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 1

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

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

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

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

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

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

$\Rightarrow$ $deff_p$ 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 <- 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]
```