Sampling and Estimation

Part 1: Introduction to Design Based Inference

Stefan Zins[1], Matthias Sand[2], and Jan-Philipp Kolb[3]

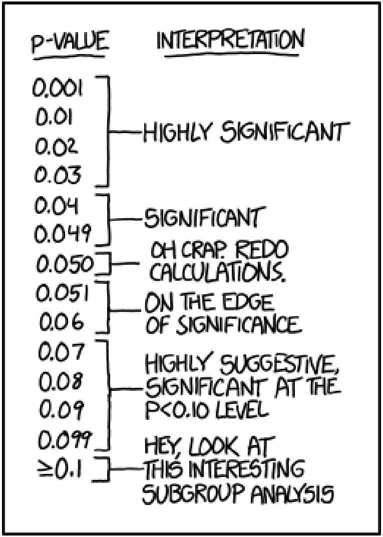
February 3, 2016

[1] Stefan.Zins@gesis.org
[2] Matthias.Sand@gesis.org
[3] Jan-Philipp.Kolb@gesis.org

Motivation

- What do you want to do?

Motivation

- 
P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

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

Motivation

- 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?

Representative Sample

What is a representative sample?

Representative Sample

What is a representative sample?
The popular concept of a representative sample is 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.

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

Estimation

$$
\begin{aligned}
\theta &= f(\mathcal{Y}) && \text{statistic of interest}\\
\hat{\theta} &= f(\mathcal{Y}, \mathscr{s}) && \text{estimator for } \theta\\
\mathrm{E}\left(\hat{\theta}\right) &= \sum_{\mathscr{s}\in\mathcal{G}} p(\mathscr{s}) f(\mathcal{Y}, \mathscr{s}) && \text{expected value of } \hat{\theta}\\
\mathrm{V}\left(\hat{\theta}\right) &= \mathrm{E}\left(\hat{\theta}^2\right) - \mathrm{E}\left(\hat{\theta}\right)^2 && \text{variance of } \hat{\theta}\\
\mathrm{MSE}\left(\hat{\theta}\right) &= \mathrm{E}\left((\hat{\theta}-\theta)^2\right) && \\
&= \left(\mathrm{E}\left(\hat{\theta}\right) - \theta\right)^2 + \mathrm{V}\left(\hat{\theta}\right) && \text{mean square error of } \hat{\theta}
\end{aligned}
$$

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

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

Expectation and Variance of a Random Sample

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

Simple Random Sampling

Without Replacement

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

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

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

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

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

Sample Mean with SRS

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

Sample Mean with SRS

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

Expected value

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

Sample Mean with SRS

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

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

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. $S_k \geqslant I_k$). The order of the selected elements is also not remembered.

$$
\begin{aligned}
\mathcal{G} &= \binom{N+n-1}{n} & p(\mathscr{s}) &= \binom{N+n-1}{n}^{-1} \\
\nu_k &= \frac{n}{N} & \pi_k &= 1 - \left(\frac{N-1}{N}\right)^n \\
\nu_{kl} &= \frac{n(n-1)}{N^2} & \pi_{kl} &= 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n \qquad \text{for } k \neq l
\end{aligned}
$$

Sample Mean with SRSWR

Expected value

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

Sample Mean with SRSWR

Expected value

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

Variance

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

Sample Mean with SRSWR

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

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

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

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

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

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

Features of Sampling Designs

- 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.

Stratified Random Sampling Notation I

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 $\mathscr{s}_h$ 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 $\mathscr{s}_h$ is denoted with n_h.

Stratified Random Sampling Notation II

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

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

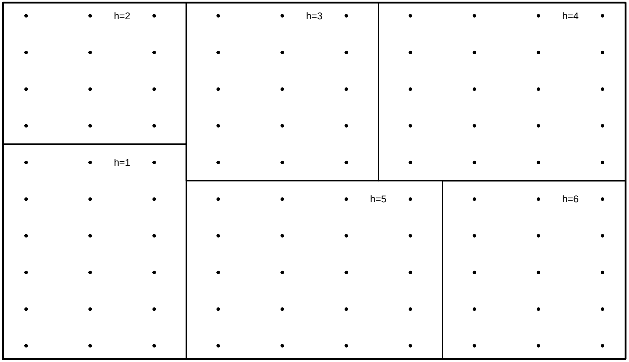
$$\sigma_h^2 = \frac{1}{N_h} \sum_{k=1}^{N_h} (y_{kh} - \mu_h)^2 \qquad \text{variance of stratum } h$$

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

Where y_{kh} as the *k*-th element in the *h*-th stratum. Sampling from stratified populations is called stratified random sampling (StrRS).

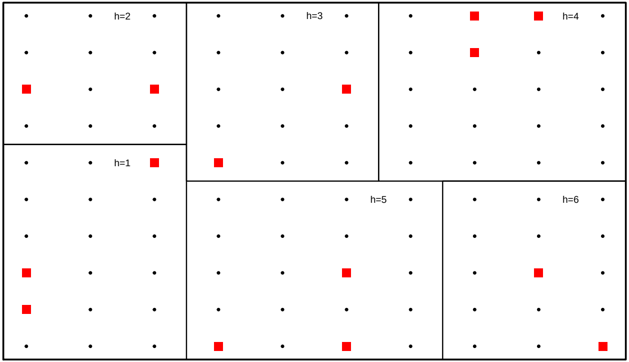
Stratification

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 from the population and their allocation is given by $n_1 = 2$ $n_2 = 3$ $n_3 = 2$ $n_4 = 3$ $n_5 = 3$ $n_6 = 2$



ESTIMATION

Estimator for the mean:

$$\overline{y}_{\text{str}} = \sum_{h=1}^{H} \gamma_h \overline{y}_h$$

where $\gamma_h = \dfrac{N_h}{N}$ and $\mathrm{E}\left(\overline{y}_{\text{str}}\right) = \mu$ for SRS and SRSWR within each stratum.

Variance and variance estimator:

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

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

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

Issues with Stratification

- 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).

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?

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.

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.

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.
- 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.
- 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

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

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.

Allocation Methods

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

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

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.

On Proportional Allocation

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

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

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

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

The Rounding Problem

- 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 $|n_h^* - n_h| \leqslant 1$ (see [1]).

Simple Systematic Sampling

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.

LITERATURE I

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.

Literature II

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