Analysis of Complex Survey Data: 3nd Assignment

February 20, 2023

```r
# these are the packages used:
library(haven)
library(tidyverse)
library(magrittr)
library(survey)
# If you haven't installed them you can install the packages by
# executing :
# install.packages(c("haven","tidyverse","magrittr","survey"))

# this string needs to match the location/path
# where you saved the four data sets on your computer:
dataPath <- "Path/to/my/data/"
# remember R uses slash '/' instead of backslash '\', as windows does,
# to separate directories.

ess6    <- read_dta(str_c(dataPath,"ESS6e02_4.dta"))
sddf_es <- read_dta(str_c(dataPath,"ESS6_ES_SDDF.dta"))
sddf_dk <- read_dta(str_c(dataPath,"ESS6_DK_SDDF.dta"))


calculate_mode <- function(x) {
  uniqx <- unique(na.omit(x))
  uniqx[which.max(tabulate(match(x, uniqx)))]
}

ess6 %>%
  filter(cntry == "DK") %>%
  replace_na(list(
    edulvlb =  median(.$edulvlb, na.rm = TRUE),
    agea    =  median(.$agea, na.rm = TRUE),
    gndr    =  median(.$gndr, na.rm = TRUE),
    region  =  calculate_mode(.$region)
```

```
  )) %>%
  mutate(
    gndr_c   = as.character(gndr),
    age_c    = as.character(cut(
      agea,
      breaks = c(15, 35, 55, Inf),
      include.lowest = TRUE
    )),
    edulvlb_c =
      case_when((edulvlb >= 0   & edulvlb <= 229) | edulvlb > 800  ~ "low",
                (edulvlb >= 311 & edulvlb <= 423) ~ "medium",
                (edulvlb >= 510 & edulvlb <= 800) ~ "high",
                TRUE ~ as.character(edulvlb) ) ,
    region_c  = as.factor(region),
    gae_c     = as.factor(str_c(gndr_c, age_c, edulvlb_c)),
    pspwght_p = pspwght * pweight * 10000,
    dweight_p = dweight * pweight * 10000,
    trstplt_c =
      case_when((trstplt >= 0   & trstplt <= 3)  ~ "low",
                (trstplt >= 4   & trstplt <= 7)  ~ "medium",
                (trstplt >= 8   & trstplt <= 10) ~ "high",
                TRUE ~ as.character(trstplt) ),
    stfeco_c =
      case_when((stfeco >= 0   & stfeco <= 3)  ~ "low",
                (stfeco >= 4   & stfeco <= 7)  ~ "medium",
                (stfeco >= 8   & stfeco <= 10) ~ "high",
                TRUE ~ as.character(stfeco) )
  ) %>%
  left_join(sddf_dk %>% select(idno, psu, stratify), by="idno") %>%
  mutate(psu = as.factor(psu),
         stratify = as.factor(stratify) )  ->
  ess_dk

ess6 %>%
  filter(cntry == "ES") %>%
  replace_na(list(
    edulvlb =  median(.$edulvlb, na.rm = TRUE),
    agea    =  median(.$agea, na.rm = TRUE),
    gndr    =  median(.$gndr, na.rm = TRUE),
    region  =  calculate_mode(.$region)
  )) %>%
  mutate(
    gndr_c   = as.character(gndr),
    age_c    = as.character(cut(
      agea,
```

```
      breaks = c(15, 35, 55, Inf),
      include.lowest = TRUE
    )),
    edulvlb_c =
      case_when((edulvlb >= 0   & edulvlb <= 229) | edulvlb > 800  ~ "low",
                (edulvlb >= 311 & edulvlb <= 423) ~ "medium",
                (edulvlb >= 510 & edulvlb <= 800) ~ "high",
                TRUE ~ as.character(edulvlb) ) ,
    region_c  = as.factor(str_sub(region, 1, 3)),
    gae_c     = as.factor(str_c(gndr_c, age_c, edulvlb_c)),
    pspwght_p = pspwght * pweight * 10000,
    dweight_p = dweight * pweight * 10000,
    trstplt_c =
      case_when((trstplt >= 0   & trstplt <= 3)  ~ "low",
                (trstplt >= 4   & trstplt <= 7)  ~ "medium",
                (trstplt >= 8   & trstplt <= 10) ~ "high",
                TRUE ~ as.character(trstplt) ),
    stfeco_c =
      case_when((stfeco >= 0   & stfeco <= 3)  ~ "low",
                (stfeco >= 4   & stfeco <= 7)  ~ "medium",
                (stfeco >= 8   & stfeco <= 10) ~ "high",
                TRUE ~ as.character(stfeco) )
  ) %>%
  left_join(sddf_es %>% select(idno, psu, stratify), by="idno") %>%
  mutate(psu = as.factor(psu),
         stratify = as.factor(stratify) )   ->
  ess_es
```

We use again data from ESS Round 6 for Spain and Denmark. Again you can executed the above R-Script, which will provide you with a data frame for Spain `ess_es` and data frame for Denmark `ess_dk`. Use these data sets to solve the estimation tasks of the assignment. Each estimation in the assignment is to be done using a complete case analysis to deal with item nonresponse. That is, elements with item non-response on a relevant item should be ignored in the estimation.

a (10 Points) What is the problem with just using symmetric confidence intervals (CIs) based on a normal approximation (i.e. the central limit theorem), e.g. $\hat{\rho} \pm \sqrt{\hat{\mathrm{V}}(\hat{\rho})} * 1.96$, if we want to construct confidence intervals for proportions? And for which kind of proportions is this problem particularly relevant?

b (20 Points) Denmark uses a Simple Random Sample, and Spain uses a two-stage sampling design with stratification at the level of the PSUs. The ESS does not provide any information on the sizes of PSU strata, thus sampling with replacement of PSUs can be assumed. The PSU identification variable is named `psu`, the variable identifying the PSU strata is named `stratify`, and the person identification variable is named `idno`, i.e. `idno` is the id of the ultimate sampling unit.

- We are interested in estimating the proportions for certain combinations of categories of `stfeco` (`How satisfied with present state of economy in country`) and `stflife` (`How satisfied with life as a whole`). Using the design weights `dweight_p` report your estimate of the following proportion:
  - **i)** The faction of persons in **Spain** that report `stfeco == 8` **and** `stflife == 5`.
  - **ii)** The faction of persons in **Denmark** that report `stfeco == 10` **and** `stflife == 0`.
- Report symmetric CIs with a confidence level of 95% for your estimated proportions in i) and ii), using a normal approximation. What problems do you see with these CIs?
- Now report Clopper-Person type CIs for complex survey for your estimated proportions in i) and ii). How do they differ from the CIs reported before? [Hint: You can extract CIs from objects returned by estimation functions of the `survey` package using `confint`]

c (20 Points) Our variables of interest are now `stfeco_c` and `trstplt_c`, which are aggregated version of `stfeco` and `trstplt` (`Trust in politicians`) respectively.

- Use the design weights `dweight_p` to estimate the joint distribution of categorical variables `stfeco_c` and `trstplt_c` for both Spain and Denmark. Report the proportions of all categories of `stfeco_c` given that `trstplt_c == low` for both Spain and Denmark data.
- Using a χ^2-test, test the following null hypothesis for Denmark and Spain: There is no association between variables `stfeco_c` and `trstplt_c`. Report your test decision, including the value of the test statistic and p-value. Name the kind of correction you applied to the χ^2-test to account for the sampling design.
- Repeat your χ^2-test for Spain, but this time use SE estimates based on a standard bootstrap with 99 resamples [Hint: set `type="bootstrap"` in `as.svrepdesign`]. Report your test decision, including the value of the test statistic and p-value.

d (20 Points) Repeat the test of the null hypotheses that for Spain and Denmark there is no association between variables `stfeco_c` and `trstplt_c`. But this time base you test decision on the comparison of log-linear models. Report your test decision, including a p-value.

e (BONUS 15 Points) Using a χ^2-test, test the following null hypothesis: The proportions of `stfeco_c` in Spain follow the same distribution as in Denmark. Report your test decision, including the value of the test statistic and p-value.

f (30 Points) Our variable of interest is `stfeco`. We treat it not as a categorical but as a **metric variable**. We are interested in the mean of `stfeco` by age, gender, and education categories. Our population of interest is the population of Spain.

i) Estimate for Spain the means of `stfeco`, treating the variable as metric, given each category of variable `gae_c` (crossing of gender, age classification, and education classification). Use for this estimation calibration weights that adjust design weights `dweight_p` to the population totals of the `gae_c` variable in Spain. Report your estimates together with SE estimates. [Hint: Remember `gae_c` is used in the ESS to compute its so-called post-stratification weights `pspwght_p`.]

ii) Repeat the before mentioned estimation, but this time use calibration weights that adjust design weights `dweight_p` not to the population totals of `gae_c` in Spain but the population totals of variable `stfgov` (How satisfied with the national government) within the categories of `gae_c` and the overall population size of Spain. Because these subtotals of `stfgov` are not available to you estimate them from the sample data using the ESS *post-stratification* weights. You can use the following R code to do this:

```r
# 1. Treat the item non-response;
#  We use a simple median imputation (not a general recommendation!)
# "svy_ps_es" is the survey design object
#   for spain with ess post-stratification weights.
# "svy_d_es" is the survey design object
#   for spain with design weights weights.
svyd_ps_es$variables$stfgov[
  is.na(svyd_ps_es$variables$stfgov)] <-
  median(svyd_ps_es$variables$stfgov,na.rm = T)
svyd_d_es$variables$stfgov[
  is.na(svyd_d_es$variables$stfgov)]  <-
  median(svyd_d_es$variables$stfgov,na.rm = T)

# 2. Calculate the totals of "stfgov" within "gae_c" categories.
pop_StfgovByGAE  <-
  coef(svyby(~ stfgov,
             by=~gae_c,
             svyd_ps_es,
             svytotal,
             na.rm = T))
```

```r
names(pop_StfgovByGAE) <-
  paste0("stfgov:gae_c",
         names(pop_StfgovByGAE))

# The interaction effect between
#  "stfgov" and "gae_c" is the  calibration variable.
# Check how the population total vector should look like:
cal_names(~ stfgov:gae_c, svyd_d_es)

# you can supply ~ stfgov:gae_c to the
#  formula argument of "calibrate"
#  to specifiy the calibraiton variable
```

Report your estimates together with SE estimates.

iii) (10 Bonus) Compare the SE estimates between your results in i) and ii) and interpret the differences. What do you think of calibrating the weights of a sample survey on totals that have been estimated from the same sample survey?

g (BONUS 30 Points) Our population of interest is again that of Spain. We are interested in the following statistics:

$$\mu_{\text{stfeco}_e} = \frac{\sum_{g\in\{1,2\}} \sum_{a\in\{[15,35],\,(35,55],\,(55,Inf]\}} \mu_{\text{stfeco}_{g\ a\ e}}}{6} \qquad \text{for } e = \text{high},\ \text{medium},\ \text{low}.$$

where $\mu_{\text{stfeco}_{g\ a\ e}}$ is the mean of `stfeco` for persons within the corssing of the g-th gender, the a-th age, and the e-th eduction cateogie of variable `gae_c`.

Use for your estimation calibration weights that adjust design weights `dweight_p` to the population totals of variables `gae_c` and `region_c` in Spain. Report your estimates for the three statistics together with an estimate for the covariance matrix of the estimators you used. [Hint: 1.) If you want `svycontrast()` to estimate a covariance matrix (`covmat=TRUE`) in combination with calibration weights it is better to use a resampling approach. 2.) You can use the `calibrate` function to calibrate the replication weights of a survey design object.]