Basic population estimates with British Social Attitudes Survey data using R

UK Data Service

June 2024

This exercise is part of the 'Introduction to the British Social Attitudes Survey (BSA)' online module. In the exercise, we examine data from the 2020 British Social Attitudes survey to find out:

- what proportion of respondents said they voted remain in the EU Referendum?
- whether people think the government should raise taxes and spend more or reduce tax and cut social expenditures?
- how much people think they'll get from the State pension?

Answers to the questions asked throughout the exercise can be found at the end of the page.

Getting started

Data can be downloaded from the UK Data Service website following registration. Download the compressed folder, unzip and save it somewhere accessible on your computer.

The examples below assume that the dataset has been saved in a new folder named *UKDS* on your Desktop (Windows computers). The path would typically be `C:\Users\YOUR_USER_NAME\Desktop\UKDS`. Feel free to change it to the location that best suits your needs.

We begin by loading the R packages needed for the exercise and set the working directory.

```
library(dplyr) ### Data manipulation functions
library(haven) ### Functions for importing data from
               ### commercial packages
library(Hmisc) ### Extra statistical functions

### Setting up the working directory
```

```
### Please adjust the setwd() command  below
### to match the location of the data on your computer

setwd("C:\Users\Your_Username_here\")

getwd()
```

```
[1] C:\Users\Your_Username_here\
```

We then open the BSA dataset in SPSS format. Stata or tab-delimited format can also be used.

```
bsa20<-read_spss(
       'UKDA-9005-spss/spss/spss25/bsa2020_archive.sav'
       )
```

1. Explore the dataset

Start by getting an overall feel for the data. Use the code below to produce a summary of all the variables in the dataset.

```
### Gives the number of rows (observations)
### and columns (variables)
dim(bsa20)
```

```
[1] 3964  210
```

```
 ### List variable names in their actual
 ### order in the dataset
names(bsa20)
```

```
  [1] "serial"       "QnrVersion"   "RespSx2cat"   "RespAgeE"     "MarStat6"
  [6] "REconFW01"    "REconFW02"    "REconFW03"    "REconFW04"    "REconFW05"
 [11] "REconFW06"    "REconFW07"    "REconFW08"    "REconFW09"    "REconFW10"
 [16] "REconFW11"    "EMPSTAT"      "Employ"       "Superv"       "EmpOCC"
 [21] "TenureE"      "SupParty"     "ClosePty"     "PARTYFW"      "Idstrng"
 [26] "RemLea"       "RemLeaCl"     "RemLeaSt"     "Politics"     "ConLabDf"
 [31] "VoteDuty"     "SocTrust"     "EngParl"      "ScotPar2"     "ECPolicy2"
 [36] "Spend1"       "Spend2"       "SocBen1"      "SOCBEN2"      "DOLE"
 [41] "TAXSPEND"     "WkMent"       "WkPhys"       "HProbRsp"     "PhsRetn"
```

```
 [46] "PhsRecov"     "MntRetn"      "MntRecov"     "HCWork21"     "HCWork22"
 [51] "HCWork23"     "HCWork24"     "HCWork25"     "HCWork26"     "HCWork28"
 [56] "HCWork29"     "HCWork213"    "HCWork214"    "HCWork215"    "HCWork27"
 [61] "CMtUnmar1"    "CMtUnmar2"    "CMtUnmar3"    "CMtUnmar4"    "CMtUnmar5"
 [66] "CMtUnmar6"    "CMtUnmar7"    "CMtUnmar8"    "CMtUnmar9"    "CMtUnmar10"
 [71] "CMtmar1"      "CMtmar2"      "CMtmar3"      "CMtmar4"      "CMtmar5"
 [76] "CMtmar6"      "CMtmar7"      "CMtmar8"      "CMtmar9"      "CMtmar10"
 [81] "ChCoSupp"     "ChMIncM"      "ChMIncF"      "ChMCont"      "RBGaran2"
 [86] "RBGGov"       "DigPCUn"      "DigPCctl"     "DigPCcon"     "DigPCrsk"
 [91] "DigGVun"      "DigGVctl"     "DigGVcon"     "DigGVrsk"     "DigPro"
 [96] "NHSSat"       "WkHmNow"      "WkHmJan"      "CovWkc"       "CovNoWkc"
[101] "CovWkr1"      "CovWkr2"      "CovWkr3"      "CovWkr4"      "CovWkr5"
[106] "CovWkr6"      "CovWk1"       "CovWk2"       "CovWk3"       "GovtWork"
[111] "GovTrust"     "CLRTRUST"     "MPsTrust"     "LoseTch"      "VoteIntr"
[116] "PtyNMat2"     "PolPart01"    "PolPart02"    "PolPart03"    "PolPart04"
[121] "PolPart05"    "PolPart06"    "PolPart07"    "PolPart08"    "PolPart09"
[126] "PolPart10"    "PolPart11"    "REFHANG"      "RefSyst"      "UnempJob"
[131] "SocHelp"      "DoleFidl"     "WelfFeet"     "welfhelp"     "morewelf"
[136] "damlives"     "proudwlf"     "Redistrb"     "BigBusnN"     "Wealth"
[141] "RichLaw"      "Indust4"      "TradVals"     "StifSent"     "DeathApp"
[146] "Obey"         "WrongLaw"     "Censor"       "NatIdGB"      "ChAttend"
[151] "DisNew2"      "DisAct"       "HEdQual2"     "HhldEdu"      "EURefV2"
[156] "EUVOTWHO"     "EURefb"       "Voted"        "Vote"         "Anybn3"
[161] "HHincome"     "Maininc5"     "REarn"        "HIncDif4"     "RetExp"
[166] "RetExpb"      "FutrWrk"      "PenKnow2"     "PenExp2"      "PenComp"
[171] "PenIntr"      "INFORET3"     "WkPKnw"       "WKPSav"       "WkPSpn"
[176] "WPSvUs"       "WPSvWw"       "WPSvEas"      "PrPKnw"       "PrPSav"
[181] "PrPSpn"       "PrPSvUs"      "PrPSvWW"      "PrPSvEas"     "NCOutcome"
[186] "Ragecat"      "Ragecat20"    "DisActDV"     "leftrigh"     "libauth"
[191] "welfare2"     "libauth2"     "leftrig2"     "welfgrp"      "REconAct20"
[196] "REconSum20"   "RaceOri4"     "LegMarStE"    "HhlAdGpd"     "HhlChlGpd"
[201] "BestNatU2"    "RetirAg3"     "ReligSum20"   "RlFamSum20"   "EmplStatDV"
[206] "RClassGP"     "serialh"      "GOR"          "gor2"         "BSA20_wt_new"
```

```
### Displays the first five
### lines of a data frame

head(bsa20)
```

```
# A tibble: 6 x 210
  serial   QnrVersion RespSx2cat RespAgeE MarStat6 REconFW01 REconFW02 REconFW03
  <dbl+lb> <dbl+lbl>  <dbl+lbl>  <dbl+lb> <dbl+lb> <dbl+lbl> <dbl+lbl> <dbl+lbl>
```

```
1 3.21e9   1 [Versio~ 2 [Male]   70       5 [Divo~ 0 [No]    0 [No]    0 [No]
2 3.21e9   1 [Versio~ 2 [Male]   66       1 [Marr~ 0 [No]    0 [No]    0 [No]
3 3.21e9   1 [Versio~ 1 [Female] 64       1 [Marr~ 0 [No]    0 [No]    0 [No]
4 3.21e9   1 [Versio~ 2 [Male]   43       1 [Marr~ 0 [No]    0 [No]    1 [Yes]
5 3.21e9   1 [Versio~ 1 [Female] 38       1 [Marr~ 0 [No]    0 [No]    1 [Yes]
6 3.21e9   1 [Versio~ 2 [Male]   77       1 [Marr~ 0 [No]    0 [No]    0 [No]
# i 202 more variables: REconFW04 <dbl+lbl>, REconFW05 <dbl+lbl>,
#   REconFW06 <dbl+lbl>, REconFW07 <dbl+lbl>, REconFW08 <dbl+lbl>,
#   REconFW09 <dbl+lbl>, REconFW10 <dbl+lbl>, REconFW11 <dbl+lbl>,
#   EMPSTAT <dbl+lbl>, Employ <dbl+lbl>, Superv <dbl+lbl>, EmpOCC <dbl+lbl>,
#   TenureE <dbl+lbl>, SupParty <dbl+lbl>, ClosePty <dbl+lbl>,
#   PARTYFW <dbl+lbl>, Idstrng <dbl+lbl>, RemLea <dbl+lbl>, RemLeaCl <dbl+lbl>,
#   RemLeaSt <dbl+lbl>, Politics <dbl+lbl>, ConLabDf <dbl+lbl>, ...
```

The above output is summarised in a `haven`- imported dataframe format also known as a 'tibble'. For a really raw output we need to convert into a 'pure' data frame. Beware, the output might be very lengthy!

```
head(data.frame(bsa20))
```

```
     serial QnrVersion RespSx2cat RespAgeE MarStat6 REconFW01 REconFW02
1 3.211e+09          1          2       70        5         0         0
2 3.211e+09          1          2       66        1         0         0
3 3.211e+09          1          1       64        1         0         0
4 3.211e+09          1          2       43        1         0         0
5 3.211e+09          1          1       38        1         0         0
6 3.211e+09          1          2       77        1         0         0
  REconFW03 REconFW04 REconFW05 REconFW06 REconFW07 REconFW08 REconFW09
1         0         0         0         0         0         0         1
2         0         0         0         0         0         0         1
3         0         0         0         0         0         0         1
4         1         0         0         0         0         0         0
5         1         0         0         0         0         0         0
6         0         0         0         0         0         0         1
  REconFW10 REconFW11 EMPSTAT Employ Superv EmpOCC TenureE SupParty ClosePty
1         0         0       1      2      1      3      10        1       NA
2         0         0       1      2      1      1       1        1       NA
3         0         0       1      1      2      1       1        1       NA
4         0         0       1      3      1      3       1        2        2
5         0         0       1      3      2      2       1        2        2
6         0         0       3     NA     NA      1       9        1       NA
  PARTYFW Idstrng RemLea RemLeaCl RemLeaSt Politics ConLabDf VoteDuty SocTrust
```

```
1       1       2     NA       NA       NA        2       NA       NA        1
2       2       3     NA       NA       NA        3       NA       NA        1
3       2       3     NA       NA       NA        3       NA       NA        1
4       2       3     NA       NA       NA        2       NA       NA        2
5       1       3     NA       NA       NA        3       NA       NA        2
6       1       2     NA       NA       NA        2       NA       NA        2
  EngParl ScotPar2 ECPolicy2 Spend1 Spend2 SocBen1 SOCBEN2 DOLE TAXSPEND WkMent
1      NA       NA        NA      2      1       1       2    1        2      1
2      NA       NA        NA      1      3       2       5    1        2      2
3      NA       NA        NA      3      1       2       3    1        2      2
4      NA       NA        NA      7      3       1       2    2        2      2
5      NA       NA        NA      7      3       2       4    2        2      1
6      NA       NA        NA     98     NA       1       4    2        3      2
  WkPhys HProbRsp PhsRetn PhsRecov MntRetn MntRecov HCWork21 HCWork22 HCWork23
1      1        1       1        2       1        2        1        1        1
2      2        1       1        3       1        2        1        0        1
3      2        1       1        2       1        2        1        1        1
4      2        2       2        3       1        2        1        1        1
5      1        1       1        2       1        2        1        1        1
6      2        2       2        2       2        2        1        0        1
  HCWork24 HCWork25 HCWork26 HCWork28 HCWork29 HCWork213 HCWork214 HCWork215
1        1        1        1        0        0         0         0         0
2        1        1        1        0        0         0         0         0
3        1        1        1        0        0         0         0         0
4        1        1        1        0        0         0         0         0
5        1        1        1        0        0         0         0         0
6        1        1        0        0        0         0         0         0
  HCWork27 CMtUnmar1 CMtUnmar2 CMtUnmar3 CMtUnmar4 CMtUnmar5 CMtUnmar6
1        0         1         2         2         1         1         1
2        0         1         1         1         3         3         1
3        0         1         1         1         3         3         1
4        0        NA        NA        NA        NA        NA        NA
5        0        NA        NA        NA        NA        NA        NA
6        0         1         1         1         3         1         8
  CMtUnmar7 CMtUnmar8 CMtUnmar9 CMtUnmar10 CMtmar1 CMtmar2 CMtmar3 CMtmar4
1         1         2         1          1      NA      NA      NA      NA
2         1         1         3          1      NA      NA      NA      NA
3         1         1         3          3      NA      NA      NA      NA
4        NA        NA        NA         NA       1       1       2       1
5        NA        NA        NA         NA       1       1       1       1
6         1         1         3          1      NA      NA      NA      NA
  CMtmar5 CMtmar6 CMtmar7 CMtmar8 CMtmar9 CMtmar10 ChCoSupp ChMIncM ChMIncF
1      NA      NA      NA      NA      NA       NA        3       1      NA
```

```
2      NA      NA      NA      NA      NA       NA        3       2      NA
3      NA      NA      NA      NA      NA       NA        2       2      NA
4       1       1       1       2       1        1       NA      NA       1
5       1       1       1       2       1        1       NA      NA       1
6      NA      NA      NA      NA      NA       NA        3       8      NA
  ChMCont RBGaran2 RBGGov DigPCUn DigPCctl DigPCcon DigPCrsk DigGVun DigGVctl
1       1        2     NA       2        2        2        1      NA       NA
2       4        2     NA       2        3        3        1      NA       NA
3       2        3     NA       3        3        3        8      NA       NA
4      NA       NA     NA      NA       NA       NA       NA       1        2
5      NA       NA     NA      NA       NA       NA       NA       3        3
6       1        1      1       1        3        1        2      NA       NA
  DigGVcon DigGVrsk DigPro NHSSat WkHmNow WkHmJan CovWkc CovNoWkc CovWkr1
1       NA       NA      2      3      NA      NA     NA       NA      NA
2       NA       NA      2      2      NA      NA     NA       NA      NA
3       NA       NA      2      3      NA      NA     NA       NA      NA
4        4        1      2      2       1       2     NA        1       0
5        3        8      1      2       3       3      1       NA       0
6       NA       NA      2      2      NA      NA     NA       NA      NA
  CovWkr2 CovWkr3 CovWkr4 CovWkr5 CovWkr6 CovWk1 CovWk2 CovWk3 GovtWork
1      NA      NA      NA      NA      NA     NA     NA     NA       NA
2      NA      NA      NA      NA      NA     NA     NA     NA       NA
3      NA      NA      NA      NA      NA     NA     NA     NA       NA
4       0       0       0       1       0      5      5      5       NA
5       0       0       0       0       1      3      3      3       NA
6      NA      NA      NA      NA      NA     NA     NA     NA       NA
  GovTrust CLRTRUST MPsTrust LoseTch VoteIntr PtyNMat2 PolPart01 PolPart02
1       NA       NA       NA      NA       NA       NA        NA        NA
2       NA       NA       NA      NA       NA       NA        NA        NA
3       NA       NA       NA      NA       NA       NA        NA        NA
4       NA       NA       NA      NA       NA       NA        NA        NA
5       NA       NA       NA      NA       NA       NA        NA        NA
6       NA       NA       NA      NA       NA       NA        NA        NA
  PolPart03 PolPart04 PolPart05 PolPart06 PolPart07 PolPart08 PolPart09
1        NA        NA        NA        NA        NA        NA        NA
2        NA        NA        NA        NA        NA        NA        NA
3        NA        NA        NA        NA        NA        NA        NA
4        NA        NA        NA        NA        NA        NA        NA
5        NA        NA        NA        NA        NA        NA        NA
6        NA        NA        NA        NA        NA        NA        NA
  PolPart10 PolPart11 REFHANG RefSyst UnempJob SocHelp DoleFidl WelfFeet
1        NA        NA      NA      NA        3       4        4        4
2        NA        NA      NA      NA        3       3        3        4
```

```
3        NA        NA      NA      NA        3       4        4        4
4        NA        NA      NA      NA        2       3        3        1
5        NA        NA      NA      NA        2       4        2        3
6        NA        NA      NA      NA        2       2        2        2
  welfhelp morewelf damlives proudwlf Redistrb BigBusnN Wealth RichLaw Indust4
1        4        2        2        1        3        4      3       5       4
2        4        3        1        2        4        3      3       4       4
3        3        3        1        1        3        3      2       3       3
4        2        4        3        3        4        2      2       2       3
5        3        3        3        2        4        2      3       3       4
6        3        3        4        2        4        4      3       5       4
  TradVals StifSent DeathApp Obey WrongLaw Censor NatIdGB ChAttend DisNew2
1        3        3        2    3        4      3       5        7       2
2        4        3        2    2        3      2       6       NA       2
3        3        3        3    2        2      2       1       NA       2
4        2        1        2    1        2      2       3        7       2
5        4        3        3    3        4      2       3       NA       2
6        1        2        3    1        3      2       3        1       2
  DisAct HEdQual2 HhldEdu EURefV2 EUVOTWHO EURefb Voted Vote Anybn3 HHincome
1     NA        2       2      NA       NA     NA     2   NA      1        2
2     NA        1      NA      NA       NA     NA     1    2      2        3
3     NA        2       1      NA       NA     NA     1    2      2        3
4     NA        4       2      NA       NA     NA     1    1      1        4
5     NA        3       2      NA       NA     NA     1    1      1        3
6     NA        1      NA      NA       NA     NA     1    1      1        9
  Maininc5 REarn HIncDif4 RetExp RetExpb FutrWrk PenKnow2 PenExp2 PenComp
1        4    NA        3     NA      NA      NA       NA      NA      NA
2        2    NA        2     NA      NA      NA       NA      NA      NA
3        2    NA        2     NA      NA      NA       NA      NA      NA
4        1     3        2      3      60       2        1    7000       4
5        1     3        3      3      65       1        2     130       2
6        1    NA        3     NA      NA      NA       NA      NA      NA
  PenIntr INFORET3 WkPKnw WKPSav WkPSpn WPSvUs WPSvWw WPSvEas PrPKnw PrPSav
1      NA       NA     NA     NA     NA     NA     NA      NA     NA     NA
2      NA       NA     NA     NA     NA     NA     NA      NA     NA     NA
3      NA       NA     NA     NA     NA     NA     NA      NA     NA     NA
4       2        2      2      1      4      1      1       1     NA     NA
5       2        2      3      1      4      1      2       2     NA     NA
6      NA       NA     NA     NA     NA     NA     NA      NA     NA     NA
  PrPSpn PrPSvUs PrPSvWW PrPSvEas NCOutcome Ragecat Ragecat20 DisActDV leftrigh
1     NA      NA      NA       NA         1       7         6        3      3.8
2     NA      NA      NA       NA         1       7         6        3      3.6
3     NA      NA      NA       NA         1       6         5        3      2.8
```

```
4     NA       NA       NA        NA          1        3          3         3       2.6
5     NA       NA       NA        NA          1        3          3         3       3.2
6     NA       NA       NA        NA          1        7          7         3       4.0
   libauth welfare2 libauth2 leftrig2 welfgrp REconAct20 REconSum20 RaceOri4
1 3.000000    2.000        2        3       1          9          6        3
2 3.333333    2.375        2        3       1          9          6        3
3 3.500000    2.125        2        2       1          9          6        3
4 4.333333    3.625        3        2       3          3          2        3
5 2.833333    3.000        2        2       2          3          2        3
6 4.000000    3.500        3        3       2          9          6        3
  LegMarStE HhlAdGpd HhlChlGpd BestNatU2 RetirAg3 ReligSum20 RlFamSum20
1         4        1         0         1       65          3          1
2         1        2         0         3       58          5          2
3         1        2         0         1       54          5          1
4         1        2         1         1       NA          3          2
5         1        2         1         2       NA          5          3
6         1        2         0         2       99          3          3
  EmplStatDV RClassGP   serialh GOR gor2 BSA20_wt_new
1          4        1 321100002   1    1    0.7099859
2          6        1 321100014   1    1    0.3145871
3          7        1 321100014   1    1    0.5649618
4          4        1 321100040   1    1    0.9355446
5          7        2 321100040   1    1    0.6830794
6          3        1 321100042   1    1    1.4006989
```

Questions

1. What is the overall sample size?

2. How many variables are there in the dataset?

Now, focus on the three variables we will use.

Note Traditional statistical software such as SPSS or Stata treat categorical variables as arbitrary numbers. Values labels are then attached, that allocate a substantive meaning to these values. R on the other hand can either directly deal with the value themselves as alphanumeric variables, or with its own version of categorical variables, known as 'factors'. There aren't straightforward ways to convert SPSS or Stata labelled categorical variables into R factors.

The `haven` package that we use here preserves the original numeric values in the data, and add attributes that can be manipulated separately and contain the labels. Attributes are a special type of R objects that have a name, and can be read using the `attr()` function. Each variable has a 'label' and 'labels' attribute. The former is the variable description, the latter the value labels.

Alternatively, haven-imported numeric variables can be converted into factors with levels (ie categories) reflecting the SPSS or Stata value labels, but with numeric values different from the original ones.

Let's examine the original variable description and value labels with the `attr()` function. We can do this variable by variable…

```
attr(bsa20$TAXSPEND,"label")
```

```
[1] "If it had to choose, should govt reduce/increase/maintain levels of taxation and spendi
```

… Or all at once:

```
t(                          # Transpose rows and columns for better readability
  bsa20 |>
        select(TAXSPEND,EUVOTWHO,PenExp2) |> # Select the relevant variables
        summarise_all(attr,"label")  # Apply the attr() function to all of them
)
```

```
         [,1]
TAXSPEND "If it had to choose, should govt reduce/increase/maintain levels of taxation and sp
EUVOTWHO "Did you vote to 'remain a member of the EU' or to 'leave the EU'?"
PenExp2  "How much do you think someone who reaches State Pension age today would receive in
```

We do the same with value labels:

```
attr(bsa20$TAXSPEND,"labels")
```

```
                                                                 Not applicable
                                                                             -1
  Reduce taxes and spend less on health, education and social benefits
                                                                              1
    Keep taxes and spending on these services at the same level as now
                                                                              2
Increase taxes and spend more on health, education and social benefits
                                                                              3
                                                                     Don't know
                                                                              8
                                                            Prefer not to answer
                                                                              9
```

```
attr(bsa20$EUVOTWHO,"labels")
```

```
                     Not applicable Remain a member of the European Union
                                 -1                                      1
           Leave the European Union                      I Don't remember
                                  2                                      3
                         Don't know                   Prefer not to answer
                                  8                                      9
```

Question 3

What do the variables measure and how?

2. Missing values

Let's now examine the distribution of our three variables. We can temporarily convert `EUVOTWHO` and `TAXSPEND` into factors using `mutate()` for a more meaningful output that include their value labels. Review the frequency tables, examining the 'not applicable' and 'don't know' categories.

```
bsa20%>%select(EUVOTWHO,TAXSPEND) %>%
       mutate(as_factor(.)) %>%
       summary()
```

```
                                 EUVOTWHO
 Not applicable                      :   0
 Remain a member of the European Union: 635
 Leave the European Union            : 463
 I Don't remember                    :   2
 Don't know                          :   0
 Prefer not to answer                :  21
 NA's                                :2843
                                                                    TAXSPEND
 Not applicable                                                        :   0
 Reduce taxes and spend less on health, education and social benefits  : 186
 Keep taxes and spending on these services at the same level as now    :1589
 Increase taxes and spend more on health, education and social benefits:2133
 Don't know                                                            :  35
 Prefer not to answer                                                  :  21
```

```
summary(bsa20$PenExp2)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0     120     160    1293     200    9999    1076
```

Question 4

Why are there so many system missing values (NA) for `EUVOTWHO` and `PenExp2` ? What does this mean when it comes to interpreting the percentages? You can use the documentation if needed.

When analysing survey data, it is sometimes convenient to recode item nonresponses such as ´Don't know´ and 'Prefer not to say' as system missing so that they do not appear in the results. An example of the syntax required to achieve this with EUVOTWHO and TAXSPEND is provided in the appendix.

Unlike some other surveys, 'Don't knows' and 'Does not apply' were not removed when weights were computed in the BSA. As a result, analyses using weights (ie when planning to use the data to make inference about the British population) need to retain these observations, otherwise estimated results might be incorrect.

3. Compare unweighted and weighted proportions

In this section, we compare unweighted and weighted prportionso for `EUVOTWHO` and `TAXSPEND`. Let's examine the unweighted responses first. In order to ensure coherence with the remainder of this exercise, we use `xtabs()` for categorical variables and `summary()` for continuous ones.

First, as mentioned above, we recode `EUVOTWHO` and `TAXSPEND`into factors, with value labels as levels using `as_factor()`

```
bsa20<-bsa20%>%mutate(
               TAXSPEND.f=as_factor(TAXSPEND,"labels"),
               EUVOTWHO.f=as_factor(EUVOTWHO,"labels")
                )
```

We can truncate factor levels respectively to 14 and 6 characters, for a more human-friendly output using `substr()`:

```
levels(bsa20$TAXSPEND.f)<-substr(levels(bsa20$TAXSPEND.f),1,14)
levels(bsa20$EUVOTWHO.f)<-substr(levels(bsa20$EUVOTWHO.f),1,6)
```

Finally, we compute the proportions:

```
round(                                  ### Rounds the results to one decimal
  100*                                  ### Converts proportions to %
    prop.table(                         ### Computes proportions
      xtabs(~TAXSPEND.f,bsa20,          ### Computes frequencies,
             drop.unused.levels = T)    ### Leaves out levels with 0 observations),
      ),
  1)
```

```
TAXSPEND.f
Reduce taxes a Keep taxes and Increase taxes     Don't know Prefer not to
           4.7           40.1           53.8            0.9            0.5
```

```
round(100*prop.table(xtabs(~EUVOTWHO.f,bsa20,drop.unused.levels = T)),1)
```

```
EUVOTWHO.f
Remain Leave  I Don' Prefer
  56.6   41.3    0.2    1.9
```

We can also examine the basic summary statistics for `PenExp2`:

```
summary(bsa20$PenExp2)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0     120     160    1293     200    9999    1076
```

What is the (unweighted) percentage of respondents who say they voted remain in the EU referendum? About 57 percent of sample members who voted in referendum said they voted to remain. This figure seems a bit high (though people do not always report accurately).

Let's compare with the weighted frequencies. We will keep using `xtabs()` for convenience. With `xtabs()`, weights are specified on the left hand side of the formula as shown below. For the record, `wtd.table()` function from the `Hmisc` package also produces weighted frequency tables.

```
xtabs(BSA20_wt_new~EUVOTWHO.f,
      data=bsa20)
```

```
EUVOTWHO.f
    Not ap     Remain     Leave      I Don'     Don't      Prefer
  0.000000 565.011079 489.146642   3.752765   0.000000  22.527320
```

We can get rid of the empty levels to improve the output:

```
xtabs(BSA20_wt_new~EUVOTWHO.f,
      data=bsa20,
      drop.unused.levels = T)
```

```
EUVOTWHO.f
    Remain     Leave      I Don'     Prefer
565.011079 489.146642   3.752765  22.527320
```

We convert the weighted frequencies into proportions and examine the results:

```
euv.wp<-round(
  100*
    prop.table(
      xtabs(BSA20_wt_new~EUVOTWHO.f,
            data=bsa20,
            drop.unused.levels = T)
      ),
  1)

euv.wp
```

```
EUVOTWHO.f
Remain Leave  I Don' Prefer
  52.3   45.3    0.3    2.1
```

Now, what proportion say they voted remain in the EU referendum?

It is about 52 percent, lower than the unweighted proportion and closer to the actual referendum results.

Do you have an idea as to why this might be the case?

A possible explanation is that those more likely to vote ‘Remain’, such as younger people tend to also be less likely to take part in surveys, and therefore their real prevalence in the population will be underestimated by unweighted proportions.

4. Confidence intervals

So far, we have just computed point estimates without worrying about their precision. Estimates precision (or uncertainty) does matter insofar as it determines how big the ranges within which 'true' population values are likely to be. These are also known as the *confidence intervals* of our estimates.

In this exercise, we will be computing confidence intervals 'by hand' and ignore the survey design (ie whether clustering or stratification were used when collecting the sample) as the information is not available in this edition of the BSA. This amounts to assuming that the sample was collected using simple random sampling - which wasn't the case - and increase the bias of our estimates.

We will explore the more reliable survey design functions provided by the `survey` package in the next exercise.

Confidence intervals for proportions

The `Hmisc` package provides `binconf()` a handy function to compute confidence intervals for proportions. We need to provide it with two parameters: the frequencies for which we would like a confidence interval, and the total number of non missing observations. `binconf()` accepts individual proportions or complete frequency tables as input.

We begin with the unweighted confidence interval for EUVOTWHO:

```
eu.ci<-binconf(xtabs(~EUVOTWHO.f,
                     bsa20,
                     drop.unused.levels = T)[1],
               sum(xtabs(~EUVOTWHO.f,bsa20)))

eu.ci
```

```
  PointEst     Lower     Upper
 0.5664585 0.5372704 0.5951927
```

We convert the output into rounded percentages for better readability:

```
round(100*
      eu.ci,
      1)
```

```
 PointEst Lower Upper
     56.6  53.7  59.5
```

We can adapt the syntax above to make it work with weighted frequencies:

```
round(100*
  binconf(xtabs(bsa20$BSA20_wt_new~EUVOTWHO.f,
                data=bsa20,
                drop.unused.levels = T)[2],
          sum(xtabs(bsa20$BSA20_wt_new~EUVOTWHO.f,
                    data=bsa20,
                    drop.unused.levels = T))),
      1)
```

```
 PointEst Lower Upper
     45.3  42.3  48.3
```

What are the differences between weighted and unweighted confidence intervals for the proportion of people who voted remain?

Let us now do the same with people's views about government tax and spending.

```
ciprop<-
round(100*
binconf(xtabs(BSA20_wt_new~TAXSPEND.f,
              data=bsa20,
              drop.unused.levels=T),
        sum(xtabs(BSA20_wt_new~TAXSPEND.f,
                  bsa20))),
1)

ciprop
```

```
 PointEst Lower Upper
      5.5   4.8   6.3
     42.8  41.3  44.3
     50.3  48.8  51.9
      0.9   0.6   1.2
      0.5   0.3   0.8
```

We can improve the layout by adding the value labels. In order to do this, we create a data frame with the results of the above computation `ciprop` and specify that the row names should be the original value labels of TAXSPEND using as_factor. We also however need to omit the first label 'Not applicable' as we removed it earlier.

```
ciprop.l<-data.frame(
           ciprop,
           row.names=levels(
                     bsa20$TAXSPEND.f
                     )[-1]
           )

ciprop.l
```

```
               PointEst Lower Upper
Reduce taxes a      5.5   4.8   6.3
Keep taxes and     42.8  41.3  44.3
Increase taxes     50.3  48.8  51.9
Don't know          0.9   0.6   1.2
Prefer not to       0.5   0.3   0.8
```

Question 5.

What proportion think government should increase taxes and spend more on health, education and social benefits?

Confidence intervals for means

Several R packages offer functions for computing confidence intervals and standard errors of means. Here, we privilege doing things by hand in order to properly understand what is happening in the background.

Under assumptions of simple random sampling, a 95% confidence interval of the mean is defined as plus or minus 1.96 times its standard error. The standard error of the mean is its standard deviation – that is, the square root of its variance – divided by the square root of the sample size.

We will be using `wtd.mean` from the `Hmisc` package to compute weighted means, and `wtd.var` for variances. We can therefore compute:

```
m.p<-wtd.mean(bsa20$PenExp2,weights=bsa20$BSA20_wt_new)
se.p<-sqrt(wtd.var(bsa20$PenExp2,weights=bsa20$BSA20_wt_new))
n<-sum(bsa20$BSA20_wt_new[!is.na(bsa20$PenExp2)])

ci<-c(m.p,m.p-1.96*(se.p/sqrt(n)),m.p+1.96*(se.p/sqrt(n)))

round(ci,1)
```

```
[1] 1305.9 1194.4 1417.5
```

Question 6

How much do people think they will get at state pension age?

Answers

1. There are 3964 cases in the dataset.
2. The total number of variables is 212.
3. *TAXSPEND records responses to the questions of whether government should reduce/increase/maintain levels of taxation and spending. There are three possible responses to the question.* EUVOTWHO records responses to the question 'Did you vote to 'remain a member of the EU' or to 'leave the EU'?' The responses are 'Remain' or 'Leave'. *PenExp2 contains responses to the question 'How much do you think someone who reaches State Pension age today would receive in pounds per week?' Responses are numeric.
4. There are two reasons for the many 'Not applicable'.
   - Routing: the question is only asked to those who said yes to a previous question (EURefV2).
   - Versions 5 and 6 - The BSA uses a split sample and the question is only asked in Versions 5 and 6.
5. Between 48.8 and 51.9% in the population say the government should increase taxes and spend more.
6. The amount people think they will get at state pension age varies between £1194 and £1417, with an average (ie mean) in the region of £1306.

Appendix: recoding nonresponses as system missing (NA)

The code below provides and example of how to recode missing values into system missing (NA) using separate variables. For ease of interpretation, we also convert the original numeric variable into labelled factors using `as_factor()`, so that they directly display the value labels.

```
bsa20<-bsa20%>%mutate(
              TAXSPEND.r=factor(as_factor(TAXSPEND,"labels"),
                                exclude = c("Prefer not to answer",
                                            "Don't know")),
              EUVOTWHO.r=factor(as_factor(EUVOTWHO,"labels"),
                                exclude = c("Prefer not to answer",
                                            "I Don't remember",
                                            "Not applicable",NA)),
              PenExp2.r=ifelse(PenExp2==-1 | PenExp2>=9998,NA,PenExp2)
                      )
### Value labels need to be truncated as they are rather lengthy!
levels(bsa20$TAXSPEND.r)<-substr(levels(bsa20$TAXSPEND.r),1,14)
levels(bsa20$EUVOTWHO.r)<-substr(levels(bsa20$EUVOTWHO.r),1,6)

levels(bsa20$TAXSPEND.r)
```

```
[1] "Reduce taxes a" "Keep taxes and" "Increase taxes"
```

```
levels(bsa20$EUVOTWHO.r)
```

```
[1] "Remain" "Leave "
```