we use a two stage sample. As PSUs the counties of California are used, the SSU are the schools. 25 PSUs are selected with probablity proportional to their number of schools. Within each selected PSU 2 schools are sampled by a SRS.

gesis Leibniz Institute for the Social Sciences

Example Two Stage Sampling

Table : Naive

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	853.819	24.221	35.251	0.000
ell	-1.475	0.616	-2.394	0.021
meals	-3.217	0.460	-6.999	0.000
mobility	1.173	1.368	0.858	0.396
stypeH	-110.767	23.125	-4.790	0.000
stypeM	7.754	45.369	0.171	0.865

Table : Design

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	853.819	25.084	34.039	0.000
ell	-1.475	0.626	-2.356	0.036
meals	-3.217	0.409	-7.866	0.000
mobility	1.173	1.257	0.933	0.369
stypeH	-110.767	49.419	-2.241	0.045
stypeM	7.754	46.304	0.167	0.870

Section 3

Sample Size Planning

Selecting a Sample Size

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 $\mathrm{CV}(\hat{\theta}) = \frac{\sqrt{\mathrm{V}(\hat{\theta})}}{\hat{\theta}}$.
Set $\mathrm{CV}(\bar{y}) = \mathrm{CV}_0$ as a precision requirement (representative!).

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

Selecting a Sample Size

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

Sample Size for Proportions

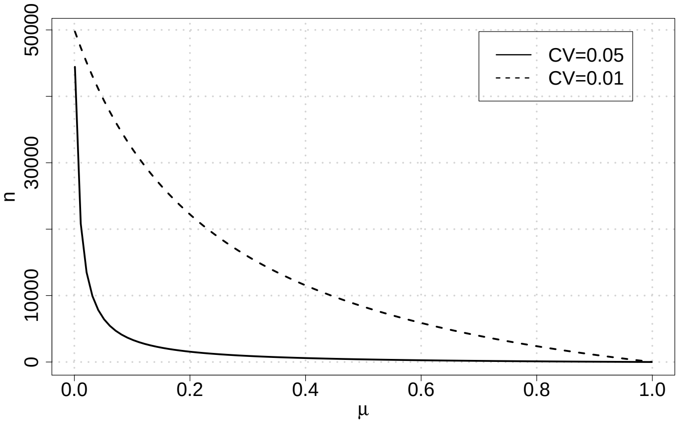
If the variable of interest is binary we have

$V(\overline{y})_{SRS} = \frac{\mu(1-\mu)}{n}\frac{N-n}{N-1}$ and

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

$\lim_{\mu \to 0} CV^2(\overline{y})_{SRS} = \infty$, thus for rare observation to meet a CV target the sample size can become very large.

Sample Size for Proportions



gesis Leibniz Institute for the Social Sciences

Sample Size for Proportions

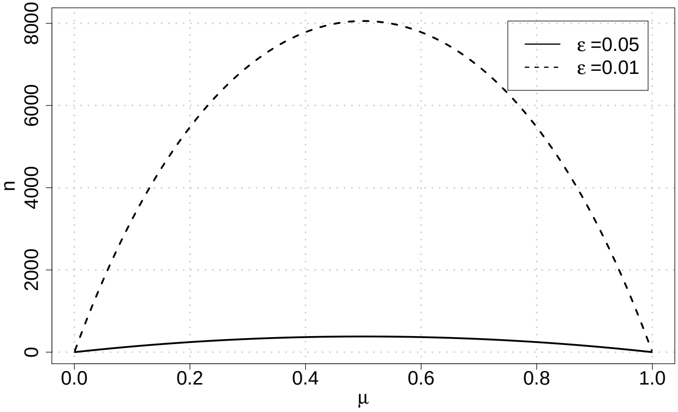
If the variable of interest is binary we have

$V(\bar{y})_{SRS} = \frac{\mu(1-\mu)}{n}\frac{N-n}{N-1}$ and

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

$\lim_{\mu\to 0} CV^2(\bar{y})_{SRS} = \infty$, thus for rare observation to meet a CV target the sample size can become very large. The target for $V(\bar{y})_{SRS}$ can be set to achieve a CI's with a maximal length of 2ϵ.

Sample Size for Proportions



Literature I

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.

Leibniz Institute for the Social Sciences

Literature II

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