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

- Introduction
- Sampling Designs
- Planing

Motivation

What kind of analysis have you done with sample survey data?
What concerns did you have while applying your analytical methods?

- What do you want to do?

Motivation

What kind of analysis have you done with sample survey data?
What concerns did you have while applying your analytical methods?

-

Motivation

What kind of analysis have you done with sample survey data?
What concerns did you have while applying your analytical methods?

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

Motivation

What kind of analysis have you done with sample survey data? What concerns did you have while applying your analytical methods?

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

- What problems do you foresee?

Section 1

Desgin Based Inferenz

Finite Population, Sample, and Sampling Design

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

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

$$\mathcal{s} \subset \mathcal{U} \quad \text{sample of size } n$$

$$\mathcal{P}(\mathcal{U}) \quad \text{all possible subsets of } \mathcal{U}$$

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

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

gesis Leibniz Institute for the Social Sciences

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 (WR)

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

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

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

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

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

Inclusion Probabilities

Construct design unbiased estimators. E.g. estimator for a total $\tau = \sum_{k \in \mathcal{U}} y_k$ $\hat{\tau} = \sum_{k \in s} \frac{y_k}{\pi_k}$ $\mathrm{E}\left(\hat{\tau}\right) = \sum_{k \in \mathcal{U}} E(I_k) \frac{y_k}{\pi_k} = \tau$ Many estimator can be written as functions of totals, which makes it possible to have design consistent estimtors for them.

$V(\theta) = f(\mathcal{Y}, \Sigma)$

not only the weights

Estimation with weights Inference requires Variance estimaion

Design Weight

Sample Mean with SRS

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

Sample Mean with SRS

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

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

Sample Mean with SRS

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

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

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

Model-based Approach

The sample data: $\mathcal{Y} = \{y_1, \ldots, y_k, \ldots, y_n\}$. 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

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

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

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

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

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

Note that there is no finite population correction.

Section 2

Sampling Designs

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.

Sampling Methods

Probability based Samples - E Design should be measurable $\pi_k > 0 \forall k \in \mathcal{U}$ $\pi_{kl} > 0 \forall k \neq l \in \mathcal{U}$ Non-probability based Samples Convenience Samples Purposive Samples Opt-in Samples (Online) Access Panels - Addertising on webpages Quota Samples The selection process is often to complex to model it Assumptions are made over the data itself (model-based inference) Probability based Samples - without any (parametric) assumptions *robouts* strategie Nonprobability based Samples - inverificable assumptions but a gain in efficiency

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 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.

Techniques for probabilitic Sampling

A set of rules (algrithm)
Simple Random Sampling All samples not sample elements have the same probability of being selected. p(s) is a constant for all s
Unequal Probability Sampling
Random Routes
Cite Tille

Systematic Sampling

Law of Large Numbers (LLN)

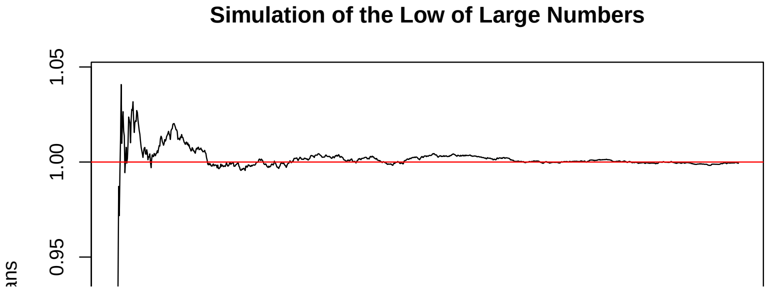
Weak law of large numbers
Suppose $\{y_1, y_2, \ldots, y_k, \ldots, y_N\}$ is a sequence of i.i.d. random variables with mean μ and $\mu \neq \infty$ and $\mu \neq -\infty$. Then for $n \rightarrow \infty$ then we have:

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

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

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.



Central Limite 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:

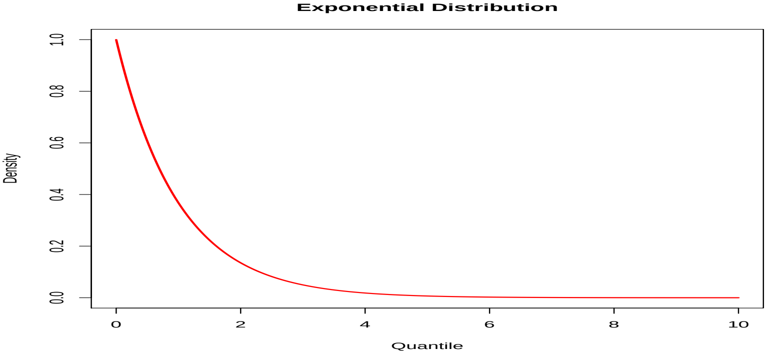
$$\frac{\bar{y} - \mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1)$$

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

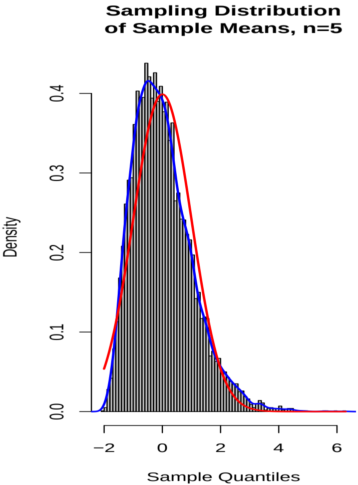
$$\left[\bar{y} - \Phi(\alpha/2)\sigma\sqrt{n}\,;\ \bar{y} + \Phi(1 - \alpha/2)\sigma\sqrt{n}\right]$$

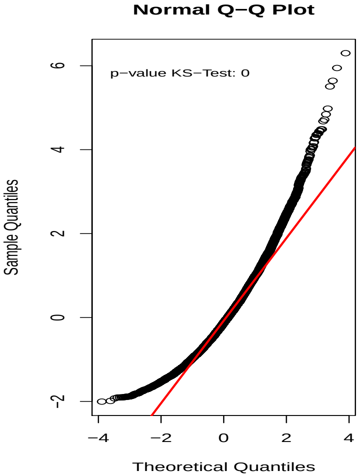
CLT Demonstration

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

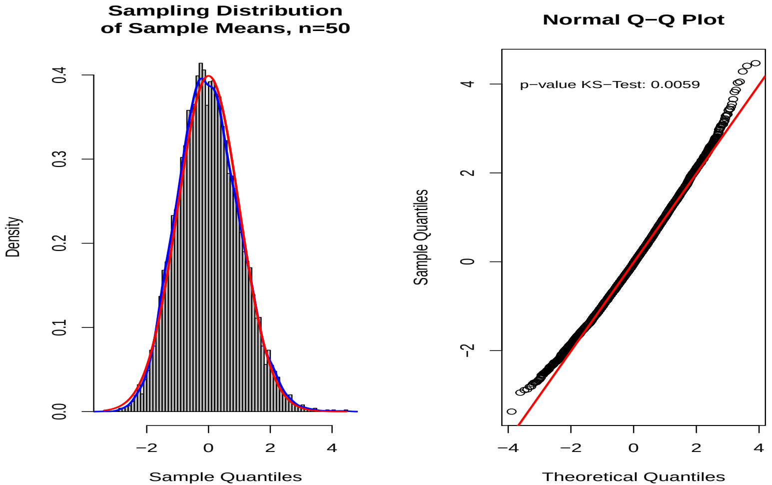


CLT Demonstration

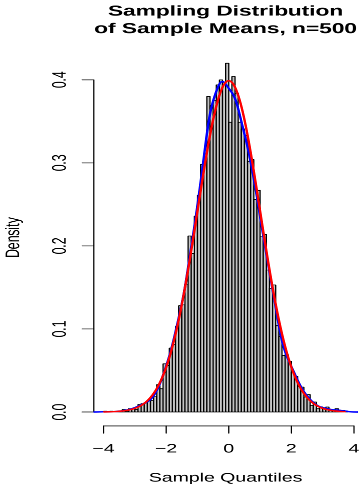


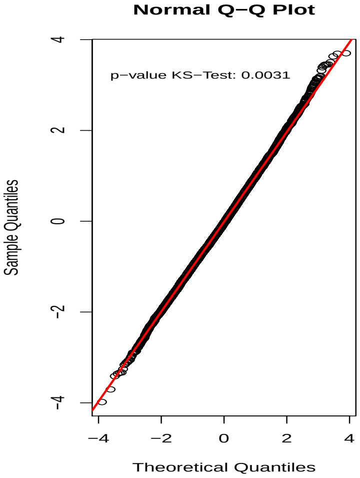


CLT Demonstration

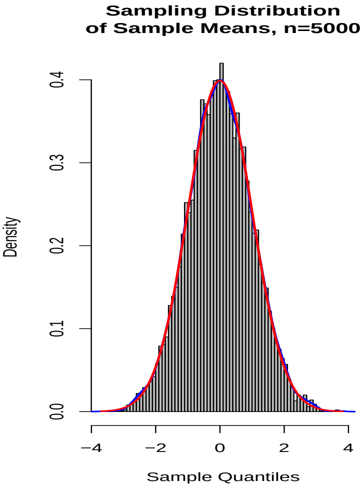


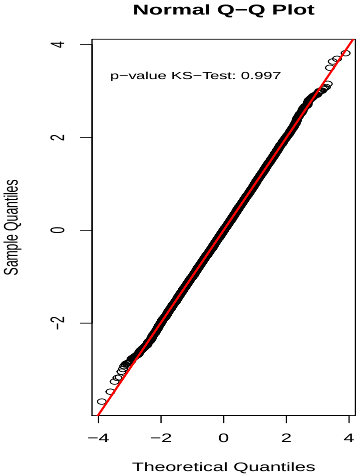
CLT Demonstration





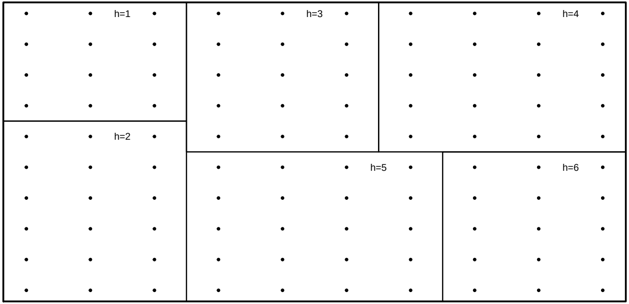
CLT Demonstration





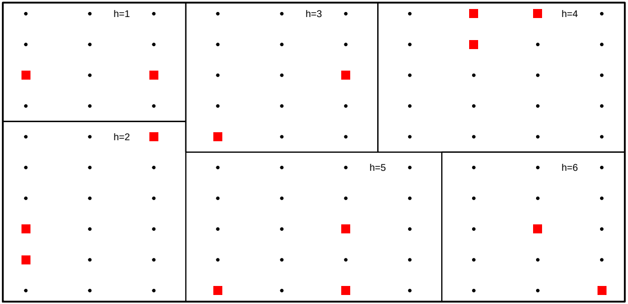
Stratification

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

Sratification can redure 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 Statification

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 Statification

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 $\text{API}_9 9$ and $\text{API}_0 0$ (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

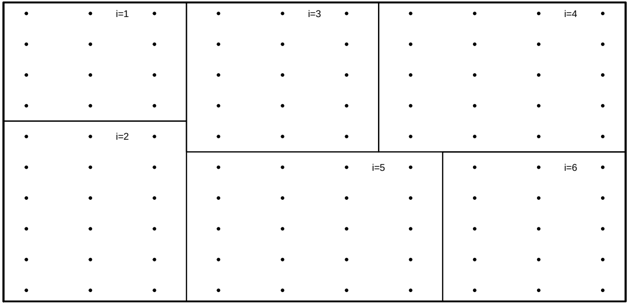
Example Statification

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

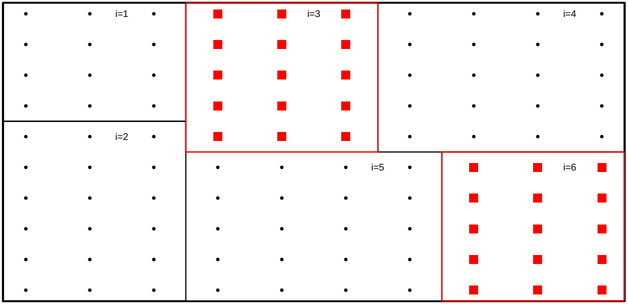
A Population of 100 elements is clustered into $N_I = 6$ cluster



gesis Leibniz Institute for the Social Sciences

Clustering

A Population of 100 elements is clustered into $N_{\text{I}} = 6$ cluster and $n_{\text{I}} = 2$ clusters are selected from the population.



Cluster Sampling

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

Cluster Sampling

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

Thus, it is very common to select clusters, so called *primary sampling units* (PSU's) that are populated by *secondary sampling units* (SSU's).

Cluster Sampling

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

Thus, it is very common to select clusters, so called *primary sampling units* (PSU's) that are populated by *secondary sampling units* (SSU's).

Cluster sampling makes it still possible to obtain unbiased estimates but it can have a big influence on the variance.

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

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 desing 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

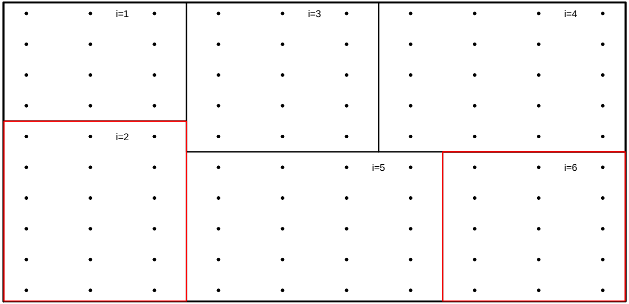
Because of the under estimation by the naive variance estimator the naive approach results in a severe under coverage. The design based approach does not under estimate the variance but their is a problem with the application of the CLT for building the confidence intervals.

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).

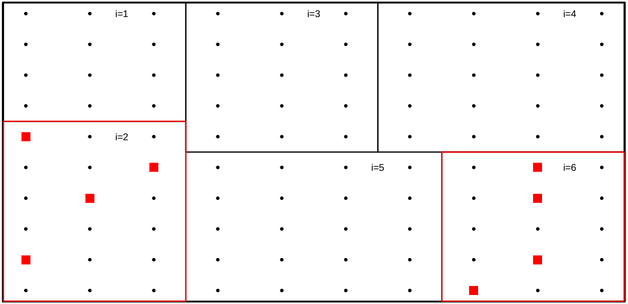
	Design	Naive
Coverage Rate	0.980	0.330

Two Stage Sampling



Two Stage Sampling

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



Two Stage Sampling

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

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

The resulting sample of SSU's is denote $\mathcal{s} = \bigcup_{i \in \mathcal{s}_{\mathrm{I}}} \mathcal{s}_i$. 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

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

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

Example

Compare Variances glm vs. lm
Mult-stage sampling by cnum + snum

Sample Size Determination

Samples Size are planned with a specific estimator in mind
Complex Problem for Multivariate Surveys
the minimum sample size under a certain precision requirements
The variance or MSE of an estimator
clustering stratification

Example

For proportions and SRS Sample

Vielen Dank für die Aufmerksamkeit