Unit 5 – STATA for Data Description
Homework

SOLUTIONS

Initialize R Studio Session.

```
setwd("~/Desktop")              # Set working directory
getwd()                         # Check working directory

## [1] "/Users/cbigelow/Desktop"

options(scipen=999)             # Turn off scientific notation
rm(list = ls())                 # Clear the Decks
```

Input data

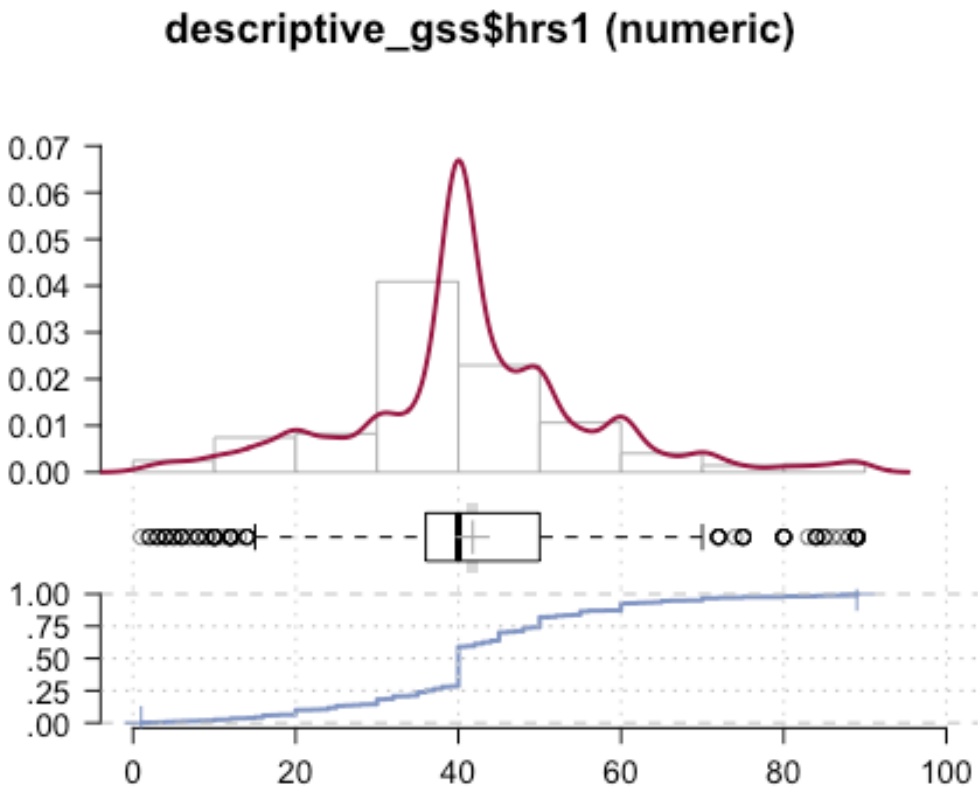
```
load(file="descriptive_gss.Rdata")
#str(descriptive_gss)                        #  Not knitting this
```

Q1A - Detailed descriptives for hrs1

```
library(stargazer)
library(summarytools)
library(FSA)
library(DescTools)
library(psych)
library(Rmisc)

# Using base
print("Method I: {base}: summary( )")
summary(descriptive_gss$hrs1)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   36.00   40.00   41.78   50.00   89.00    1036


# Using package stargazer
# What could go wrong: {stargaze} works on dataframes only
descriptive_gss <- as.data.frame(descriptive_gss)
print("METHOD II:  {stargazer}: stargazer( )")
stargazer::stargazer(descriptive_gss[c("hrs1")],type="text",summary.stat=c("n", "mean", "sd", "min", "p25"
, "median", "p75", "max"),title="descriptive_gss.Rdata")

##
## descriptive
## ==========================================================================
## Statistic   N    Mean  St. Dev.  Min  Pctl(25) Median Pctl(75)  Max
## --------------------------------------------------------------------------
## hrs1      1,729 41.777  14.623  1.000  36.000  40.000  50.000  89.000
## --------------------------------------------------------------------------
```

```
# Using package summarytools - default
print("METHOD III:  {summarytools}:  descr( )")
summarytools:: escry(descriptive_gss$hrs1, transpose=TRUE)

## Descriptive Statistics
## descriptive_gss$hrs1
## N: 2765
##
##               Mean   Std.Dev    Min      Q1   Median      Q3     Max     MAD     IQR      CV
## ---------- ------- --------- ------ ------- -------- ------- ------- ------- ------- ------
##       hrs1   41.78     14.62   1.00   36.00    40.00   50.00   89.00   10.38   14.00   0.35
##
## Table: Table continues below
##
##
##
##              Skewness   SE.Skewness   Kurtosis   N.Valid   Pct.Valid
## ---------- ---------- ------------- ---------- --------- -----------
##       hrs1       0.28          0.06       1.31   1729.00       62.53



# Using package summarytools - user selects stats
print("METHOD IV:  {summarytools}:  descr( )")
summarytools::descr(descriptive_gss$hrs1, stats = c("N.Valid","min", "med","mean","sd","max"),
                    transpose=TRUE, title="descriptive_gss.Rdata")

## Descriptive Statistics
## descriptive_gss$hrs1
## N: 2765
##
##              N.Valid    Min   Median    Mean   Std.Dev     Max
## ---------- --------- ------ -------- ------- --------- -------
##       hrs1   1729.00   1.00    40.00   41.78     14.62   89.00



# Using package FSA
print("Method V:  {FSA}: Summarize( )")
FSA::Summarize(hrs1 ~ 1,data=descriptive_gss,digits=2, na.rm=TRUE)

##       n  nvalid    mean      sd     min      Q1  median      Q3     max
## 2765.00 1729.00   41.78   14.62    1.00   36.00   40.00   50.00   89.00
```

```r
# Using package DescTools
print("Method VI: {DescTools}: Desc( )")
DescTools::Desc(descriptive_gss$hrs1)

## ------------------------------------------------------------------------
## descriptive_gss$hrs1 (numeric)
##
##   length      n    NAs  unique     0s   mean  meanCI
##    2'765  1'729  1'036      76      0  41.78   41.09
##          62.5%  37.5%           0.0%          42.47
##
##      .05    .10    .25  median    .75    .90     .95
##    16.00  21.80  36.00   40.00  50.00  60.00   68.00
##
##    range     sd  vcoef     mad    IQR   skew    kurt
##    88.00  14.62   0.35   10.38  14.00   0.28    1.31
##
## lowest : 1.0, 2.0 (3), 3.0 (4), 4.0 (5), 5.0 (4)
## highest: 85.0 (3), 86.0, 87.0, 88.0 (2), 89.0 (17)
##
## heap(?): remarkable frequency (30.8%) for the mode (= 40)
```



```
# Using package psych
print("Method VI: {psych}: describe( )")
psych::describe(descriptive_gss$hrs1)

##    vars    n  mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 1729 41.78 14.62     40   41.64 10.38   1  89    88 0.28     1.31
##      se
## X1 0.35



# Using package Rmisc
print("Method VII: {Rmisc}: summarySE() ")
Rmisc::summarySE(data=descriptive_gss,"hrs1", na.rm=TRUE)

##    .id    N     hrs1       sd        se        ci
## 1 <NA> 1729 41.77675 14.62304 0.3516739 0.6897512
```

Q1B - In one table, descriptive statistics for hrs1, age, and wwwhr

```
library(stargazer)
library(tidyverse)
library(summarytools)

# Using package stargazer
# What could go wrong: {stargaze} works on dataframes only
print("METHOD I:  {stargazer}: stargazer( )")
descriptive_gss <- as.data.frame(descriptive_gss)
stargazer::stargazer(descriptive_gss[c("hrs1","age","wwwhr")],type="text",summary.stat=c("n", "mean", "sd"
, "min", "p25", "median", "p75", "max"),title="descriptive_gss.Rdata")

##
## descriptive
## ===============================================================================
## Statistic   N    Mean  St. Dev.  Min   Pctl(25) Median Pctl(75)   Max
## -------------------------------------------------------------------------------
## hrs1      1,729 41.777  14.623  1.000   36.000  40.000  50.000  89.000
## age       2,751 46.283  17.370  18.000  32.000  44.000  58.000  89.000
## wwwhr     1,574 5.908   8.867   0.000   1.000   3.000   7.000   112.000
## -------------------------------------------------------------------------------



# Using package summarytools
print("METHOD II: {summarytools}: descr( )")
q1bvars <- descriptive_gss %>% select("hrs1", "age", "wwwhr")
summarytools::descr(q1bvars,stats = c("N.Valid","min", "med","mean","sd","max"),
                    transpose=TRUE, title="descriptive_gss.Rdata")

## Descriptive Statistics
## q1bvars
## N: 2765
##
##              N.Valid     Min   Median    Mean   Std.Dev      Max
## ----------- --------- ------- -------- ------- --------- --------
##         age   2751.00   18.00    44.00   46.28     17.37    89.00
##        hrs1   1729.00    1.00    40.00   41.78     14.62    89.00
##       wwwhr   1574.00    0.00     3.00    5.91      8.87   112.00
```

Q1C - Descriptives for hr1, separately for groups defined by polviews

```
library(summarytools)
library(FSA)
library(Rmisc)

# Using package summarytools
print("METHOD I:  {summarytools}")
with(descriptive_gss, stby(data =hrs1, INDICES = polviews,
 FUN =  escry,
 stats = c("N.Valid","min", "med","mean","sd","max"),transpose=TRUE))

## Descriptive Statistics
## hrs1 by polviews
## Data Frame: descriptive_gss
## N: 210
##
##                             N.Valid     Min   Median    Mean   Std.Dev     Max
## ------------------------- --------- ------- -------- ------- --------- -------
##               conservative    125.00    2.00    40.00   43.90     15.49   89.00
##          extremely liberal     27.00    5.00    40.00   37.89     16.83   72.00
##       extrmly conservative     25.00   10.00    40.00   37.00     10.26   52.00
##                    liberal     93.00   12.00    40.00   43.84     16.46   89.00
##                   moderate    336.00    2.00    40.00   42.15     13.46   89.00
##       slghtly conservative    129.00    9.00    40.00   42.36     16.23   89.00
##           slightly liberal    110.00    5.00    40.00   42.84     11.88   70.00


# Using package FSA - group variable must be factor
print("METHOD II:  {FSA}: Summarize( )")
descriptive_gss$polviewsF <- factor(descriptive_gss$polviews)
FSA::Summarize(hrs1~polviewsF,data=descriptive_gss,na.rm=TRUE)

##              polviewsF   n nvalid     mean       sd min    Q1 median    Q3
## 1         conservative 210    125 43.89600 15.49392   2 40.00     40 50.00
## 2    extremely liberal  47     27 37.88889 16.82565   5 22.00     40 50.00
## 3 extrmly conservative  41     25 37.00000 10.25914  10 35.00     40 42.00
## 4             liberal 143     93 43.83871 16.46359  12 35.00     40 55.00
## 5            moderate 522    336 42.15476 13.45617   2 39.75     40 48.00
## 6 slghtly conservative 209    129 42.36434 16.22543   9 36.00     40 50.00
## 7     slightly liberal 159    110 42.83636 11.87977   5 40.00     40 49.75
##   max
## 1  89
## 2  72
## 3  52
## 4  89
## 5  89
## 6  89
## 7  70
```

```
# Using package Rmisc  (NOTE:  ci value is half of width of 95% CI for mean)
print("METHOD III: {Rmisc}: summarySE( )")
Rmisc::summarySE(data=descriptive_gss,measurevar="hrs1",groupvars=c("polviewsF"),na.rm=TRUE)

##              polviewsF   N     hrs1       sd        se        ci
## 1         conservative 125 43.89600 15.49392 1.3858187 2.7429234
## 2    extremely liberal  27 37.88889 16.82565 3.2380982 6.6560062
## 3 extrmly conservative  25 37.00000 10.25914 2.0518285 4.2347658
## 4              liberal  93 43.83871 16.46359 1.7071947 3.3906361
## 5             moderate 336 42.15476 13.45617 0.7340943 1.4440153
## 6 slghtly conservative 129 42.36434 16.22543 1.4285697 2.8266692
## 7     slightly liberal 110 42.83636 11.87977 1.1326916 2.2449580
## 8                 <NA> 884 41.15271 14.77189 0.4968325 0.9751104
```

Q2A - One way frequencies and relative frequency tables for: marital, sex, and satjob7

```
library(summarytools)
freq(descriptive_gss$marital)

## Frequencies
## descriptive_gss$marital
## Type: Character
##
##                      Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------------- ------ --------- -------------- --------- --------------
##            divorced    445     16.09          16.09     16.09          16.09
##             married   1269     45.90          61.99     45.90          61.99
##       never married    708     25.61          87.59     25.61          87.59
##           separated     96      3.47          91.07      3.47          91.07
##             widowed    247      8.93         100.00      8.93         100.00
##                <NA>      0                                0.00         100.00
##               Total   2765    100.00         100.00    100.00         100.00

freq(descriptive_gss$sex)

## Frequencies
## descriptive_gss$sex
## Type: Character
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------ ------ --------- -------------- --------- --------------
##       female   1537     55.59          55.59     55.59          55.59
##         male   1228     44.41         100.00     44.41         100.00
##         <NA>      0                                0.00         100.00
##        Total   2765    100.00         100.00    100.00         100.00
```

```
freq(descriptive_gss$satjob7)
```

```
## Frequencies
## descriptive_gss$satjob7
## Type: Character
##
##                                        Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------------------------------- ------ --------- -------------- --------- --------------
##                  completely dissatisfied     11      1.34           1.34      0.40           0.40
##                     completely satisfied    127     15.49          16.83      4.59           4.99
##                      fairly dissatisfied     47      5.73          22.56      1.70           6.69
##                         fairly satisfied    264     32.20          54.76      9.55          16.24
##       neither satisfied nor dissatisfied     53      6.46          61.22      1.92          18.16
##                       very dissatisfied     29      3.54          64.76      1.05          19.20
##                          very satisfied    289     35.24         100.00     10.45          29.66
##                                    <NA>   1945                               70.34         100.00
##                                   Total   2765    100.00         100.00    100.00         100.00
```

Q2B - Two Way Cross Tabulation with row percentages of sex and polviews

```
library(summarytools)
library(gmodels)

# Using package summarytools
# What could go wrong - it does not work to preface this with summarytools::
print("METHOD I:  {summarytools}: ctable( )")
with(descriptive_gss, ctable(polviews, sex, prop = "r", totals = FALSE))
```

```
## Cross-Tabulation, Row Proportions
## polviews * sex
## Data Frame: descriptive_gss
##
## ---------------------- ----- ------------ -------------
##                    sex        female          male
##              polviews
##           conservative         103 (49.0%)   107 (51.0%)
##      extremely liberal          23 (48.9%)    24 (51.1%)
##   extrmly conservative          29 (70.7%)    12 (29.3%)
##                liberal          72 (50.3%)    71 (49.7%)
##               moderate         278 (53.3%)   244 (46.7%)
##   slghtly conservative         109 (52.2%)   100 (47.8%)
##       slightly liberal          88 (55.3%)    71 (44.7%)
##                   <NA>         835 (58.2%)   599 (41.8%)
## ---------------------- ----- ------------ -------------
```

```r
# Using package gmodels
print("METHOD II:  {gmodels}: CrossTable( )")
CrossTable(descriptive_gss$polviews,descriptive_gss$sex,digits=2,
         prop.r=TRUE,prop.c=FALSE,prop.t=FALSE,prop.chisq=FALSE,
         dnn=c("Political Views","Gender"))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  1331
##
##
##                      | Gender
##      Political Views |    female |      male | Row Total |
## ---------------------|-----------|-----------|-----------|
##         conservative |       103 |       107 |       210 |
##                      |      0.49 |      0.51 |      0.16 |
## ---------------------|-----------|-----------|-----------|
##    extremely liberal |        23 |        24 |        47 |
##                      |      0.49 |      0.51 |      0.04 |
## ---------------------|-----------|-----------|-----------|
## extrmly conservative |        29 |        12 |        41 |
##                      |      0.71 |      0.29 |      0.03 |
## ---------------------|-----------|-----------|-----------|
##              liberal |        72 |        71 |       143 |
##                      |      0.50 |      0.50 |      0.11 |
## ---------------------|-----------|-----------|-----------|
##             moderate |       278 |       244 |       522 |
##                      |      0.53 |      0.47 |      0.39 |
## ---------------------|-----------|-----------|-----------|
## slghtly conservative |       109 |       100 |       209 |
##                      |      0.52 |      0.48 |      0.16 |
## ---------------------|-----------|-----------|-----------|
##     slightly liberal |        88 |        71 |       159 |
##                      |      0.55 |      0.45 |      0.12 |
## ---------------------|-----------|-----------|-----------|
##         Column Total |       702 |       629 |      1331 |
## ---------------------|-----------|-----------|-----------|
##
##
```