gesis Leibniz Institute for the Social Sciences

Sample Theory

Epidemiological Study Design and Statistical Methods

Stefan Zins - GESIS

12.12.2017

Content

- Introduction
- Sampling Designs
- Planing

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
$\mathscr{s} \subset \mathcal{U}$	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, $p : \mathcal{G} \mapsto (0, 1]$.

Leibniz Institute for the Social Sciences

Estimation

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

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

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

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

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

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

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.

Examples

Does this
Difference of mean Regression

Inclusion Probabilities

$V(\theta) = f(y, \Sigma)$
not only the weights
Estimation with weights Inference requires Variance estimaion
Design Weight

Sampling Frames

Access to target population
Address Samples (Register that list all sampling units)
Telelephon Samples (Not fully known but all possible entries)

Sampling Methods

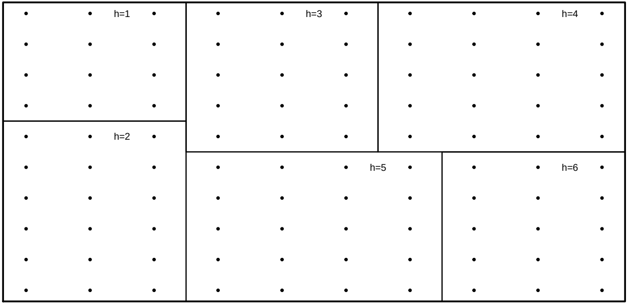
Probability based Samples
Known and Accesable Sampling Frame Desgin should be measureable $\pi_k > 0 \forall k \in \mathcal{U}$ $\pi_{kl} > 0 \forall k \neq l \in \mathcal{U}$
Nonprobability based Samples
Examples Convenience Samples Purposive Samples Opt-in Samples (Online) Access Panels - Addertising on webpages Quota Samples
The selection process is often to complex to model it
Assumptions are made over the data itself (model-based inference)

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
Systematic Sampling
Random Routes
Cite Tille

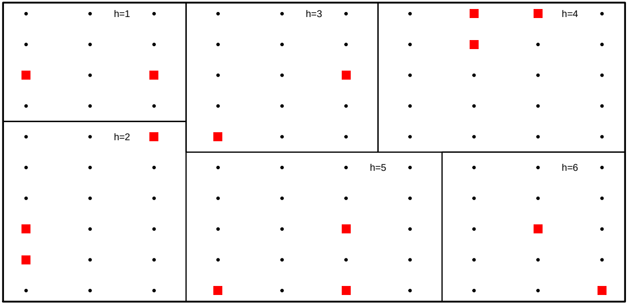
Stratification

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

Sratification can redure the sampling variance of estimators. 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.

Optimal Stratfication

Allocation Methods

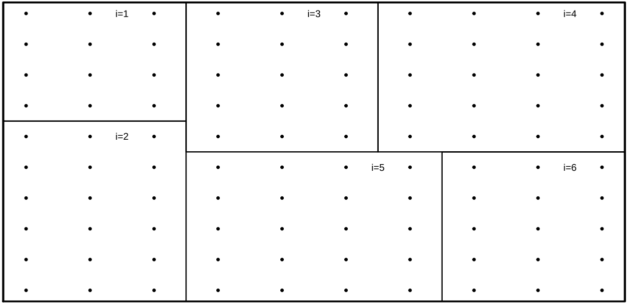
For all $h = 1, \ldots, H$

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

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

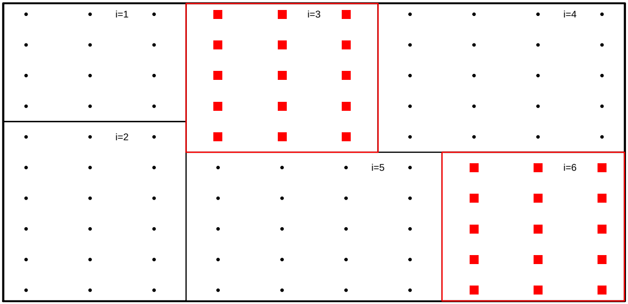
Clustering

A Population of 100 elements is clustered into $N_{\mathrm{I}} = 6$ 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.

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.

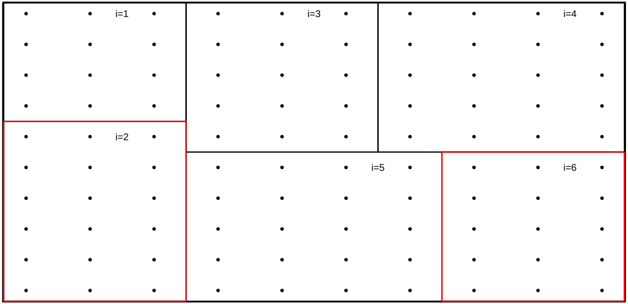
Thus, it is very common to select clusters, so called *primary sampling units* (PSU's) that are populated by *secondary sampling units* (SSU's).

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

Example

Compare Variances
Cluster sampling by dnum Stratified Sampling dnum

Two Stage Sampling



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.

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

Leibniz Institute
for the Social Sciences

Vielen Dank für die Aufmerksamkeit

Mitglied der Leibniz Gemeinschaft