Sampling, Weighting and Estimation
Exercise 3

Stefan Zins, Matthias Sand
and Jan-Philipp Kolb

GESIS - Leibniz Institute
for the Social Sciences

February 1, 2016

Exercise 3

1. Download the data set for Germany of the 5th ESS-Round (Country File and Sampling Data)

   `http://www.europeansocialsurvey.org/data/country.html?c=germany`

2. Estimate the design effect using the variables dweight,PSU and agea (model based approach)

   Advice: The variable PSU has to be a factor

3. Calculate the effective sample size

Design Effects: Model Based Approach

Model based approach

$$d\hat{e}ff = d\hat{e}ff_p * d\hat{e}ff_c = n\frac{\sum_{h=1}^{I} d_h^2 n_h}{(\sum_{h=1}^{I} d_h n_h)^2} * (1 + (b^* - 1)\rho)$$

$$\hat{\rho}^{AOV} = \frac{MSB - MSW}{MSB + (K-1)MSW}$$

$$MSB = \frac{SSB}{I-1}; \quad MSW = \frac{SSW}{n-I}; \quad K = \frac{1}{I-1}(n - \sum_{h=1}^{I} \frac{n_h^2}{n});$$

$$b^* = \frac{\sum_{l=1}^{L}(\sum_{i=1}^{n_h} w_{li})^2}{\sum_{l=1}^{L}\sum_{i=1}^{n_h} w_{li}^2}$$

n_h is the number of units per cluster; b^* is the average cluster size; ρ reflects the Intraclass Correlation Coefficient (ICC)

⇒ *deff*$_p$ captures the design effect due to unequal inclusion probabilities

Design Effects: Model Based Approach

Obtaining *MSB*, *MSW* and b^*:

```
Ger.d <- read.spss("ESS5DE.spss/ESS5DE.sav",
                   to.data.frame = TRUE,
                   use.value.labels = TRUE)
Ger.ctry <- read.spss("ESS5_DE_SDDF.spss/ESS5_DE_SDDF.por",
                      to.data.frame = TRUE,
                      use.value.labels = TRUE)

colnames(Ger.d)[5] <- "IDNO"
Ger <- merge(Ger.d,Ger.ctry,by="IDNO", all.x = TRUE)
Ger$PSU <- as.factor(Ger$PSU)
n <- nrow(Ger)
L <- length(unique(Ger$PSU))
```

Design Effects: Model Based Approach

Obtaining ***MSB***, ***MSW*** and b^*:

```
## deffc
b <- sum(tapply(Ger$dweight,Ger$PSU,
                function(x)sum(x)^2))/sum(Ger$dweight^2)
# Calculate an anova for the regression model Age by PSU
# (Coule also be any other Variable)
SS <- anova(lm(as.numeric(Ger$agea)~Ger$PSU))
#  MSB and MSW are the means of SSB and SSW
MSB <- SS$`Mean Sq`[1]
MSW <- SS$`Mean Sq`[2]
```

Exercise 4

1. Download the following R-Script: `https://github.com/BernStZi/SamplingAndEstimation/blob/short/tutorial/Samples_for_EX4.R` to generate a Multistage- and a Cluster- Sample for the belgianmunicipalities data set
2. Calculate the mean income of the population
3. Estimate the mean income from both samples, using the `survey` package and compare the results

Multistage- and Cluster-Samples with the survey package

```
surv <- svydesign(id=~Commune+id,fpc=~prob1+prob2,
                  data=Data.be,pps="brewer")
```

- In *Exercise 1* we had a single-stage sample, therefore the argument `id` has been set to 0 or 1

⇒ In case of a multi-stage sampling approach, every sampling stage has to be defined

  ⇒ PSU: *Commune*; SSU: *id*

- This also applies for the `fpc`-argument

⇒ *prob1* reflects the porbability of inclusion for each PSU in the sample and *prob2* the probability of inclusion for each SSU

Note: altough $\mathit{prob}1 * \mathit{prob}2 = \frac{n}{N}$ in this sample, it cannot be treated like a SRS