gesis

Leibniz Institute
for the Social Sciences

Sample Theory

Epidemiological Study Design and Statistical Methods

Stefan Zins - GESIS

12.12.2017

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?

Motivation

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

Motivation

- Did you ever work with sample data?
- What kind of analysis have you done with sample data?
- What concerns did you have while applying your analytically methods?

Section 1

Inference

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:

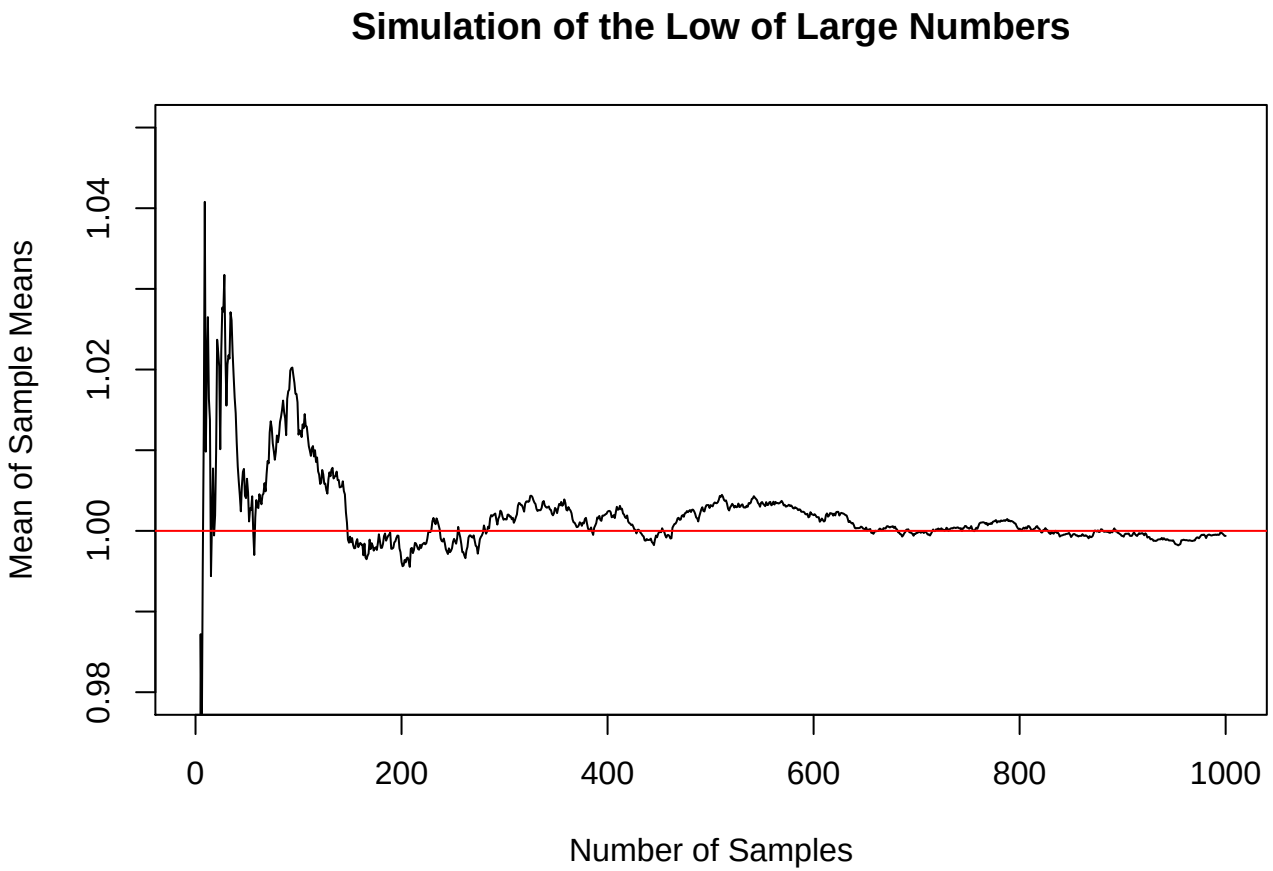
$$\bar{y} \xrightarrow{P} \mu$$

If the LLN holds we can have unbiased estimates, as our estimates will converge in probability to their expected (true) value.

LLN Demonstration

Suppose all y_k 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 $V(y_i) = \sigma^2 < \infty \; \forall \; i = 1, \ldots, N$. Then for $n \to \infty$ then we have:

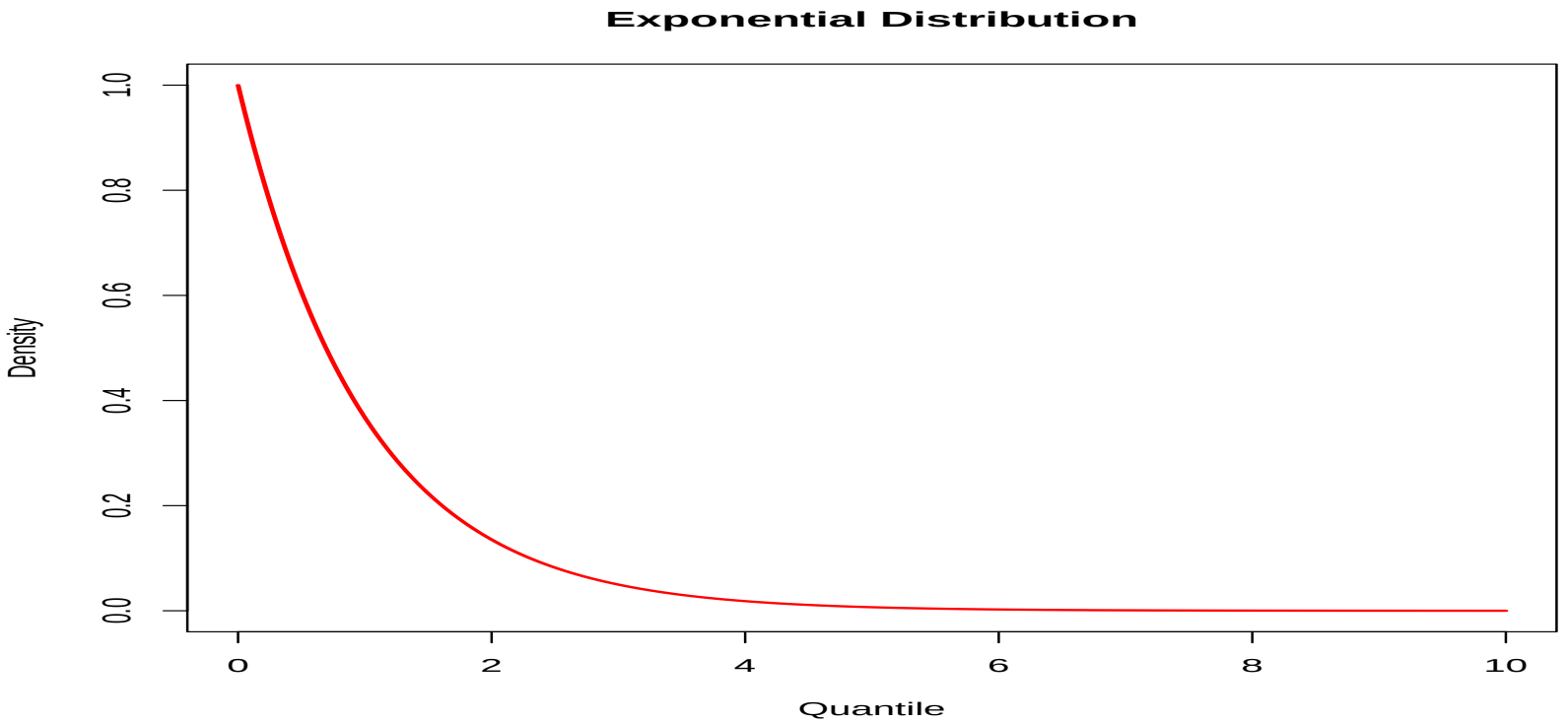
$$\sqrt{n}(\bar{y} - \mu) \xrightarrow{d} N\left(0, \sigma^2\right)$$

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

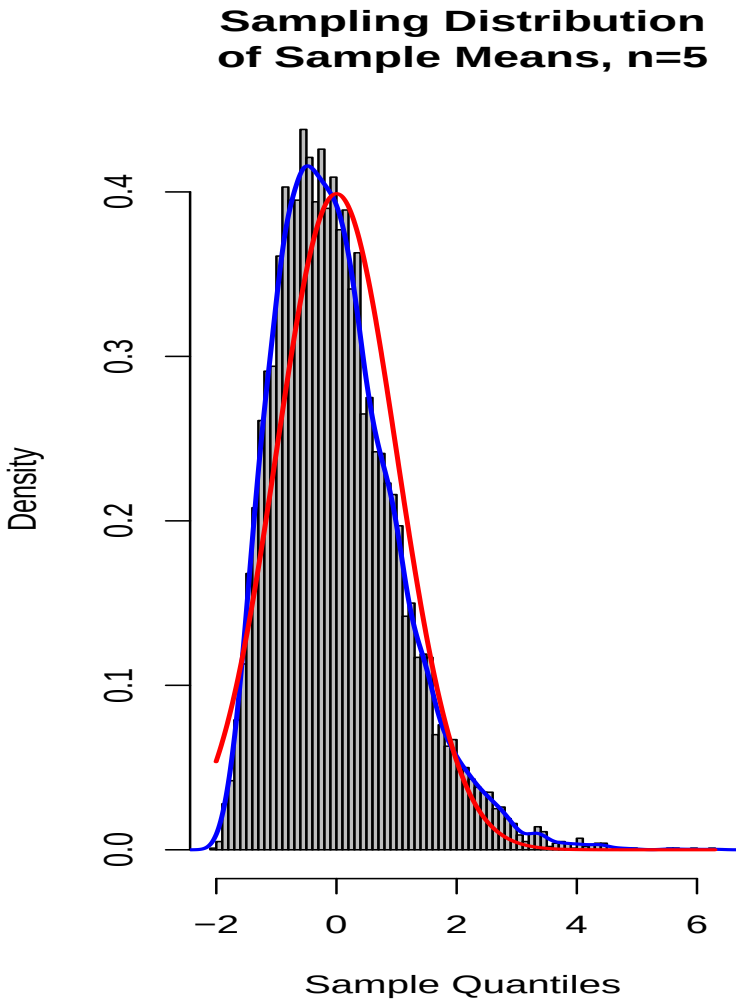
$$\left[\bar{y} + \Phi(\alpha/2)\sigma\sqrt{n}\,;\, \bar{y} + \Phi(1-\alpha/2)\sigma\sqrt{n}\right]$$

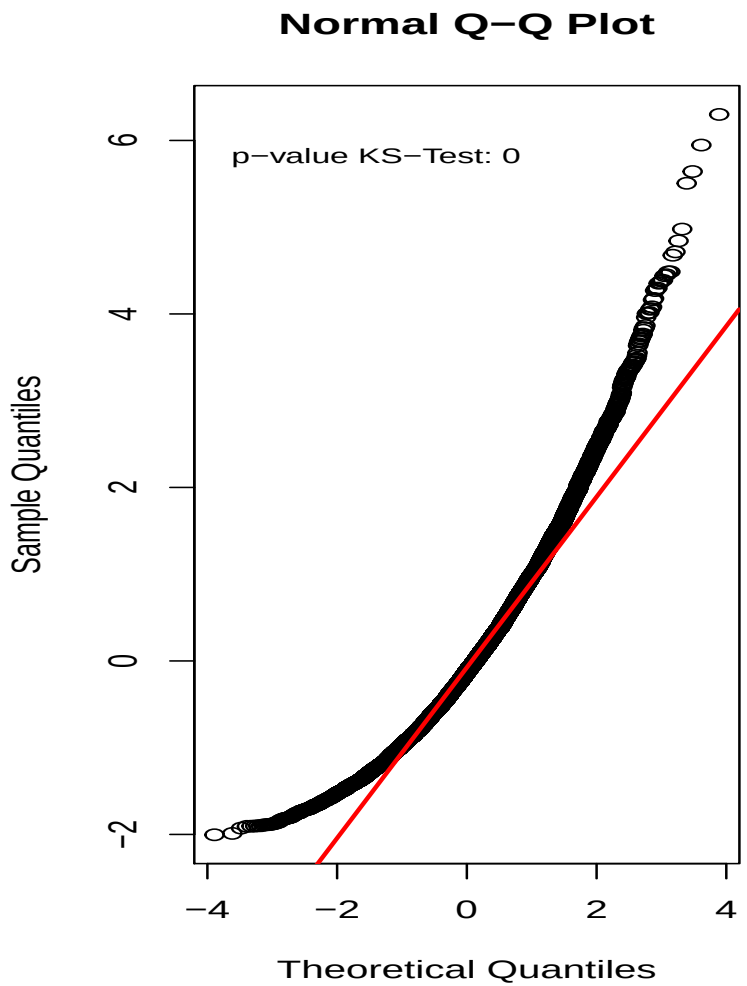
CLT Demonstration

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

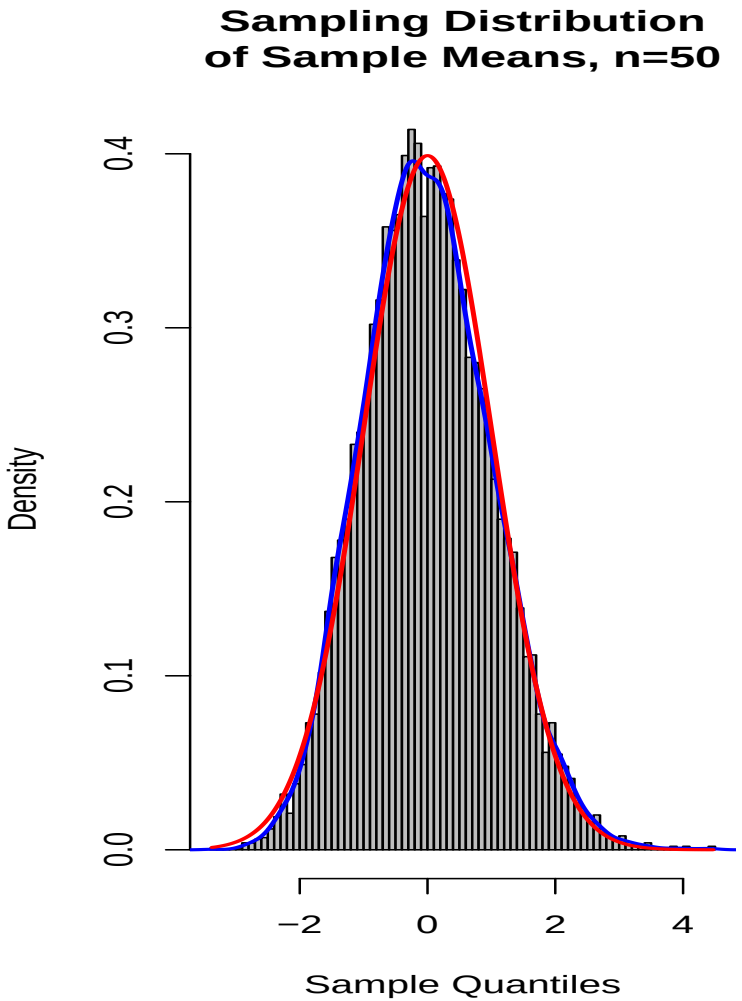


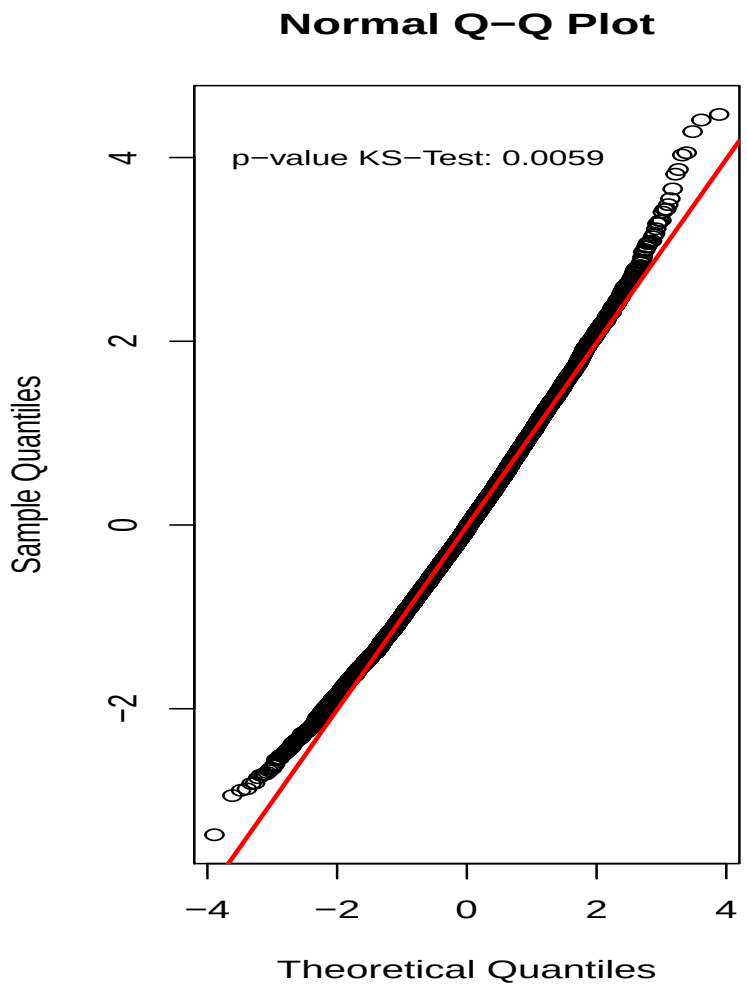
CLT Demonstration



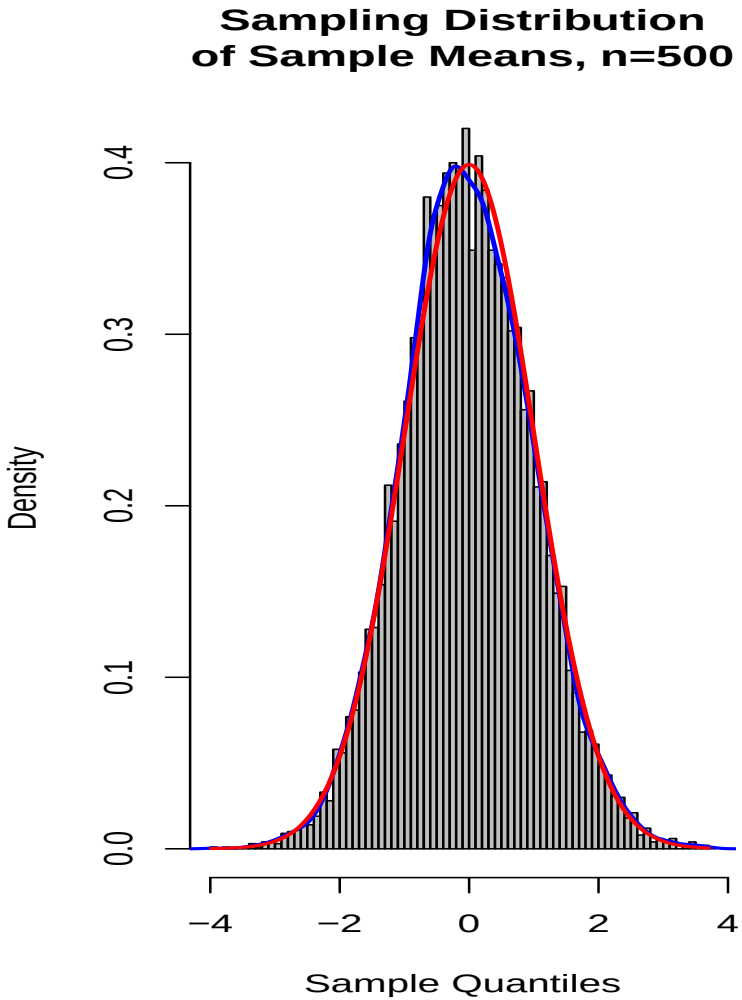


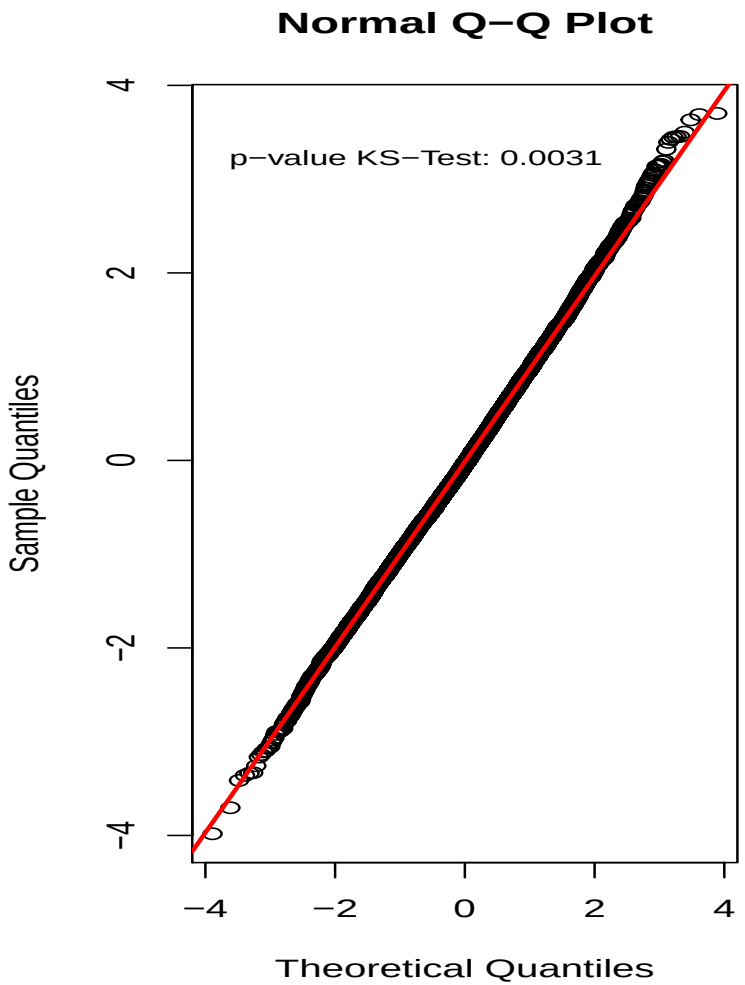
CLT Demonstration



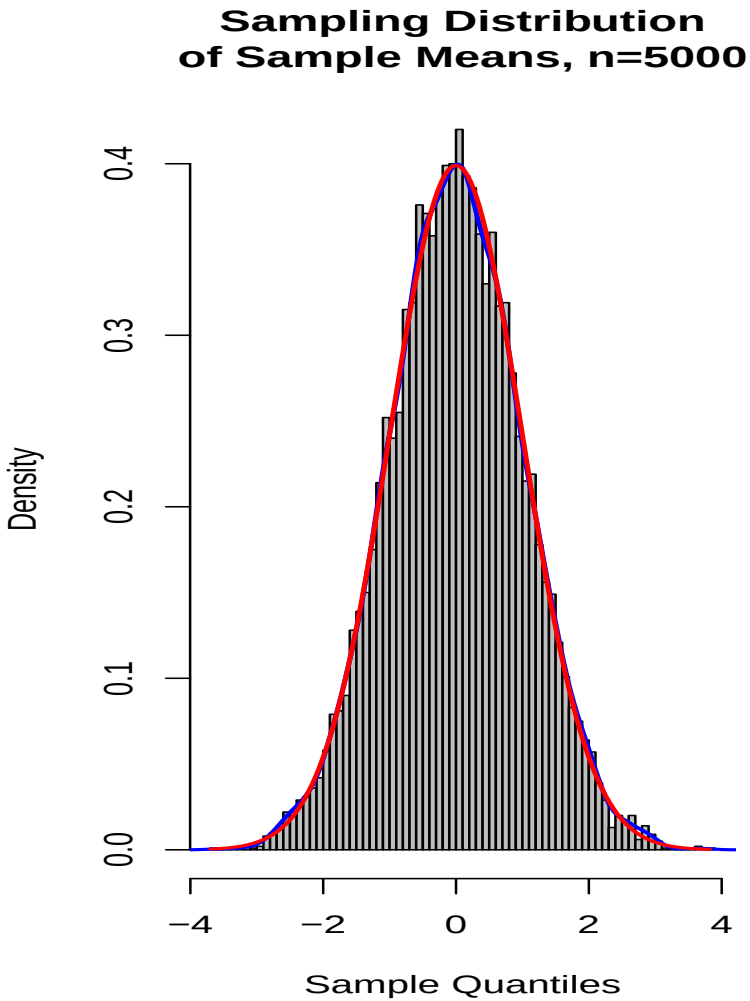


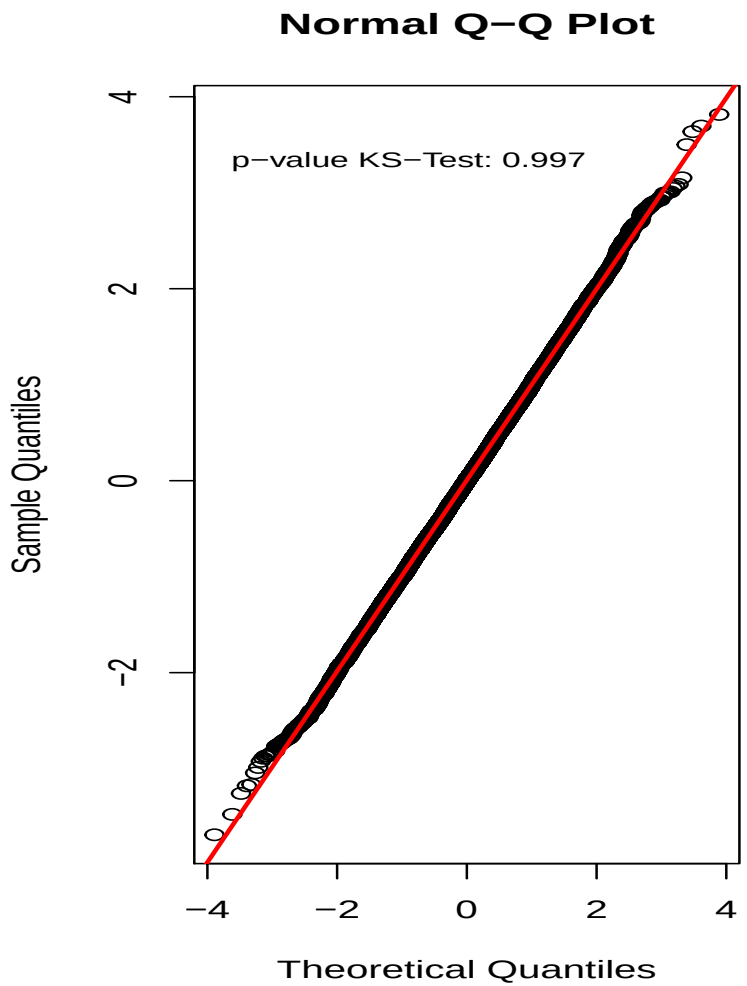
CLT Demonstration





CLT Demonstration





Section 2

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

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

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

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

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

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

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

Inclusion Probabilities I

$$I_k = \begin{cases} 1 & \text{if } k \in s \\ 0 & \text{else} \end{cases}$$ 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 II

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

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

$\pi_k^{-1} = d_k$ is also called the design weight of element k.
$V(\hat{\theta}) = f(\mathcal{Y}, \Sigma)$, with $\Sigma = (\mathrm{E}(I_k I_l) - \mathrm{E}(I_k)\,\mathrm{E}(I_l))_{k,l=1,\ldots,N}$. 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 $V(\hat{\theta})$.

Sample Mean with SRS

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

Sample Mean with SRS

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

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

Sample Mean with SRS

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

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

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

Model-based Approach

The sample data: $\{y_1, y_2, \ldots, y_k, \ldots, y_n\}$, a sequence of i.i.d. random variables with

$$y_k \sim N(\mu, \sigma^2) \quad k = 1 \ldots, n.$$

Model-based Approach

The sample data: $\{y_1, y_2, \ldots, y_k, \ldots, y_n\}$, a sequence of i.i.d. random variables with

$$y_k \sim N(\mu, \sigma^2) \quad k = 1 \ldots, n.$$

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

Model-based Approach

The sample data: $\{y_1, y_2, \ldots, y_k, \ldots, y_n\}$, a sequence of i.i.d. random variables with

$$y_k \sim N(\mu, \sigma^2) \quad k = 1 \ldots, n.$$

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

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

Model-based Approach

The sample data: $\{y_1, y_2, \ldots, y_k, \ldots, y_n\}$, a sequence of i.i.d. random variables with

$$y_k \sim N(\mu, \sigma^2) \quad k = 1 \ldots, n.$$

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

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

Note that there is no finite population correction.

Section 3

Sampling Designs

Representative Sample

What is a representative sample?

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.

Representative Sample

However, what do we actually want?
We want to estimate a statistic of interest with a certain level of precision.
The term *representativ* does not come up in survey statistics

Elements of a Sampling Design

- Sampling Frame(s)
  - Access to the target population is of major importance for the selection of any sample

Elements of a Sampling Design

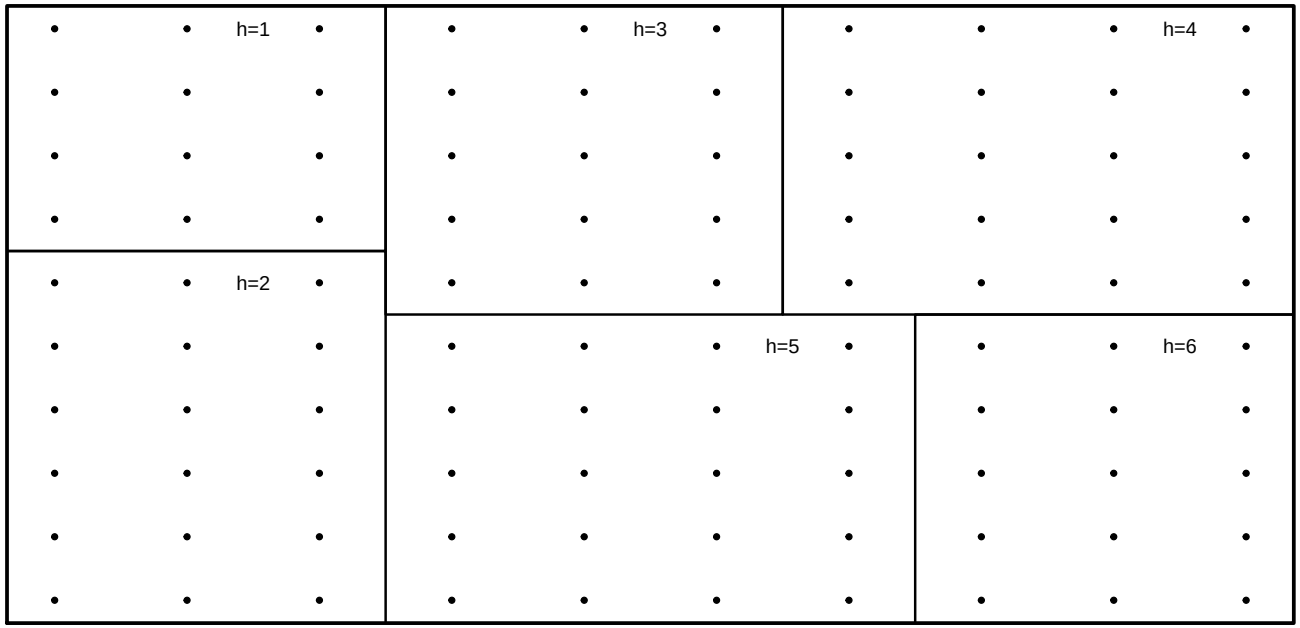
- Sampling Frame(s)
  - Access to the target population is of major importance for the selection of any sample
- Sampling Algorithm
  - A software implentation is needed

Elements of a Sampling Design

- Sampling Frame(s)
  - Access to the target population is of major importance for the selection of any sample
- Sampling Algorithm
  - A software implentation is needed
- Inklusion Probabilities
  - $\pi_k > 0 \; \forall \; k \in \mathcal{U}$
  - $\pi_{kl} > 0 \; \forall \; k \neq l \in \mathcal{U}$

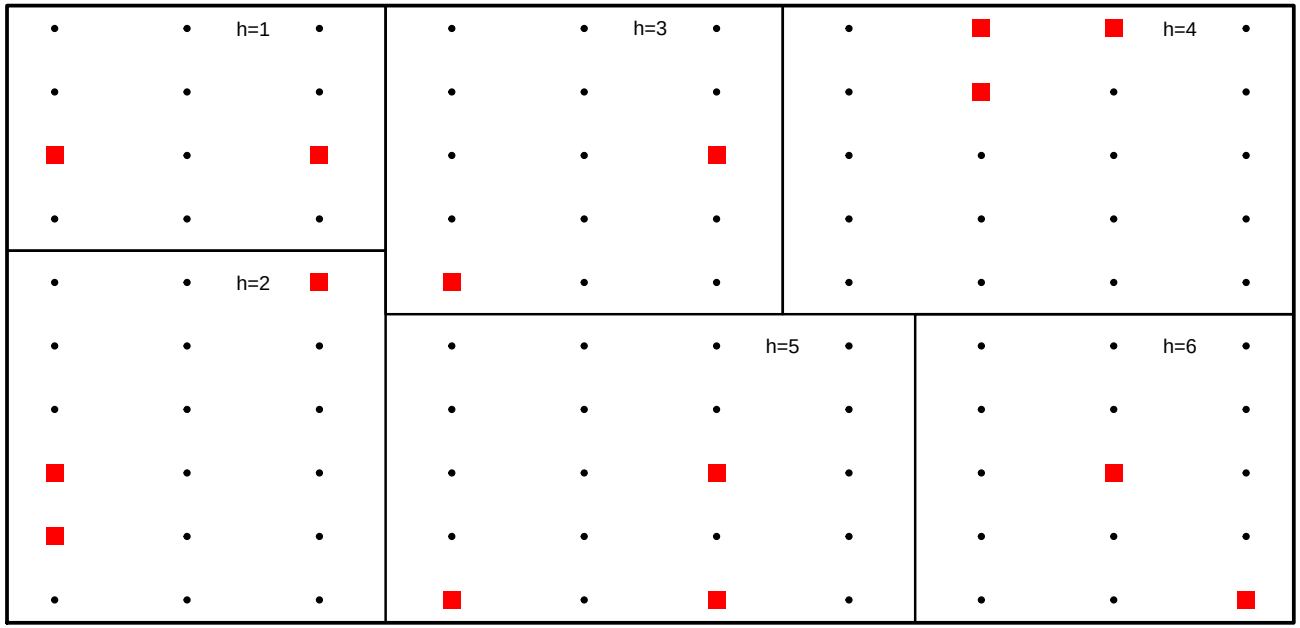
Stratification

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	$H-1$	SSB $= \sum_{h=1}^{H} N_h(\mu_h - \mu)^2$
Within strata	$N-H$	SSW $= \sum_{h=1}^{H}(N_h - 1)V_h^2$
Total, about μ_y	$N-1$	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. $\dfrac{\tau_h}{\tau} n$

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

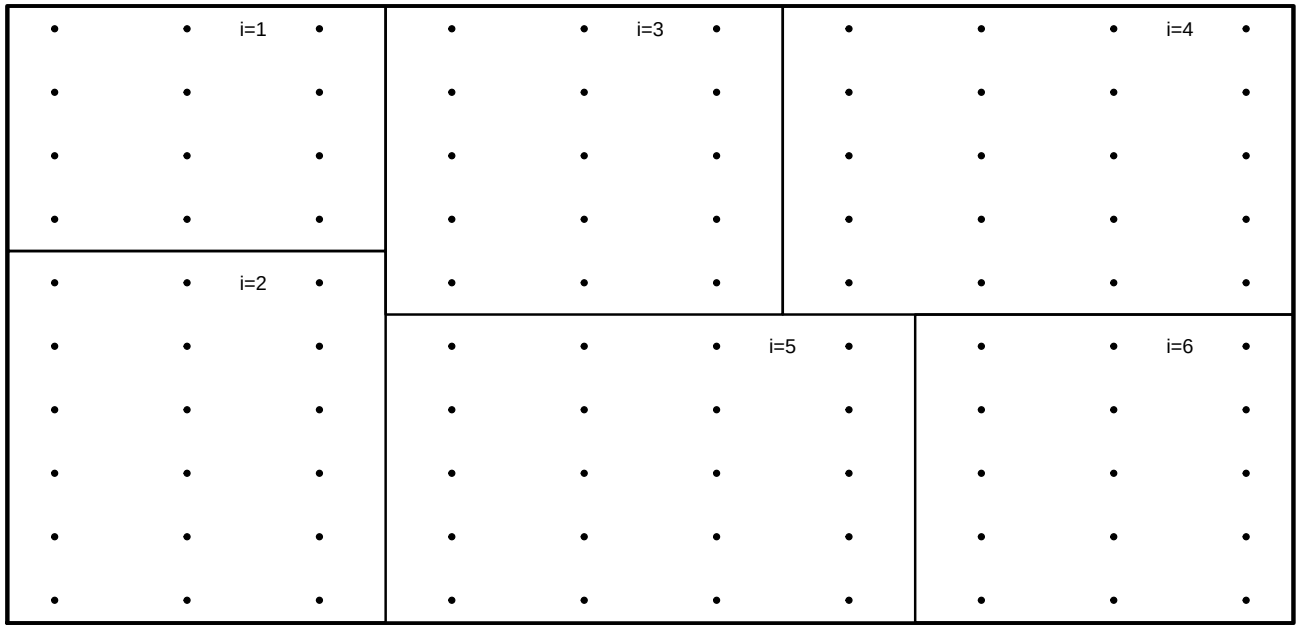
Example Stratification

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

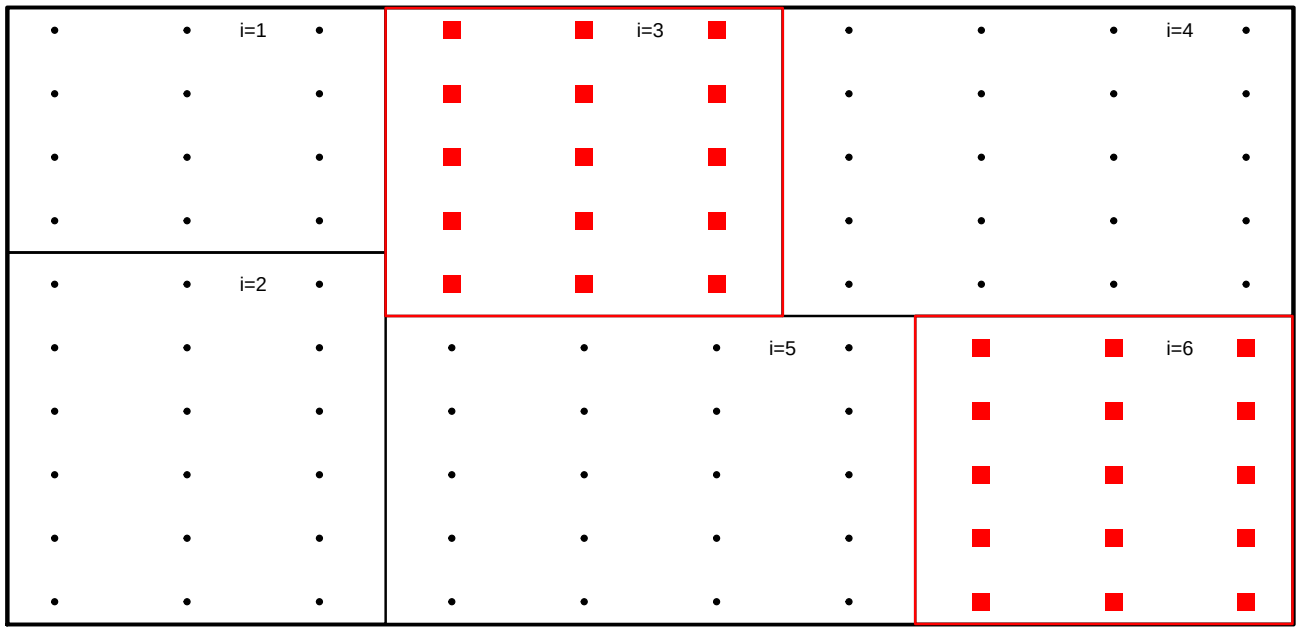
Clustering

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



Clustering

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



Cluster Sampling

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

Cluster Sampling

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

Cluster Sampling

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

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).

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

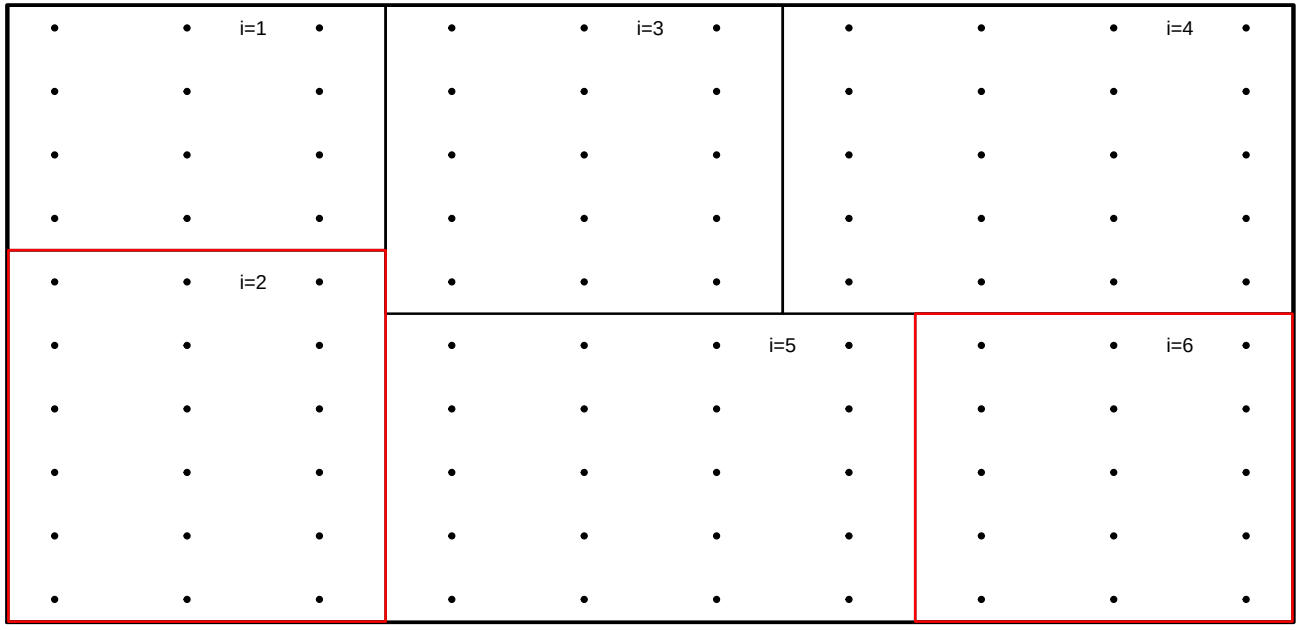
Because of the under estimation by the naive variance estimator the naive approach results in a severe under coverage. The design based approach does not under estimate the variance but their is a problem with the application of the CLT for building the confidence intervals.

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).

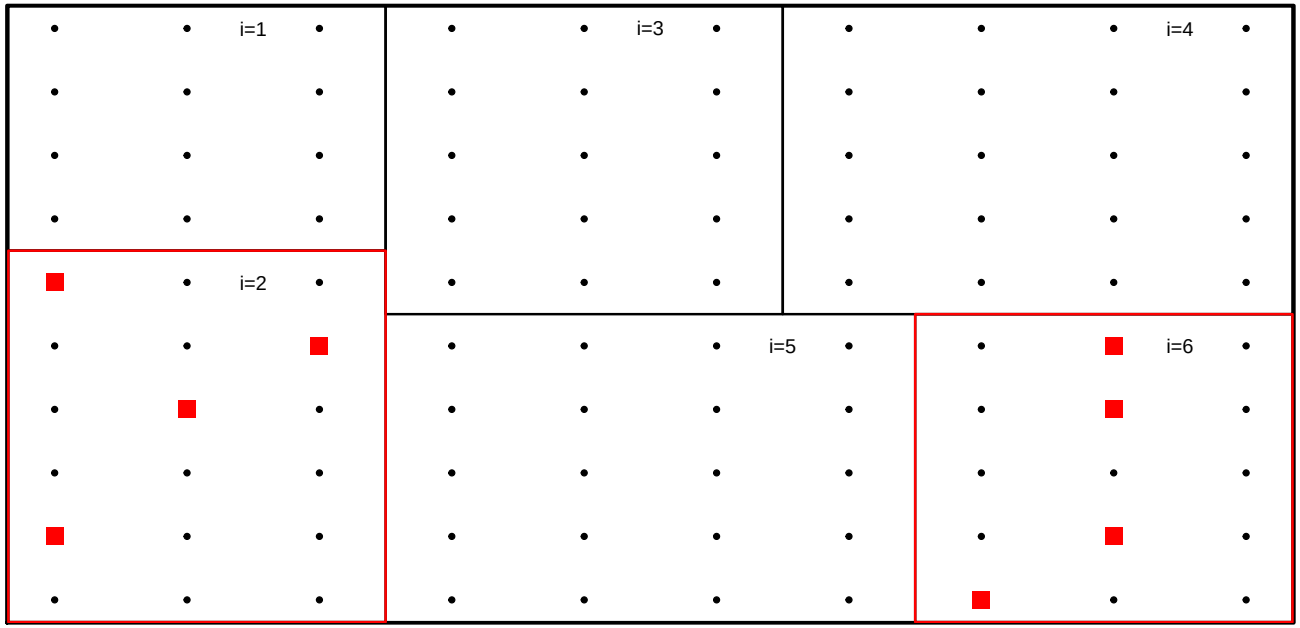
	Design	Naive
Coverage Rate	0.980	0.330

Two Stage Sampling



Two Stage Sampling

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



Section 4

Calibration Weights

Calibrating Design Weights I

The general idea is to exploit the relationship between auxiliary variables and the variable of interest to improve the efficiency of estimators.

Calibrating Design Weights I

The following problem is solved with weight calibration:
For a give design $p(.)$ and a sample $\mathcal{s}$ weights w_k for all $k \in \mathcal{s}$ have to be found that minimize

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

subject to constraints

$$\sum_{k \in \mathcal{s}} w_k \mathbf{x}_k = \sum_{k \in \mathcal{U}} \mathbf{x}_k = \tau_x$$

where $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kQ})^\top$ 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.

Calibrating Design Weights II

The calibrated weight of element k w_k can be described as $w_k = d_k g_k$, where g_k is the adjustment made to the design weight d_k. The goal is the keep g_k as close to one as possible, because the design weights have the desirable property to enable unbiased estimation.

Calibrating Design Weights II

The calibrated weight of element k w_k can be described as $w_k = d_k g_k$, where g_k is the adjustment made to the design weight d_k. The goal is the keep g_k as close to one as possible, because the design weights have the desirable property to enable unbiased estimation. Common methods to compute calibration weights are:

Calibrating Design Weights II

The calibrated weight of element k w_k can be described as $w_k = d_k g_k$, where g_k is the adjustment made to the design weight d_k. The goal is the keep g_k as close to one as possible, because the design weights have the desirable property to enable unbiased estimation. Common methods to compute calibration weights are:

- Post-stratification

Calibrating Design Weights II

The calibrated weight of element k w_k can be described as $w_k = d_k g_k$, where g_k is the adjustment made to the design weight d_k. The goal is the keep g_k as close to one as possible, because the design weights have the desirable property to enable unbiased estimation. Common methods to compute calibration weights are:

- Post-stratification
- Raking

Calibrating Design Weights II

The calibrated weight of element k w_k can be described as $w_k = d_k g_k$, where g_k is the adjustment made to the design weight d_k. The goal is the keep g_k as close to one as possible, because the design weights have the desirable property to enable unbiased estimation. Common methods to compute calibration weights are:

- Post-stratification
- Raking
- General linear Calibration

Calibrating Design Weights II

The calibrated weight of element k w_k can be described as $w_k = d_k g_k$, where g_k is the adjustment made to the design weight d_k. The goal is the keep g_k as close to one as possible, because the design weights have the desirable property to enable unbiased estimation. Common methods to compute calibration weights are:

- Post-stratification
- Raking
- General linear Calibration

$$g_k = 1 + \left(\left(\sum_{k \in \mathcal{U}} \mathbf{x}_k - \sum_{k \in s} d_k \mathbf{x}_k \right)^\top \left(\sum_{k \in s} d_k c_k \mathbf{x}_k (\mathbf{x}_k)^\top \right)^{-1} \right)^\top c_k \mathbf{x}_k$$

Example Calibration Weights

Again we sample Californian schools from the API data set.

- Design: We take a sample schools 100 with π_k's proportional to their size (i.e. number of students enrolled) and use stratification by school type (i.e. Elementary, Middle, or High School).
- Calibration: As an auxiliary variable for calibration we use a classification of schools based of their API_{99} and API_{00} values (using *k-means*).
- Estimation: We are interested in estimating the mean value of API_{00}
- Simulation: We repeat the sampling process 2500 times to assess the bias of point estimates and the coverage rate of confidence intervals (CR)

Example Calibration Weights

Table: Expected Values of Point and Variance Estimators

	Naive	Design	Cal.Naive	Cal.Design
Pest	645.9028	664.7626	664.6393	664.6393
Vest	12.8075	15.0301	15.3799	5.8138

Table: Bias and Coverage Rates

	Naive	Design	Cal.Naive	Cal.Design
Bias	-0.0284	-0.0001	-0.0002	-0.0002
CR	0.6828	0.9436	1.0000	0.9372

Lessons Learned

- The sampling design matters!

Lessons Learned

- The sampling design matters!
  - Especially for variance estimation and thus for hypothesis testing

Lessons Learned

- The sampling design matters!
  - Especially for variance estimation and thus for hypothesis testing
- Using design weights is not equivalent to considering the sampling design

Lessons Learned

- The sampling design matters!
  - Especially for variance estimation and thus for hypothesis testing
- Using design weights is not equivalent to considering the sampling design
- Calibration weights can reduce the variance of estimators

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.

Literature II


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