Sampling, Weighting and Estimation
Exercise 3

Stefan Zins, Matthias Sand
and Jan-Philipp Kolb

GESIS - Leibniz Institute
for the Social Sciences

February 3, 2016

Exercise 3a

gesis Leibniz Institute for the Social Sciences

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

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

2. Create a `svydesign` object to estimate the mean of the variable `agea`
3. To acknowledge that the sample has been collected by a multi stage design, estimate the design effect of your estimate above using the PSU-Indicator variable

   Advice: The variable PSU has to be a factor
4. Calculate the effective sample size

Design Effects: Model Based Approach

gesis Leibniz Institute for the Social Sciences

Model based approach

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

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

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

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

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

$\Rightarrow$ $deff_p$ 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.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))
```

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

```
## deffc
b.star <- sum(tapply(Ger$dweight,Ger$PSU,
                     function(x)sum(x)^2))/sum(Ger$dweight^2)
# Calculate an anova for the regression model Age by PSU
# (Could also be any other Variable)
lin.mod <- lm(as.numeric(Ger$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]
```

Exercise 3b

gesis Leibniz Institute for the Social Sciences

1. Execute 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 dataset
2. Your workspace now contains the datasets `true_ income`, `Data.be` and `Data.be2`. `true_ income` resembles the mean of the income variable for the population of the belgianmunicipalities dataset. `Data.be` is a multistage sample with 80 PSUs and 300 individual datapoints whithin each PSU. `Data.be2` is a clustersample of 10 communes
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 $prob1 * prob2 = n/N$ in this sample, it cannot be treated like a SRS

```
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

$\Rightarrow$ 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