Introduction to the British Social Attitudes Survey

R and SPSS Practical Exercises

UK Data Service

UK Data Service

Table of contents

1. Introduction

This repository contains the Quarto source files of the practical exercises for the *Introduction to Social Attitudes Surveys* UKDS Data Skills Module.

A key aim of the BSA is to track the views and opinions of the public on national issues over time. The BSA questionnaire has core questions that are repeated in most years. These cover different topics such as politics, welfare, poverty, health, education, equalities, and employment. In addition, the interview questionnaire consists of various background and demographic questions. The rest of the questionnaire includes a series of non-core questions (modules) on a number of social, political, economic and moral topics, which are included in the survey less frequently.

Direct links to the HTML pages of the exercises on GitHub Pages:

- R version:
  - Basic population estimates with BSA data using R
  - Survey design informed inference with BSA data using R
- SPSS version:
  - Basic population estimates with BSA data using SPSS
  - Survey design informed inference with BSA data using SPSS

Part I.

R Exercises

2. Using R with weights and survey design variables

This exercise is part of the 'Introduction to the British Social Attitudes Survey (BSA)' online module. In this exercise, we will practice statistical inference with data from the British Social Attitudes Survey (BSA) 2017 using weights and survey design variables.

Please note that at the time of writing this document only some of the BSA editions include survey design variables. For more information about inference from social surveys, including cases where weights and/or survey design variables are not available, please consult our guidelines.

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

2.1. 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

The code below will need to be adjusted in order to match the location of the data on your computer.

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
library(survey) ### Survey design functions

### Setting up the working directory
### Change the setwd() command  to match the location of the data on your computer
### if required

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

getwd()

# Opening the BSA dataset in SPSS format
bsa17<-read_spss("data/UKDA-8450-spss/spss/spss25/bsa2017_for_ukda.sav")
```

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

2.2. 1. Identifying the survey design and variables

We first need to find out about the survey design that was used in the BSA 2017, and the design variables available in the dataset. Such information can usually be found in the documentation that comes together with the data under the `mrdoc/pdf` folder or in the data catalogue pages for the data on the UK Data Service website.

Question 1

What is the design that was used in this survey (i.e. how many sampling stages were there, and what were the units sampled). What were the primary sampling units; the strata (if relevant)?

Now that we are a bit more familiar with the way the survey was designed, we need to try and identify the design variables we can include when producing estimates. The information can usually be found in the data documentation or the data dictionary available in the BSA documentation.

Question 2

What survey design variables are available? Are there any that are missing – if so which ones? What is the name of the weights variables?

2.3. 2. Specifying the survey design

We need to tell R about the survey design. In practice this often means specifying the units selected at the initial sampling stage ie the *Primary Sampling Units*, as well as the strata. This is achieved with the `svydesign()` command. In effect this command creates a copy of the dataset with the survey design information attached, that can then subsequently be used for further estimation.

```
bsa17.s<-svydesign(ids=~Spoint,        ### Primary Sampling Units
                   strata=~StratID,    ### Strata if stratified design
                   weights=~WtFactor, ### Weights
                   data=bsa17)         ### The dataset
class(bsa17.s)
```

```
[1] "survey.design2" "survey.design"
```

```
summary(bsa17.s) ### Warning: very long output
```

```
Stratified 1 - level Cluster Sampling design (with replacement)
With (372) clusters.
svydesign(ids = ~Spoint, strata = ~StratID, weights = ~WtFactor,
    data = bsa17)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2645  0.8288  1.0983  1.2386  1.6236  3.3318
Stratum Sizes:
           101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117
obs         18  22  30  18  16  21  22  37  10  22  19  35  23  19  19  21  25
```

```
design.PSU   2   2   3   2   2   2   2   3   2   3   2   3   2   2   2   2   2
actual.PSU   2   2   3   2   2   2   2   3   2   3   2   3   2   2   2   2   2
           118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134
obs         12  12  32  40  25  21  23  26  23  18  34  23  20  29  39  19  30
design.PSU   2   2   3   3   3   2   2   2   3   2   2   2   2   3   3   2   3
actual.PSU   2   2   3   3   3   2   2   2   3   2   2   2   2   3   3   2   3
           135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151
obs         20  10  21  12  26  16  20  17  21  24  30  30  18  29  24  19  28
design.PSU   2   2   2   2   3   2   2   2   2   3   2   3   2   3   2   3   2
actual.PSU   2   2   2   2   3   2   2   2   2   3   2   3   2   3   2   3   2
           152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168
obs         18   8  23  33  14  23  17  39  13  22  16  19  21  18  26  13  14
design.PSU   2   2   2   3   2   2   2   3   2   2   2   2   2   2   3   2   2
actual.PSU   2   2   2   3   2   2   2   3   2   2   2   2   2   2   3   2   2
           169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185
obs         22  20   8  22  31  22  24  19  38  20  29  24  29  21  23  32  36
design.PSU   2   2   2   2   2   2   2   2   3   2   2   2   3   2   2   3   3
actual.PSU   2   2   2   2   2   2   2   2   3   2   2   2   3   2   2   3   3
           186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202
obs         24  22  43  38  38  47  34  15  22  35  17  20  20  21  21  43  35
design.PSU   3   2   3   3   3   3   3   2   2   3   2   2   2   2   3   3   3
actual.PSU   3   2   3   3   3   3   3   2   2   3   2   2   2   2   3   3   3
           203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219
obs         28  25  19  18  28  15  21  30  24  33  24  22  30  24  44  18  26
design.PSU   3   3   2   2   2   2   2   2   2   3   2   2   3   2   3   2   2
actual.PSU   3   3   2   2   2   2   2   2   2   3   2   2   3   2   3   2   2
           220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236
obs         22  28  20  27  34  33  41  24  23  26  17  23  36  20  45  32  27
design.PSU   2   2   2   3   2   3   3   2   2   2   2   2   3   2   3   3   3
actual.PSU   2   2   2   3   2   3   3   2   2   2   2   2   3   2   3   3   3
           237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253
obs         33  25  39  31  29  33  20  43  22  24  26  29  37  22  27  25  43
design.PSU   3   3   3   3   2   2   2   3   2   2   2   2   3   2   2   2   3
actual.PSU   3   3   3   3   2   2   2   3   2   2   2   2   3   2   2   2   3
           254 255 256 257 258 259
obs          7  32  26  25  28  35
design.PSU   2   3   2   2   2   3
actual.PSU   2   3   2   2   2   3
Data variables:
  [1] "Sserial"             "Spoint"              "StratID"
  [4] "WtFactor"            "OldWt"               "GOR_ID"
  [7] "ABCVer"              "Country"             "househlde"
 [10] "hhtypee"             "Rsex"                "RAgeE"
 [13] "RAgeCat"             "RAgeCat2"            "RAgecat3"
 [16] "RAgecat4"            "RAgecat5"            "RSexAge"
 [19] "RSexAge2"            "MarStat"             "Married"
 [22] "legmarste"           "ChildHh"             "nch415e"
 [25] "nch318e"             "hhch04e"             "hhch511e"
 [28] "hhch1215e"           "hhch1617e"           "rch04e"
 [31] "rch511e"             "rch1215e"            "rch1617e"
```

```
 [34] "ownche"          "reconacte"       "RLastJob"
 [37] "seconacte"       "Readpap"         "WhPaper"
 [40] "paptype"         "TVNews"          "WebNews"
 [43] "WNwSite1"        "WNwSite2"        "SMNews"
 [46] "Internet"        "IntPers"         "MedResI"
 [49] "SupParty"        "ClosePty"        "PartyIDN"
 [52] "Partyid1"        "PartyId2"        "PartyID3"
 [55] "PtyAlleg"        "Idstrng"         "Politics"
 [58] "Coalitin"        "ConLabDf"        "VoteSyst"
 [61] "ScotPar2"        "ECPolicy2"       "GovTrust"
 [64] "Monarchy"        "MiEcono"         "MiCultur"
 [67] "Spend1"          "Spend2"          "SocSpnd1"
 [70] "SocSpnd2"        "SocSpnd3"        "SocSpnd4"
 [73] "SocSpnd5"        "SocSpnd6"        "Dole"
 [76] "TaxSpend"        "IncomGap"        "SRInc"
 [79] "CMArran"         "RBGaran2"        "SepInvol"
 [82] "SepServ"         "WkMent"          "WkPhys"
 [85] "HProbRsp"        "PhsRetn"         "PhsRecov"
 [88] "MntRetn"         "MntRecov"        "HCWork21"
 [91] "HCWork22"        "HCWork23"        "HCWork24"
 [94] "HCWork25"        "HCWork26"        "HCWork27"
 [97] "HCWork28"        "HCWork29"        "NatFrEst"
[100] "FalseBn2"        "RepFrau3"        "RepWho1"
[103] "RepWho2"         "RepWho3"         "RepWho4"
[106] "RepWho5"         "RepWho6"         "RepWho7"
[109] "RepWho8"         "RepWho9"         "RepWho10"
[112] "WhyNRep1"        "WhyNRep2"        "WhyNRep3"
[115] "WhyNRep4"        "WhyNRep5"        "WhyNRep6"
[118] "WhyNRep7"        "WhyNRep8"        "WhyNRep9"
[121] "BFPnsh1"         "BFPnsh2"         "BFPnsh3"
[124] "BFPnsh4"         "BFPnsh5"         "BFPnsh6"
[127] "BFPnsh7"         "BFPnsh8"         "BFPnsh9"
[130] "BFPnsh10"        "BFPnsh11"        "AwrPB"
[133] "AdminPn2"        "LosofBen"        "AwrCRec"
[136] "GovDoBF"         "ImpHDoc"         "ImpHPar"
[139] "ImpHBeha"        "ImpHFam"         "ImpHEd"
[142] "ImpHJob"         "ImpHNeig"        "ImpHArea"
[145] "ImpHSafe"        "RespoHl2"        "HomsBult"
[148] "YSBEmpl"         "YSBTrans"        "YSBGreen"
[151] "YSBSch"          "YSBAfRnt"        "YSBAfOwn"
[154] "YSBDesig"        "YSBShops"        "YSBMedic"
[157] "YSBLibry"        "YSBLeis"         "YSBFinan"
[160] "YSBOther"        "YSBDeps"         "YSBNone"
[163] "HousGSD"         "Buldres"         "EdSpnd1c"
[166] "EdSpnd2c"        "VocVAcad"        "ATTD151"
[169] "ATTD152"         "ATTD153"         "ATTD154"
[172] "ATTD155"         "ATTD156"         "ATTD157"
[175] "ATTD158"         "ATTD81"          "ATTD82"
[178] "ATTD83"          "ATTD84"          "ATTD85"
[181] "ATTD86"          "ATTD87"          "ATTD88"
```

```
[184] "GCSEFur"     "GCSEWrk"     "ALevFur"
[187] "ALevWrk"     "HEdOpp"      "ChLikUn2"
[190] "HEFee"       "FeesUni"     "FeesSub"
[193] "Himp"        "PREVFR"      "TRFPB6U"
[196] "TRFPB9U"     "TrfPb10u"    "TrfConc1"
[199] "DRIVE"       "carnume"     "CycDang"
[202] "Bikeown2"    "BikeRid"     "TRAVEL1"
[205] "TRAVEL2"     "TRAVEL3"     "TRAVEL4a"
[208] "TRAVEL6"     "airtrvle"    "CCTrans1"
[211] "CCTrans2"    "CCTrans3"    "CCTrans4"
[214] "CCTrans5"    "CCTrans6"    "CCTrans7"
[217] "CCTrans8"    "CCTrans9"    "CCALowE"
[220] "CCACar"      "CCAPLANE"    "CCBELIEV"
[223] "EUBrld"      "EUExInf2"    "EUExUne2"
[226] "EUExIm2"     "EUExEco2"    "EUImpSov"
[229] "LeavEUI"     "EUconte"     "EUcontu"
[232] "EUconth"     "EULtop1"     "EULtop2"
[235] "EULtop3"     "NHSSat"      "WhySat1"
[238] "WhySat2"     "WhySat3"     "WhySat4"
[241] "WhySat5"     "WhySat6"     "WhySat7"
[244] "WhySat8"     "WhySat9"     "WhySat10"
[247] "WhyDis1"     "WhyDis2"     "WhyDis3"
[250] "WhyDis4"     "WhyDis5"     "WhyDis6"
[253] "WhyDis7"     "WhyDis8"     "WhyDis9"
[256] "WhyDis10"    "GPSat"       "DentSat"
[259] "InpatSat"    "OutpaSat"    "AESat"
[262] "CareSat3"    "NHSFProb"    "NHS5yrs"
[265] "NHSNx5Yr"    "NHSAcc"      "NHSImp"
[268] "AEtravel"    "CareNee2"    "PaySocia"
[271] "CarePa2"     "SocFutur"    "Tranneed"
[274] "Prejtran"    "PMS"         "HomoSex"
[277] "SSRel"       "RSuperv"     "rocsect2e"
[280] "REmpWork"    "REmpWrk2"    "SNumEmp"
[283] "WkJbTim"     "ESrJbTim"    "SSrJbTim"
[286] "WkJbHrsI"    "ExPrtFul"    "EJbHrCaI"
[289] "SJbHrCaI"    "RPartFul"    "S2PartFl"
[292] "Remplyee"    "UnionSA"     "TUSAEver"
[295] "NPWork10"    "RES2010"     "RES2000"
[298] "SLastJb2"    "S2Employ"    "S2Superv"
[301] "S2ES2010"    "S2ES2000"    "rjbtype"
[304] "REconSum"    "REconPos"    "RNSEGGrp"
[307] "RNSocCl"     "RNSSECG"     "RClass"
[310] "RClassGp"    "RSIC07GpE"   "seconsum"
[313] "S2NSEGGp"    "S2NSSECG"    "S2NSocCl"
[316] "S2Class"     "S2ClassG"    "WAGMIN"
[319] "RESPPAY"     "TRCURJM"     "TRCURJN"
[322] "TRMRSJM"     "TRMRSJN"     "TRDIFJM"
[325] "TRDIFJN"     "PHOURS"      "REGHOUR"
[328] "WRKCON"      "JBMRESP"     "JBMWH1"
[331] "JBMWH2"      "JBMWH3"      "JBMWH4"
```

```
[334] "JBMWH5"    "JBMWH6"    "JBMWH7"
[337] "JBMWH8"    "FLEXHRS"   "MgCWld"
[340] "MgMWld"    "ChgAsJb1"  "ChgAsJb2"
[343] "ChgAsJb3"  "ChgJbTim"  "RetExp"
[346] "RetExpb"   "DVRetAge"  "PenKnow2"
[349] "RPenSrc1"  "RPenSrc2"  "RPenSrc3"
[352] "whrbrne"   "NatIdGB"   "NatId"
[355] "tenure2e"  "RentPrf1"  "HAWhat"
[358] "HAgdbd"    "HANotFM"   "LikeHA"
[361] "HAYwhy"    "HANwhy"    "HsDepnd"
[364] "ResPres"   "ReligSum"  "RlFamSum"
[367] "ChAttend"  "bestnatu2" "raceori4"
[370] "DisNew2"   "DisAct"    "DisActDV"
[373] "Knowdis1"  "Knowdis2"  "Knowdis3"
[376] "Knowdis4"  "Knowdis5"  "Knowdis6"
[379] "Knowdis7"  "DisPrj"    "Dis100"
[382] "tea3"      "HEdQual"   "HEdQual2"
[385] "HEdQual3"  "EUIdent"   "BritID2"
[388] "Voted"     "Vote"      "EURefV2"
[391] "EUVOTWHO"  "EURefb"    "AnyBN3"
[394] "MainInc5"  "HHIncD"    "HHIncQ"
[397] "REarnD"    "REarnQ"    "SelfComp"
[400] "knwbdri"   "knwexec"   "knwclea"
[403] "knwhair"   "knwhr"     "knwlaw"
[406] "knwmech"   "knwnurs"   "knwpol"
[409] "knwtchr"   "incdiffs"  "incdsml"
[412] "govldif"   "socblaz"   "whoprvhc"
[415] "whoprvca"  "actgrp"    "actpol"
[418] "actchar"   "govnosa2"  "hhldjob"
[421] "hhmsick"   "hdown"     "hadvice"
[424] "hsococc"   "hlpmny"    "hlpjob"
[427] "hlpadmin"  "hlplive"   "hlpill"
[430] "lckcomp"   "isolate"   "leftout"
[433] "peopadvt"  "peoptrst"  "trstcrts"
[436] "trstprc"   "helpeldy"  "helpslf1"
[439] "helpfrnd"  "fampress"  "reltdemd"
[442] "ffrangr"   "eatout"    "newfrnd"
[445] "pplcont"   "pplftf"    "parcont"
[448] "sibcon2"   "chdcon2"   "othcont"
[451] "frndcont"  "contint"   "ltsgnhth"
[454] "depres"    "diffpile"  "acgoals"
[457] "lifesat2"  "makeem"    "langgs"
[460] "helpslf2"  "payback"   "domconv"
[463] "sitwhr"    "hmecont"   "religcon"
[466] "spseedu"   "ben3000"   "ben3000d"
[469] "falcatch"  "uniaff"    "unicar"
[472] "bothearn"  "sexrole"   "womworka"
[475] "womworkb"  "parlvmf2"  "gendwrk"
[478] "gendmath"  "gendcomp"  "sxbstrm"
[481] "sxbintm"   "sxbstrw"   "sxbintw"
```

```
[484] "sxblaw"                "sxbprov"               "sxboffb"
[487] "sxbnoone"              "sxboth"                "sxbcc"
[490] "carwalk2"              "carbus2"               "carbike2"
[493] "shrtjrn"               "plnallow"              "plnterm"
[496] "plnenvt"               "plnuppri"              "cartaxhi"
[499] "carallow"              "carreduc"              "carnod2"
[502] "carenvdc"              "resclose"              "res20mph"
[505] "resbumps"              "ddnodrv"               "ddnklmt"
[508] "specamsl"              "specammo"              "specamtm"
[511] "speedlim"              "speavesc"              "mobdsafe"
[514] "mobddang"              "mobdban"               "mobdlaw"
[517] "eutrdmv"               "consvfa"               "labrfa"
[520] "libdmfa"               "ukipfa"                "rthdswa2"
[523] "rthdsaw2"              "rthdsca2"              "rthdssa2"
[526] "rthdsprd"              "eqrdisab"              "nhsoutp2"
[529] "nhsinp2"               "bodimr"                "bodimop"
[532] "girlwapp"              "tprwrong2"             "eulunem"
[535] "eulimm"                "eulecon"               "eulwork"
[538] "eullowi"               "eulmlow"               "eulnhs"
[541] "jbernmny"              "jbenjoy"               "topupchn"
[544] "topupnch"              "topuplpa"              "worknow"
[547] "losejob"               "jbgdcurr"              "robots"
[550] "robown"                "voteduty"              "welfhelp"
[553] "morewelf"              "unempjob"              "sochelp"
[556] "dolefidl"              "welffeet"              "damlives"
[559] "proudwlf"              "redistrb"              "BigBusnn"
[562] "wealth"                "richlaw"               "indust4"
[565] "tradvals"              "stifsent"              "deathapp"
[568] "obey"                  "wronglaw"              "censor"
[571] "leftrigh"              "libauth"               "welfare2"
[574] "libauth2"              "leftrig2"              "welfgrp"
[577] "eq_inc_deciles"        "eq_inc_quintiles"      "eq_bhcinc2_deciles"
[580] "eq_bhcinc2_quintiles"
```

2.4. 3. Mean age and its 95% confidence interval

We can now produce a first set of estimates using this information and compare them with those we would have got without accounting for the survey design. We will compute the average (ie mean) age of respondents in the sample. We will need to use `svymean()`

```
svymean(~RAgeE,bsa17.s)
```

```
        mean     SE
RAgeE 48.313 0.4236
```

By default `svymean()` computes the standard error of the mean. We need to embed it within `confint()` in order to get a confidence interval.

```
confint(svymean(~RAgeE,bsa17.s)) ### Just the confidence interval...
```

```
         2.5 %  97.5 %
RAgeE 47.48289 49.1433
```

```
round(
  c(
    svymean(~RAgeE,bsa17.s),
    confint(svymean(~RAgeE,bsa17.s))
    ),
  1) ### ... Or both, rounded
```

```
RAgeE
 48.3  47.5  49.1
```

What difference would it make to the estimates and 95% CI to compute respectively, an unweighted mean, as well as a weighted mean without accounting for the survey design?

There are different ways of computing 'naive estimates' in R. Below we demonstrate how to do it ´by hand' for greater transparency.

Base R provides a function for computing the variance of a variable: `var()`. Since we know that:

- The standard deviation of the mean is the square root of its variance
- The standard error of a sample mean is its standard deviation divided by the square root of the sample size
- A 95% confidence interval is the sample mean respectively minus and plus 1.96 times its standard error. It is then relatively straightforward to compute unweighted and 'casually weighted' confidences intervals for the mean.

```
### Unweighted means and CI
u.m<- mean(bsa17$RAgeE)
u.se<-sqrt(var(bsa17$RAgeE))/sqrt(length(bsa17$RAgeE))
u.ci<-c(u.m - 1.96*u.se,u.m + 1.96*u.se)
round(c(u.m,u.ci),1)
```

```
[1] 52.2 51.6 52.8
```

```
### Weighted means and CI without survey design
w.m<- wtd.mean(bsa17$RAgeE,bsa17$WtFactor)
w.se<-sqrt(wtd.var(bsa17$RAgeE,bsa17$WtFactor))/sqrt(length(bsa17$RAgeE))
w.ci<-c(w.m - 1.96*w.se,w.m + 1.96*w.se)
round(c(w.m,w.ci),1)
```

```
[1] 48.3 47.7 48.9
```

Question 3

What are the consequences of not accounting for the sample design; not using weights and accounting for the sample design when:

- *inferring the mean value of the population age?*

- *inferring the uncertainty of our estimate of the population age?*

2.5. 4. Computing a proportion and its 95% confidence interval

We can now similarly estimate the distribution of a categorical variable in the population by computing proportions (or percentages), for instance, the proportion of people who declare themselves interested in politics. This is the `Politics` variable. It has five categories that we are going to recode into 'Significantly' (interested) and 'Not' (significantly), for simplicity.

The BSA regards 'don't know' and 'refusal' responses as valid but since in this case there is only one 'don't know' and no 'refusal', we can safely ignore these categories and recode them as system missing. As before, we prefer using `xtabs()` over `table()` as it allows us to ignore unused factor levels.

```
attr(bsa17$Politics,"label")     ### Phrasing of the question
```

```
[1] "How much interest do you have in politics?"
```

```
xtabs(~as_factor(Politics),
      data=bsa17,
      drop.unused.levels = T) ### Sample distribution
```

```
as_factor(Politics)
... a great deal,      quite a lot,             some,    not very much,
              739               982              1179               708
 or, none at all?        Don`t know
              379                 1
```

```
bsa17$Politics.s<-ifelse(bsa17$Politics==1 | bsa17$Politics==2,
                         "Significantly",NA)
bsa17$Politics.s<-ifelse(bsa17$Politics>=3 & bsa17$Politics<=5,
                         "Not Interested",bsa17$Politics.s)
bsa17$Politics.s<-as.factor(bsa17$Politics.s)

rbind(xtabs(~as_factor(Politics.s),
      data=bsa17,
      drop.unused.levels = T) ,
      round(
        100*prop.table(
          xtabs(~as_factor(Politics),
          data=bsa17,
          drop.unused.levels = T)
```

```
          ),
        1)
)
```

```
     ... a great deal, quite a lot,  some, not very much, or, none at all?
[1,]            2266.0       1721.0 2266.0         1721.0           2266.0
[2,]              18.5         24.6   29.6           17.8              9.5
     Don`t know
[1,]       1721
[2,]          0
```

Changes in a data frame are not automatically transferred into **svydesign** objects used for inferences. We therefore need to recreate it each time we create or recode a variable.

```
rbind(round(xtabs(WtFactor~Politics.s,bsa17),
            1),
      round(100*
              prop.table(
                xtabs(WtFactor~Politics.s,bsa17))
            ,1)
)
```

```
     Not Interested Significantly
[1,]         2270.6        1715.2
[2,]           57.0          43.0
```

```
bsa17.s<-svydesign(ids=~Spoint,
                   strata=~StratID,
                   weights=~WtFactor,
                   data=bsa17)

rbind(round(svytable(~Politics.s,
                     bsa17.s),1),
      round(100*prop.table(
        svytable(~Politics.s,
                 bsa17.s)),1)
)
```

```
     Not Interested Significantly
[1,]         2270.6        1715.2
[2,]           57.0          43.0
```

As with the mean of age earlier, we can see that the weighted and unweighted point estimates of the proportion of respondents significantly interested in politics differ, even if slightly, and that weighted point estimates do not differ irrespective of the survey design being accounted for.

Let us now examine the confidence intervals of these proportions. Traditional statistical software usually compute these without telling us about the underlying computations going on. By

contrast, doing this in R requires more coding, but in the process we gain a better understanding of what is actually estimated.

Confidence intervals for proportion of categorical variables are usually computed as a sequence of binomial/dichotomic estimations – ie one for each category. In R this needs to be specified explicitly via the `svyciprop()` and `I()` functions. The former actually computes the proportion and its confidence interval (by default 95%), whereas the latter allows us to define the category we are focusing on (in case of non dichotomic variable).

```
svyciprop(~I(Politics.s=="Significantly"),
          bsa17.s)
```

```
                                        2.5% 97.5%
I(Politics.s == "Significantly") 0.430 0.411 0.450
```

```
round(100*
        c(prop.table(
          svytable(~Politics.s,bsa17.s))[2],
attr(svyciprop(~I(Politics.s=="Significantly"),
               bsa17.s),"ci")),1
)
```

```
Significantly          2.5%         97.5%
         43.0          41.1          45.0
```

Question 4

What is the proportion of respondents aged 17-34 in the sample, as well as its 95% confidence interval? You can use ***RAgecat5***

2.6. 5. Domain (ie subpopulation) estimates

Computing estimates for specific groups of a sample (for example the average age of people who reported being interested in politics) is not much more difficult than doing it for the sample as a whole. However doing it as part of an inferential analysis requires some caution. Calculating weighted estimates for a subpopulation, amounts to computing second order estimates ie an estimate for a group whose size needs to be estimated first. Therefore, attempting this while leaving out of the rest of the sample might yield incorrect results. This is why using survey design informed functions is particularly recommended in such cases.

The `survey` package function`svyby()` makes such domain estimation relatively straightforward. For instance, if we would like to compute the mean age of BSA respondents by Government Office Regions, we need to specify:

- The outcome variable whose estimate we want to compute: ie `RAgeE`
- The grouping variable(s) `GOR_ID`
- The estimate function we are going to use here: `svymean`, the same as we used before
- And the type of type of variance estimation we would like to see displayed ie standard errors or confidence interval

```
bsa17$gor.f<-as_factor(bsa17$GOR_ID)
bsa17.s<-svydesign(ids=~Spoint,
                   strata=~StratID,
                   weights=~WtFactor,
                   data=bsa17)

round(svyby(~RAgeE,
            by=~gor.f,
            svymean,
            design=bsa17.s,
            vartype = "ci")[-1],1)
```

```
                           RAgeE ci_l ci_u
A North East                46.1 43.6 48.6
B North West                49.6 47.3 52.0
D Yorkshire and The Humber  48.0 45.2 50.8
E East Midlands             48.6 45.9 51.3
F West Midlands             48.1 45.0 51.2
G East of England           49.0 46.0 52.0
H London                    45.0 43.0 46.9
J South East                48.0 45.1 50.8
K South West                53.4 51.5 55.2
L Wales                     49.1 45.1 53.1
M Scotland                  47.3 44.7 50.0
```

Note: we used `[-1]` from the object created by `svyby()` in order to remove a column with alphanumeric values (the region names), so that we could round the results without getting an error.

Our inference seem to suggest that the population in London is among the youngest in the country, and that those in the South West are among the oldest – their respective 95% confidence intervals do not overlap. We should not feel so confident about differences between London and the South East for example, as the CIs partially overlap.

We can follow a similar approach with proportions: we just need to specify the category of the variable we are interested in as an outcome, for instance respondents who are significantly interested in politics, and replace `svymean` by `svyciprop`.

```
round(
      100*
      svyby(~I(Politics.s=="Significantly"),
            by=~gor.f,
            svyciprop,
            design=bsa17.s,
            vartype = "ci")[-1],
            1)
```

```
                           I(Politics.s == "Significantly") ci_l ci_u
A North East                                            33.4 26.6 40.9
B North West                                            42.1 36.3 48.2
```

```
D Yorkshire and The Humber                35.6 29.1 42.6
E East Midlands                           36.9 32.9 41.1
F West Midlands                           36.3 31.5 41.5
G East of England                         47.2 41.4 53.1
H London                                  54.2 47.2 61.1
J South East                              44.6 38.7 50.8
K South West                              46.5 39.4 53.8
L Wales                                   38.6 27.7 50.7
M Scotland                                42.7 36.0 49.8
```

Question 5

What is the 95% confidence interval for the proportion of people interested in politics in the South West? Is the proportion likely to be different in London? In what way? What is the region of the UK for which the precision of the estimates is likely to be the smallest?

When using `svyby()`, we can define domains or subpopulations with several variables, not just one. For example, we could have looked at gender differences in political affiliations by regions. However, as the size of subgroups decrease, so does the precision of the estimates as their confidence interval widens, to a point where their substantive interest is not meaningful anymore.

Question 6

Using interest in politics as before, and three category age `RAgecat5` *(which you may want to recode as a factor in order to improve display clarity):*

- Produce a table of results showing the proportion of respondents significantly interested in Politics by age group

- Assess whether the age difference in interest for politics is similar for each gender?

- Based on the data, is it fair to say that men aged under 35 tend to be more likely to declare themselves interested in politics than women aged 55 and above?

2.7. Answers

Question 1 The 2017 BSA is a three stage stratified random survey, with postcode sectors, adresses and individuals as the units selected at each stage. Primary sampling units were furthermore stratified according to geographies (sub regions), population density, and proportion of owner-occupiers. Sampling rate was proportional to the size of postcode sectors (ie number of addresses)

Question 2 From the Data Dictionary it appears that the primary sampling units (sub regions) are identified by`Spoint` and the strata by`StratID`. The weights variable is`WtFactor`. Addresses are not provided but could be approximated with a household identifier.

Question 3 Not using weights would make us overestimate the mean age in the population (of those aged 16+) by about 4 years. This is likely to be due to the fact that older respondents are more likely to take part to surveys. Using survey design variables does not alter the value of the estimated population mean. However, not accounting for them would lead us to overestimate the precision/underestimate the uncertainty of our estimate with a narrower confidence interval – by about plus and minus 2 months .

Question 4 The proportion of 17-25 year old in the sample is 28.5 and its 95%confidence interval 26.5, 30.6

Question 5 The 95% confidence interval for the proportion of people interested in politics in the South West is 39.4, 53.8. By contrast, it is likely to be 47.2, 61.1 in London. The region with the lowest precision of estimates (ie the widest confidence interval) is Wales, with a 23 percentage point difference between the upper and lower bounds of the confidence interval.

Question 6

```
bsa17$RAgecat5.f<-as_factor(bsa17$RAgecat5)
bsa17$Rsex.f<-as_factor(bsa17$Rsex)


bsa17.s<-svydesign(ids=~Spoint,
                   strata=~StratID,
                   weights=~WtFactor,
                   data=bsa17)

round(
      100*
      svyby(~I(Politics.s=="Significantly"),
            by=~RAgecat5.f+Rsex.f,
            svyciprop,
            design=bsa17.s,
            vartype = "ci")[c(-8,-4),c(-2,-1)],
      1)
```

```
             I(Politics.s == "Significantly") ci_l ci_u
17-34.Male                               42.9 37.7 48.2
35-54.Male                               50.8 46.6 54.9
55+.Male                                 57.8 53.9 61.6
17-34.Female                             26.3 22.0 31.1
35-54.Female                             34.1 30.6 37.8
55+.Female                               43.0 39.6 46.5
```

Older respondents both male and female tend to be more involved in politics than younger ones.

The confidence intervals for the proportion of men under 35 and women above 55 interested in politics overlap; it is unlikely that they differ in the population.

3. Basic population estimates with BSA data using R

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.

3.1. 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'
       )
```

3.2. 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 s
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?

3.3. 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.

3.4. 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.

3.5. 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.

3.5.1. 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?

3.6. 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.

3.7. 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 "
```

Part II.

SPSS Exercises

4. Using SPSS with weights and survey design variables

This exercise is part of the 'Introduction to the British Social Attitudes Survey (BSA)' online module. In this exercise, we will practice statistical inference with data from the British Social Attitudes Survey (BSA) 2017 using weights and survey design variables.

Please note that at the time of writing this document only some of the BSA editions include survey design variables. For more information about inference from social surveys, including cases where weights and/or survey design variables are not available, please consult our guidelines.

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

4.1. 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

4.2. 1. Identifying the survey design and variables

We first need to find out about the survey design that was used in the BSA 2017, and the design variables available in the dataset. Such information can usually be found in the documentation that comes together with the data under the `mrdoc/pdf` folder or in the data catalogue pages for the data on the UK Data Service website.

Question 1 What is the design that was used in this survey (ie how many stages were there, and what were the units sampled). What were the primary sampling units; the strata (if relevant)?

Now that we are a bit more familiar with the way the survey was designed, we need to try and identify the design variables we can include when producing estimates. The information can usually be found in the user manual or the data dictionary available in the BSA documentation.

Question 2 What survey design variables are available? Are there any that are missing – if so which ones? What is the name of the weights variables?

4.3. 2. Specifying the survey design

Let us first open the 2017 BSA dataset.

```
CD 'C:\Users\YOUR_USER_NAME\Desktop\UKDS'.
  GET
    FILE=' UKDA-8450-spss\spss\spss25\bsa2017_for_ukda.sav'.
```

In principle, we should tell SPSS that we are working with a three stage stratified cluster sample. In practice however, we only have information about the initial ie primary sampling units. This is achieved with the `CSPLAN` command through we create a plan file which contains the survey design information.

```
CSPLAN ANALYSIS
  /PLAN FILE='bsa17_SPSS_design.csaplan'
  /PLANVARS ANALYSISWEIGHT=WtFactor
  /SRSESTIMATOR TYPE=WOR
  /PRINT PLAN
  /DESIGN STAGELABEL='S1' STRATA=StratID CLUSTER=Spoint
  /ESTIMATOR TYPE=WR.
```

4.4. 3. Mean age and its 95% confidence interval

We can now produce a first set of estimates using this design and compare them with those we would have got without accounting for it. We will compute the average (ie mean) age of respondents in the sample, as well as the proportion of male and female respondents aged over 55. We will need to use `/CSDESCRIPTIVES`

```
DATASET ACTIVATE DataSet1.
* Complex Samples Descriptives.
CSDESCRIPTIVES
  /PLAN FILE='bsa17_SPSS_design.csaplan'
  /SUMMARY VARIABLES=RAgeE
  /MEAN
  /STATISTICS SE CIN(95)
  /MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.
```

Under the `/STATISTICS` command we can request either or both the standard error of the mean and its 95% confidence interval.

What difference would it make to the estimates and 95% CI to compute respectively, an unweighted mean, as well as a weighted mean without accounting for the survey design?

Unweighted means and CI

```
DESCRIPTIVES VARIABLES=RAgeE
  /STATISTICS=MEAN SEMEAN.
```

SPSS does not provide an option for computing confidence intervals in this case, but we know that a 95% confidence interval is the sample mean respectively minus and plus 1.96 times its standard error. Using the SPSS output, we can compute it ourselves as 1.96*.2872=about .56 years, that is close to 7 months.

Weighted means and CI without survey design

```
WEIGHT BY WtFactor.
DESCRIPTIVES VARIABLES=RAgeE
  /STATISTICS=MEAN SEMEAN.
WEIGHT OFF.
```

Question 3 What are the consequences of weighing but not accounting for the sample design; not using weights and accounting for the sample design when:

- inferring the mean value of the population age?
- inferring the uncertainty of our estimate of the population age?

4.5. 4. Computing a proportion and its 95% confidence interval

We can now similarly compute an estimate of a proportion (or percentage) of a categorical variable in the population. For instance, the proportion of people who declare themselves interested in politics. This is the `Politics` variable. It has five categories that we are going to recode into 'Significantly' (interested) and 'Not' (significantly) in order to simplify the analysis.

The BSA regards 'don't know' and 'refusal' responses as valid but since in this case there is only one 'don't know' and no 'refusal', we can safely ignore these categories and recode them as system missing.

```
FREQUENCIES VARIABLES=Politics
  /ORDER=ANALYSIS.

RECODE Politics (9=SYSMIS) (1 thru 2=1) (3 thru 5=2) INTO Politics.s.
EXECUTE.

VARIABLE LABELS
Politics.s   "Whether significantly interested in politics".
VALUE LABELS
Politics.s
1  "Significant"
2  "Not significant".
EXECUTE.

FREQUENCIES VARIABLES=Politics.s
  /ORDER=ANALYSIS.

WEIGHT BY WtFactor.
FREQUENCIES VARIABLES=Politics.s
  /ORDER=ANALYSIS.
WEIGHT OFF.
```

As with the mean of age earlier, we can see that the weighted and unweighted point estimates of the proportion of respondents significantly interested in politics change, even if slightly, and that they remain the same when survey design is accounted for.

With the help of `CSTABULATE` we can examine frequencies, proportions and confidence intervals of these proportions accounting for the survey design. As before, the point estimates do not further change once survey design is accounted for.

```
* Complex Samples Frequencies.
CSTABULATE
  /PLAN FILE='bsa17_SPSS_design.csaplan'
  /TABLES VARIABLES=Politics.s
  /CELLS POPSIZE TABLEPCT
  /STATISTICS CIN(95)
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

Question 4 What is the proportion of respondents aged 17-34 in the sample, as well as its 95% confidence interval? You can use `RAgecat5`

4.6. 5. Domain (ie subpopulation) estimates

Although computing estimates for specific groups (for example the average age of people who report being interested in politics) is not conceptually different from doing it for the sample as a whole, doing it with weights as part of an inferential analysis requires some caution. Calculating weighted estimates for a subpopulation while the rest of the sample is left out of the analysis might yield incorrect results. This is why using survey design informed functions is particularly recommended when doing such analyses.

The SPSS command `CSDESCRIPTIVES` that we used above makes such domain estimation relatively straightforward. If we would like to compute the mean age of BSA respondents by government office regions, we need to specify:

- The outcome variable whose estimate we want to compute: ie `RAgeE`
- The grouping variable(s) `GOR_ID`
- And the type of type of variance estimation we would like to see displayed ie standard errors or confidence interval

```
 * Complex Samples Descriptives.
CSDESCRIPTIVES
  /PLAN FILE='bsa17_SPSS_design.csaplan'
  /SUMMARY VARIABLES=RAgeE
  /SUBPOP TABLE=GOR_ID DISPLAY=LAYERED
  /MEAN
  /STATISTICS CIN(95)
  /MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.
```

Our inference seem to suggest that the population in London is among the youngest in the country, and that those in the South West are among the oldest – their respective 95% confidence

intervals do not overlap. We should not feel so confident about differences between London and the South East for example, as the CIs partially overlap.

We can also examine proportions for subpopulations. In order to do this, we need to specify the category of the variable we are interested in as an outcome.For instance, the syntax below uses respondents who are significantly interested in politics:

```
* Complex Samples Frequencies.
CSTABULATE
  /PLAN FILE='bsa17_SPSS_design.csaplan'
  /TABLES VARIABLES=Politics.s
  /SUBPOP TABLE=GOR_ID DISPLAY=LAYERED
  /CELLS TABLEPCT
  /STATISTICS CIN(95)
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

Question 5 What is the 95% confidence interval for the proportion of people interested in politics in the South West? Is the proportion likely to be different in London? In what way? What is the region of the UK for which the precision of the estimates is likely to be the smallest?

When using `CSTABULATE`, we can define domains or subpopulations with several variables, not just one. For example, we could look at gender differences in political affiliations by regions. However, as the size of subgroups decrease, so does the precision of the estimates as their confidence interval widens, to a point where their substantive interest is not meaningful anymore.

Question 6 Using interest in politics as before, and three category age `RAgecat5`:

- Produce a table of results showing the proportion of respondents significantly interested in Politics by age group and gender
- Assess whether the age difference in interest for politics is similar for each gender?
- Based on the data, is it fair to say that men aged under 35 tend to be more likely to declare themselves interested in politics than women aged 55 and above?

4.7. Answers

Question 1 The 2017 BSA is a three stage stratified random survey, with postcode sectors, adresses and individuals as the units selected at each stage. Primary sampling units were stratified according to geographies (sub regions), population density, and proportion of owner-occupiers. Sampling rate was proportional to the size of postcode sectors (ie number of addresses).

Question 2 From the Data Dictionary it appears that the primary sampling units (sub regions) are identified by `Spoint` and the strata by `StratID`. The weights variable is `WtFactor`. Addresses are not provided but could be approximated with a household identifier.

Question 3 Not using weights would make us overestimate the mean age in the population (of those aged 16+) by about 4 years. This is likely to be due to the fact that older respondents are more likely to take part to surveys. Using survey design variables does not alter the value of the estimated population mean. However, not accounting for it would lead us to overestimate the precision/underestimate the uncertainty of our estimate with a narrower confidence interval – by about plus or minus 3 months.

Question 4 The proportion of 17-34 year old in the sample is 28.5 and its 95% confidence interval 26.5, 30.6

Question 5 The 95% confidence interval for the proportion of people interested in politics in the South West is 39.8-53.4. By contrast, it is 47.6-60.8 in London. The region with the lowest precision of estimates (ie the widest confidence interval) is Wales, with a 20 percentage point difference between the upper and lower bounds of the confidence interval.

Question 6

```
CSTABULATE
  /PLAN FILE='bsa17_SPSS_design.csaplan'
  /TABLES VARIABLES=Politics.s
  /SUBPOP TABLE=Rsex BY RAgecat5 DISPLAY=LAYERED
  /CELLS TABLEPCT
  /STATISTICS CIN(95)
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

Subpopulation Tables

Whether significantly interested in politics

Person 1 SEX	Age of respondent(grouped) <3-category> dv			Estimate	95% Confidence Interval Lower	Upper
Male	17-34	% of Total	Significant	42.9%	37.8%	48.1%
			Not significant	57.1%	51.9%	62.2%
			Total	100.0%	100.0%	100.0%
	35-54	% of Total	Significant	50.8%	46.6%	54.9%
			Not significant	49.2%	45.1%	53.4%
			Total	100.0%	100.0%	100.0%
	55+	% of Total	Significant	57.8%	53.9%	61.6%
			Not significant	42.2%	38.4%	46.1%
			Total	100.0%	100.0%	100.0%
	DK/Ref	% of Total	Not significant	100.0%	100.0%	100.0%
			Total	100.0%	100.0%	100.0%
Female	17-34	% of Total	Significant	26.3%	22.1%	31.1%
			Not significant	73.7%	68.9%	77.9%
			Total	100.0%	100.0%	100.0%
	35-54	% of Total	Significant	34.1%	30.7%	37.8%
			Not significant	65.9%	62.2%	69.3%
			Total	100.0%	100.0%	100.0%
	55+	% of Total	Significant	43.0%	39.6%	46.5%
			Not significant	57.0%	53.5%	60.4%
			Total	100.0%	100.0%	100.0%
	DK/Ref	% of Total	Significant	100.0%	100.0%	100.0%
			Total	100.0%	100.0%	100.0%

Figure 4.1.: SPSS output for POLITICS.s by Rsex and RAgecat5

Older respondents both male and female tend to be more involved in politics than younger ones.

The confidence intervals for the proportion of men under 35 and women above 55 interested in politics overlap; it is unlikely that they differ in the population.

5. Basic population estimates with BSA data using SPSS

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.

5.1. 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.

You need to set the folder as your working directory in SPSS. To do this, you need to add the correct file path to the folder on your computer to the code below.

```
* Setting up the working directory
* Change the command below  to match yours:

cd "C:\Users\YOUR_USER_NAME\Desktop\UKDS".
show DIRECTORY.
```

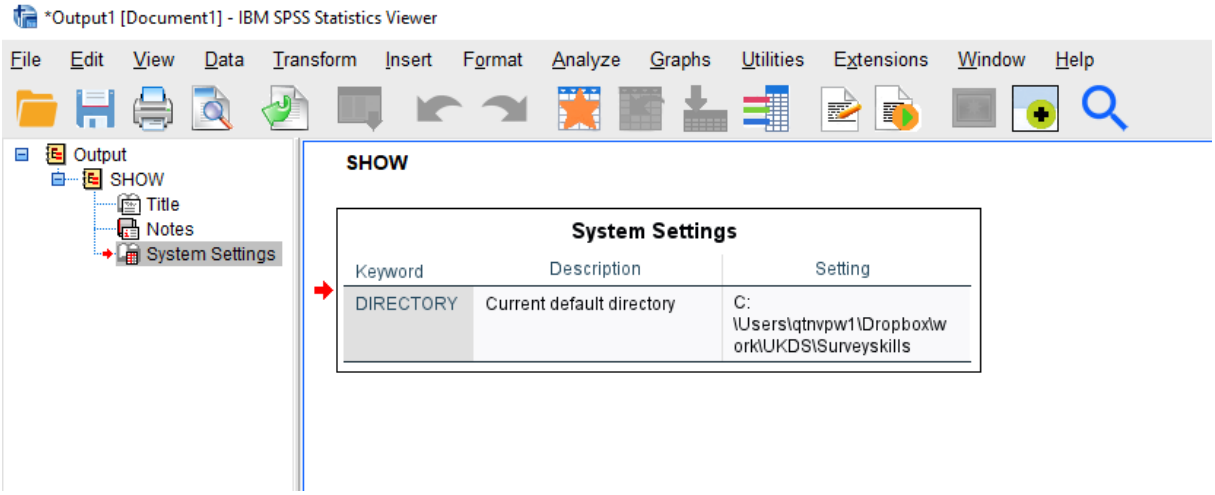
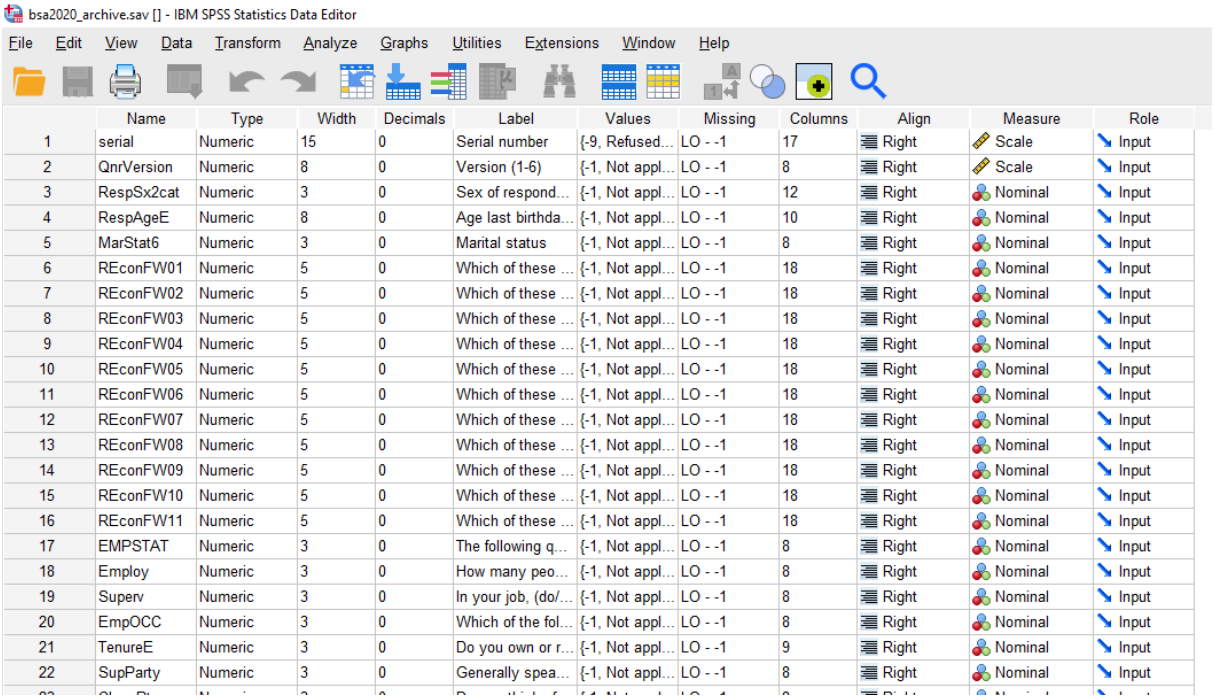


Figure 5.1.: Output of the show DIRECTORY command

If you have your working directory saved to the folder location, the following code should open the BSA dataset.

```
GET FILE='BSA\UKDA-9005-spss\spss\spss25\bsa2020_archive.sav'.
```



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	serial	Numeric	15	0	Serial number	{-9, Refused...	LO - -1	17	Right	Scale	Input
2	QnrVersion	Numeric	8	0	Version (1-6)	{-1, Not appl...	LO - -1	8	Right	Scale	Input
3	RespSx2cat	Numeric	3	0	Sex of respond...	{-1, Not appl...	LO - -1	12	Right	Nominal	Input
4	RespAgeE	Numeric	8	0	Age last birthda...	{-1, Not appl...	LO - -1	10	Right	Nominal	Input
5	MarStat6	Numeric	3	0	Marital status	{-1, Not appl...	LO - -1	8	Right	Nominal	Input
6	REconFW01	Numeric	5	0	Which of these ...	{-1, Not appl...	LO - -1	18	Right	Nominal	Input
7	REconFW02	Numeric	5	0	Which of these ...	{-1, Not appl...	LO - -1	18	Right	Nominal	Input
8	REconFW03	Numeric	5	0	Which of these ...	{-1, Not appl...	LO - -1	18	Right	Nominal	Input
9	REconFW04	Numeric	5	0	Which of these ...	{-1, Not appl...	LO - -1	18	Right	Nominal	Input
10	REconFW05	Numeric	5	0	Which of these ...	{-1, Not appl...	LO - -1	18	Right	Nominal	Input
11	REconFW06	Numeric	5	0	Which of these ...	{-1, Not appl...	LO - -1	18	Right	Nominal	Input
12	REconFW07	Numeric	5	0	Which of these ...	{-1, Not appl...	LO - -1	18	Right	Nominal	Input
13	REconFW08	Numeric	5	0	Which of these ...	{-1, Not appl...	LO - -1	18	Right	Nominal	Input
14	REconFW09	Numeric	5	0	Which of these ...	{-1, Not appl...	LO - -1	18	Right	Nominal	Input
15	REconFW10	Numeric	5	0	Which of these ...	{-1, Not appl...	LO - -1	18	Right	Nominal	Input
16	REconFW11	Numeric	5	0	Which of these ...	{-1, Not appl...	LO - -1	18	Right	Nominal	Input
17	EMPSTAT	Numeric	3	0	The following q...	{-1, Not appl...	LO - -1	8	Right	Nominal	Input
18	Employ	Numeric	3	0	How many peo...	{-1, Not appl...	LO - -1	8	Right	Nominal	Input
19	Superv	Numeric	3	0	In your job, (do/...	{-1, Not appl...	LO - -1	8	Right	Nominal	Input
20	EmpOCC	Numeric	3	0	Which of the fol...	{-1, Not appl...	LO - -1	8	Right	Nominal	Input
21	TenureE	Numeric	3	0	Do you own or r...	{-1, Not appl...	LO - -1	9	Right	Nominal	Input
22	SupParty	Numeric	3	0	Generally spea...	{-1, Not appl...	LO - -1	8	Right	Nominal	Input

Figure 5.2.: BSA dataset in SPSS Variables View

5.2. 1. Explore the dataset

Start by getting an overall feel for the dataset. Either inspect variables and cases in the data editor or use the code below to produce a summary of all the variables in the dataset.

```
CODEBOOK all.
```

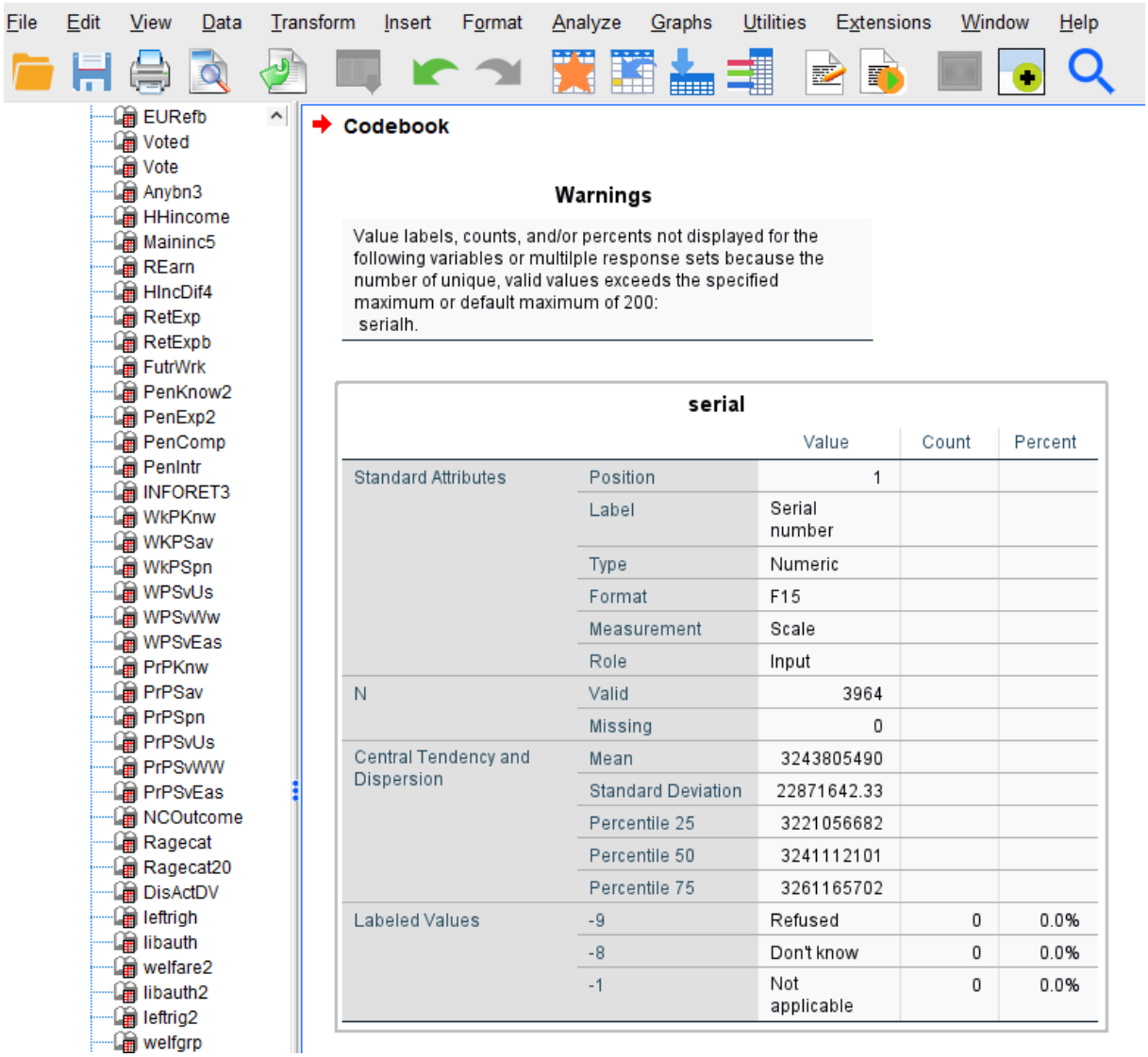


Figure 5.3.: SPSS codebook output for the first variables

Questions

1. What is the overall sample size? 2. How many variables are in the dataset?

Now, focus on the three variables we will use.

```
CODEBOOK TAXSPEND EUVOTWHO PenExp2.
```

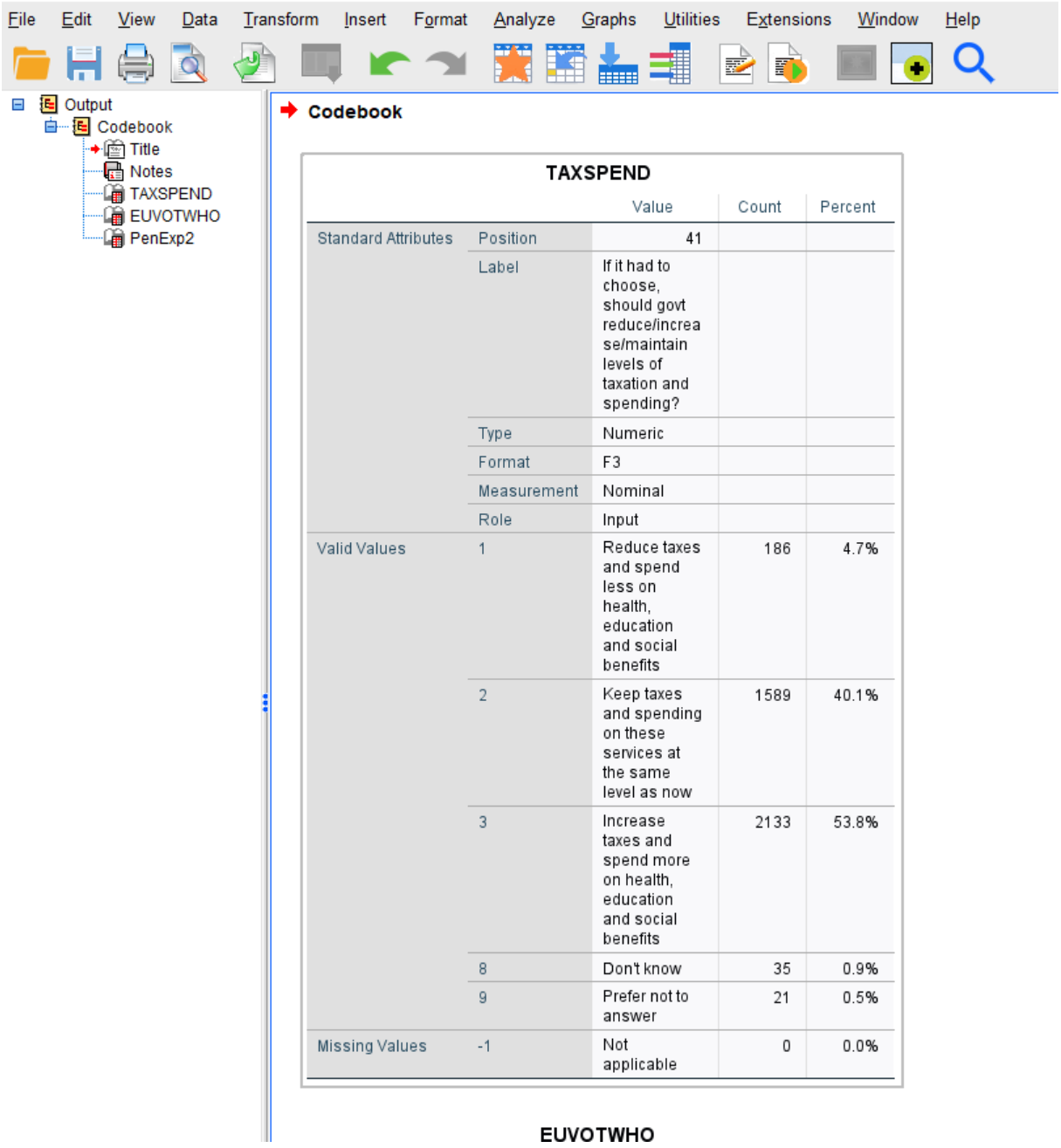


Figure 5.4.: SPSS codebook output for TAXSPEND

Questions 3

What do the variables measure and how?

5.3. 2. Missing values

Review the frequency tables, examining the not applicable and don't know categories.

Question 4

Why for EUVOTWHO are there so many not applicable? Note, you can use the documentation to check if needed. What does this mean when it comes to interpreting the percentages?

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.

5.4. 3. Compare unweighted and weighted frequencies

Let's examine the weighted responses.

```
WEIGHT Off.
*This line is probably not unnecessary as we have not applied a weight yet; it has been incl
FREQUENCIES VARIABLES=TAXSPEND EUVOTWHO
  /BARCHART PERCENT
  /ORDER=ANALYSIS.
*Here, we use the FREQUENCIES command for the categorical variables and the EXAMINE command
EXAMINE VARIABLES=PenExp2
  /PLOT HISTOGRAM
  /STATISTICS DESCRIPTIVES
  /MISSING LISTWISE
  /NOTOTAL.
```

Frequencies

Statistics

		If it had to choose, should govt reduce/increase/maintain levels of taxation and spending?	Did you vote to 'remain a member of the EU' or to 'leave the EU'?
N	Valid	2963	2246
	Missing	1025	1742

Frequency Table

If it had to choose, should govt reduce/increase/maintain levels of taxation and spending?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Reduce taxes and spend less on health, education and social benefits	118	3.0	4.0	4.0
	Keep taxes and spending on these services at the same level as now	928	23.3	31.3	35.3
	Increase taxes and spend more on health, education and social benefits	1805	45.3	60.9	96.2
	(None)	80	2.0	2.7	98.9
	Don`t know	31	.8	1.0	100.0
	Refusal	1	.0	.0	100.0
	Total	2963	74.3	100.0	
Missing	skip, version off route	1025	25.7		
Total		3988	100.0		

Did you vote to 'remain a member of the EU' or to 'leave the EU'?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Remain a member of the European Union	1163	29.2	51.8	51.8
	Leave the European Union	1046	26.2	46.6	98.4
	Prefer not to say	13	.3	.6	98.9
	Don`t remember	7	.2	.3	99.2
	DK/Refusal	17	.4	.8	100.0
	Total	2246	56.3	100.0	
Missing	skip, version off route	1002	25.1		
	Item not applicable	740	18.6		

Figure 5.5.: SPSS output for Frequency distribution of TAXSPEND and EUVOTWHO

What is the (unweighted) percent who say they voted remain in the EU referendum? The answer is about 58 percent of those who voted in the referendum say they voted to remain. This figure seems a bit high (though people do not always report accurately).

Let's add the weight.

```
*The weight is added by the command below. It will remain on for all subsequent analyses.
WEIGHT BY BSA20_wt_new.
FREQUENCIES VARIABLES=TAXSPEND EUVOTWHO
   /ORDER=ANALYSIS.
EXAMINE VARIABLES=PenExp2
  /PLOT HISTOGRAM
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.

*To stop weighting the data you can use the following command.
WEIGHT off.
```

Descriptives

			Statistic	Std. Error
How much do you think someone who reaches State Pension age today would receive in pounds per week?	Mean		1292.65	58.185
	95% Confidence Interval for Mean	Lower Bound	1178.56	
		Upper Bound	1406.73	
	5% Trimmed Mean		878.12	
	Median		160.00	
	Variance		9777451.114	
	Std. Deviation		3126.892	
	Minimum		0	
	Maximum		9999	
	Range		9999	
	Interquartile Range		80	
	Skewness		2.421	.046
	Kurtosis		3.877	.091

How much do you think someone who reaches State Pension age today would receive in pounds per week?

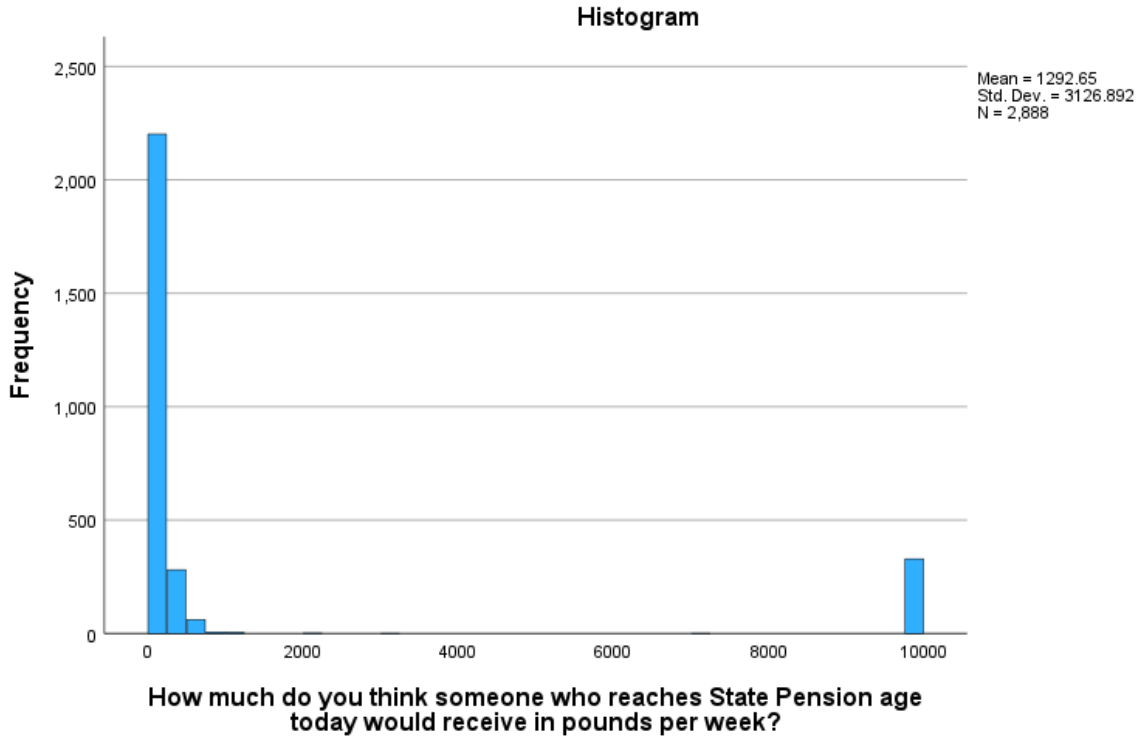


Figure 5.6.: SPSS output for Frequency distribution of TAXSPEND and EUVOTWHO

Now, what proportion say they voted remain in the EU referendum? It is about 54 percent, lower than the unweighted proportion and closer to the actual referendum results.

5.5. 4. Confidence intervals

Add confidence intervals to the bar charts and mean to indicate uncertainty due to sampling error.

```
WEIGHT BY BSA20_wt_new.
GRAPH
  /BAR(SIMPLE)=PCT BY TAXSPEND
  /INTERVAL CI(95.0).

GRAPH
  /BAR(SIMPLE)=PCT BY EUVOTWHO
  /INTERVAL CI(95.0).

EXAMINE VARIABLES=PenExp2
  /PLOT NONE
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

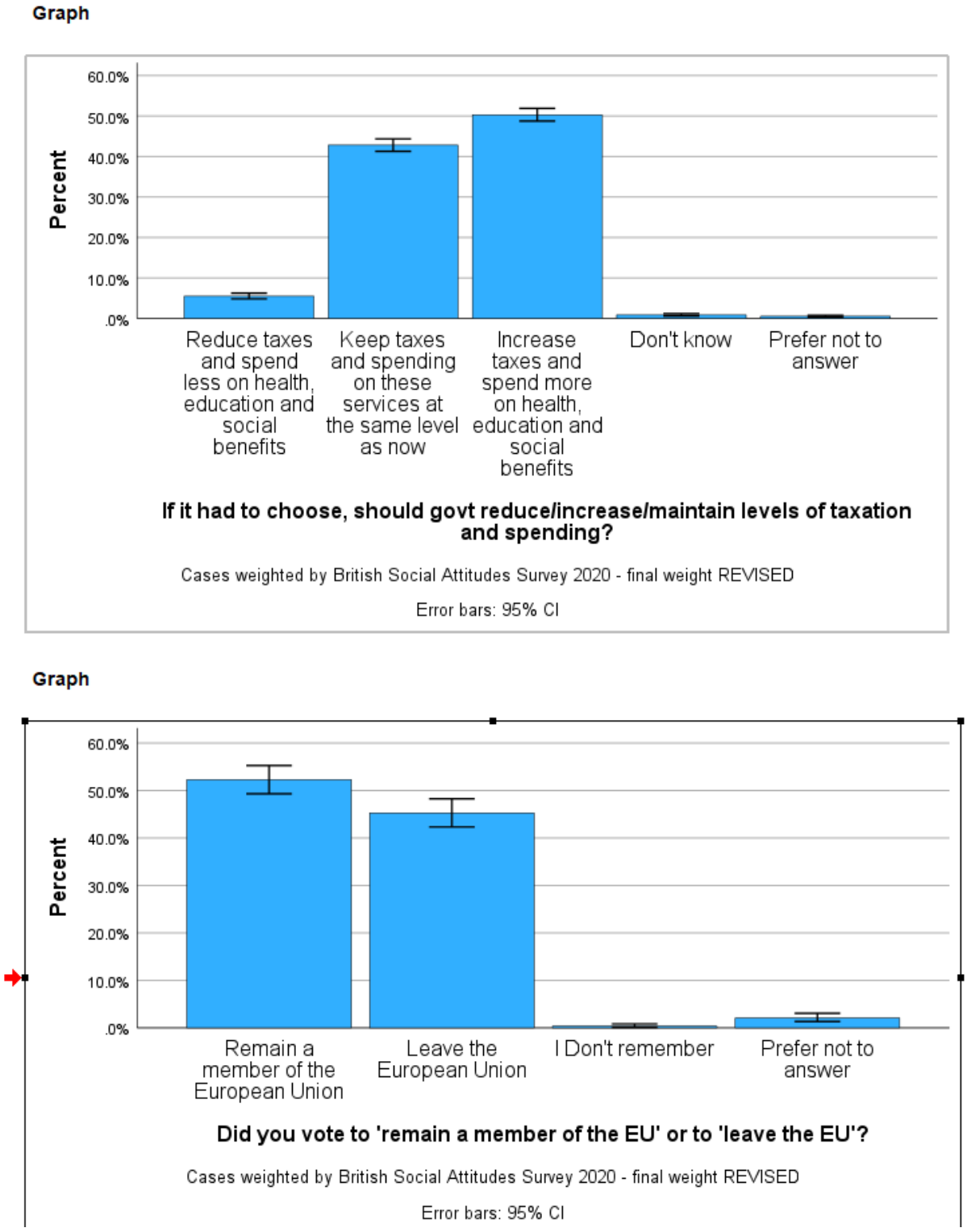


Figure 5.7.: SPSS output for GRAPH BAR of TAXSPEND and EUVOTWHO

Question 5

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

Question 6

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

Additional question

Select two variables that interest you and examine their distribution.

5.6. Answers

1. There are 3964 cases in the dataset.
2. The total number of variables is 213.
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. About 50% say the government should increase taxes and spend more.
6. The amount people think they will get at state pension age varies between £0 and £7000, with an average in the region between £170 and £184

5.7. Appendix: recoding nonresponses as system missing

The code below provides and example of how to recode missing values including 'don't knows' and 'prefer not to say' into system missing.

The SPSS syntax below includes the command, the variables and the relevant missing values in (). Note, you can set missing values more than 1 at a time if they have the same missing value pattern.

```
COMPUTE EUVOTWHO_m=EUVOTWHO.
COMPUTE TAXSPEND_m=TAXSPEND.
COMPUTE PenExp2_m=PenExp2.

MISSING VALUES PenExp2_m (-1, 9998, 9999).
MISSING VALUES TAXSPEND_m (-1, 8, 9).
MISSING VALUES EUVOTWHO_m (-1, 3 THRU 9).
```