

SAMPLING, WEIGHTING AND ESTIMATION

EXERCISE 3

Stefan Zins, Matthias Sand
and Jan-Philipp Kolb

GESIS - Leibniz Institute
for the Social Sciences

February 1, 2016

- 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

MODEL BASED APPROACH

$$\hat{deff} = \hat{deff}_p * \hat{deff}_c = n \frac{\sum_{h=1}^l d_h^2 n_h}{(\sum_{h=1}^l d_h n_h)^2} * (1 + (b^* - 1)\rho)$$

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

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

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

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

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

```
Ger.d <- read.spss("ESS5DE.spss/ESS5DE.sav",  
                  to.data.frame = TRUE,  
                  use.value.labels = TRUE)  
Ger.ctrtry <- 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.ctrtry, by="IDNO", all.x = TRUE)  
Ger$PSU <- as.factor(Ger$PSU)  
n <- nrow(Ger)  
L <- length(unique(Ger$PSU))
```

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

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

- 1 Download the following R-Script: https://github.com/BernStZi/SamplingAndEstimation/blob/short/tutorial/Samples_for_EX3b.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 probability of inclusion for each PSU in the sample and *prob2* the probability of inclusion for each SSU
- Note:** although $prob1 * prob2 = n/N$ in this sample, it cannot be treated like a SRS

MULTISTAGE- AND CLUSTER-SAMPLES WITH THE `survey` PACKAGE

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

- `pps` should be used to define the design information; usually the second order probability of inclusion
- ⇒ If the second order probability of inclusion are unknown (or too complex to calculate), a brewer approximation can be applied to estimate the joint inclusion probabilities