RECSM: Quantitative Methods in Social Research

Day 1 - 02 07 2025

Burak Sonmez

Please skim through all of the instructions before getting started.

Note: This document also walks you through installing R and RStudio.

1. Install R

- Go to https://cran.r-project.org/
- Select "Download R for Windows"
- Click "install R for the first time"
- Click "Download R 4.0.3 for Windows"
- Select "Download R for (Mac) OS X"
- Click "R-4.0.3.pkg"

2. Install RStudio

- Go to https://www.rstudio.com/products/rstudio/download/
- Select "RStudio 1.3.1093 - Windows 7+ (64-bit)" for Windows users
- Select "RStudio 1.3.1093 - macOS 10.13+ (64-bit)" for Mac users

3. Problem set 1

Step 1: What is the result of summing all numbers from 1 to 250?

Step 2: Calculate the square root of 1500 and round it to the nearest possibility.

Step 3: For example, suppose the yearly number of MSc students at UPF during the period 2010 to 2020 is 75, 152, 230, 141, 292, 79, 211, 137, 156, and 111, respectively. What are the mean, the variance, and the standard deviation?

$$Hint: \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

4. Problem set 2

Step 1: First download the folder (RECSM workshop) on your computer

Step 2: Create a new file called "day1.R" in your RECSM workshop folder and write all the solutions in it.

Step 3: Clear the workspace and set the working directory to your RECSM workshop folder.

Step 4: Load the Asylum dataset into your R environment and inspect the name of variables. The asylum data is about the perception of asylum seeker size in the UK.

Table 1: The Codebook

Variable	Description
asylum	How many do you think are asylum seekers from Syria out of every 100 asylees in Britain?
sex	1 = male, 2 = female
age	Age of respondent
news	Do you normally read any daily morning newspaper at least three times a week?
religious	Do you regard yourself as belonging to any particular religion?
urban	Population density, 4 categories (highest density is 4, lowest is 1)
hhincome	Income bands for household, high number = high HH income
partyid	1 = Tories, 2 = Labour, 3 = SNP, 4 = Greens, 5 = Ukip, 6 = BNP, 7 = other

Step 5: What is the level of measurement for each variable in the Asylum dataset?

Step 6: Calculate the correct measure of central tendency for asylum, hhincome, news.

Step 7: Calculate the correct measure of dispersion for asylum, hhincome, news.

Step 8: How many respondents identify with the Labour?

Step 9: Calculate the variance and standard deviation of asylum for each party identification.

Step 10: Find the party identification of the oldest and youngest respondents.

Step 11: Find the 20th, 40th, 60th and 80th percentiles of age.

Step 12: Create a box plot for asylum grouped by the religious variable to show the difference between asylum for people who believe in any religion and people who do not.

Step 13:What is the mean of asylum for men and for women?

5. Problem set 3

Step 1: Using the Asylum dataset, please try to answer to what extent political differences explain perceptional biases regarding Syrian asylum seekers.

Step 2: Report summary statistics of the variables (asylum, age, news, sex). Make sure that you define the variables of news and sex as a factor.

Step 3: Explore the relationship between sex and asylum perception through a boxplot, using `ggplot`. You can customise your ggplot boxplot by adding more layers (e.g. labels, title, aesthetic specifications). Please state your interprations/observations from plotting.

Step 4: Please conduct hypothesis testing to confirm whether the mean differences in asylum perception are statistically significant between male and female respondents. Hint: (T-test)

- **Null Hypothesis**: The average perception of Syrian asylum seeker size is the same between male and female respondents
- **Alternative Hypothesis**: The average perception of Syrian asylum seeker size differs between different sex categories.

Solutions for problem sets

```
#Remove objects from the enviroment
rm(list=ls())
#Set your working directory
setwd("~/Downloads/RECSM workshop")

#1)
sum_1_to_250 <- seq(1,250, 1)
sum_results <- sum(sum_1_to_250)
sum_results
```

```
## [1] 31375
```

```
#2)
round(sqrt(1500))
```

```
## [1] 39
```

```
#3)
msc <- c(75, 152, 230, 141, 292, 79, 211, 137, 156, 111)
mean(msc)
```

```
## [1] 158.4
```

```
var(msc)
```

```
## [1] 4672.933
```

```
sqrt(var(msc))
```

```
## [1] 68.35886
```

```
sqrt( sum( (msc - mean(msc))^2 /(length(msc)-1)))
```

```
## [1] 68.35886
```

```
sd(msc)
```

```
## [1] 68.35886
```

```r
load("asylum_data.RData")
names(asy)
```

```
## [1] "asylum"    "sex"       "age"       "news"      "religious" "urban"
## [7] "hhincome"  "partyid"
```

```r
#base R solution

#correct measure of central tendency for asylum, hhincome, news
mean(asy$asylum)
```

```
## [1] 29.03051
```

```r
mean(asy$hhincome)
```

```
## [1] 9.586273
```

```r
mean(asy$news)
```

```
## [1] 0.4537655
```

```r
##correct measure of dispersion for asylum, hhincome, news
sd(asy$asylum)
```

```
## [1] 21.06331
```

```r
sd(asy$hhincome)
```

```
## [1] 4.639529
```

```r
mean(asy$news)
```

```
## [1] 0.4537655
```

```r
##Forcing R not to use exponential/scientific notation
options(scipen = 999)

#How many respondents identify with the Labour?
asy$partyid <- factor(asy$partyid, labels = c("Tories", "Labour", "SNP", "Greens", "Ukip", "BNP", "othe
r"))

#just look at the frequency table or filter it
table(asy$partyid)
```

```
##
## Tories Labour    SNP Greens   Ukip    BNP  other
##    284    280     16     23     31     32    383
```

```r
length(which(asy$partyid=="Labour"))
```

```
## [1] 280
```

```r
#Calculate the variance and standard deviation of asylum for each party identification
# Tories
var(asy$asylum[asy$partyid=="Tories"])
```

```
## [1] 431.8308
```

```r
sd(asy$asylum[asy$partyid=="Tories"])
```

```
## [1] 20.78054
```

```r
#Find the party identification of the oldest and youngest respondents
# range for age
range(asy$age)
```

```
## [1] 17 99
```

```r
oldest <- which(asy$age == max(asy$age))
youngest <- which(asy$age == min(asy$age))

# party affiliation of those respondents
asy$partyid[oldest]
```

```
## [1] other  Labour
## Levels: Tories Labour SNP Greens Ukip BNP other
```

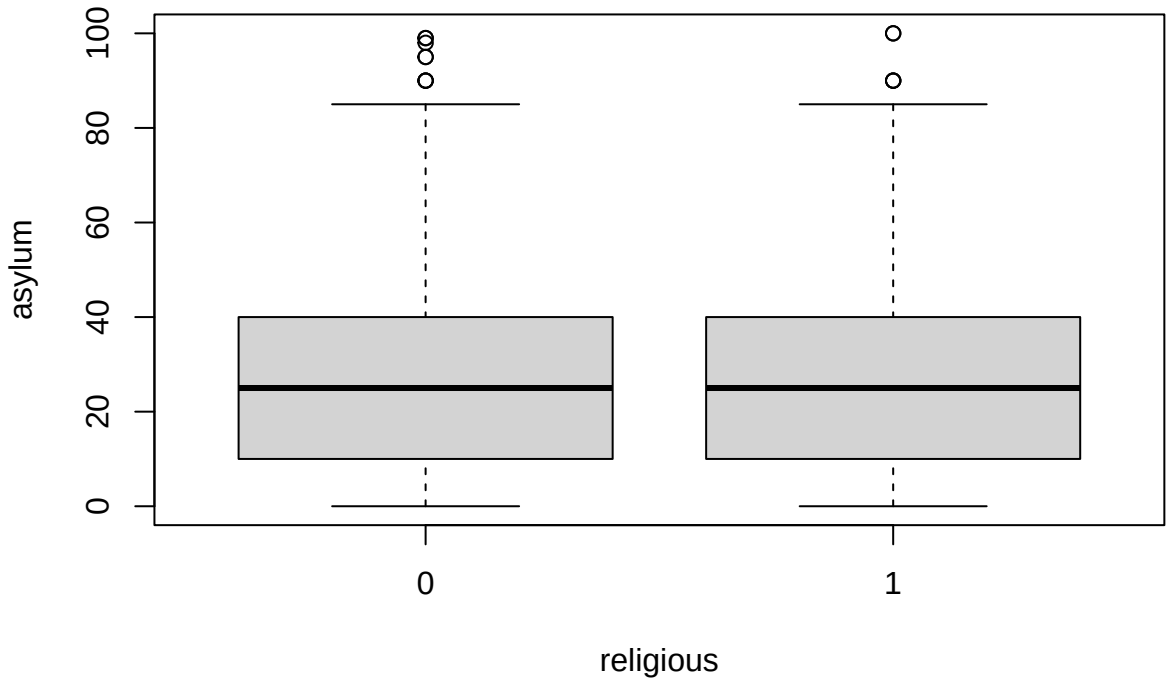
```r
asy$partyid[youngest]
```

```
## [1] other
## Levels: Tories Labour SNP Greens Ukip BNP other
```

```r
#Find the 20th, 40th, 60th and 80th percentiles of age
quantile(asy$age, c(.2, .4, .6, .8))
```

```
## 20% 40% 60% 80%
##  33  44  55  66
```

```r
#a box plot for asylum grouped by the religious variable
boxplot( asylum ~ religious, data = asy)
```




```
#What is the mean of asylum for men and for women
mean(asy$asylum[asy$sex==2]) - mean(asy$asylum[asy$sex==1])
```

```
## [1] 8.253937
```

```
##Instal the packages first if those packages are not installed before
library(ggplot2)
```

```
#To what extent political differences explain perceptional biases regarding Syrian asylum seekers.
by(asy$asylum, asy$partyid, summary)
```

```
## asy$partyid: Tories
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   20.00   27.73   40.00   99.00
## ------------------------------------------------------------
## asy$partyid: Labour
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    10.0    25.0    27.9    40.0   100.0
## ------------------------------------------------------------
## asy$partyid: SNP
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   30.00   30.00   33.75   36.25   60.00
## ------------------------------------------------------------
## asy$partyid: Greens
```

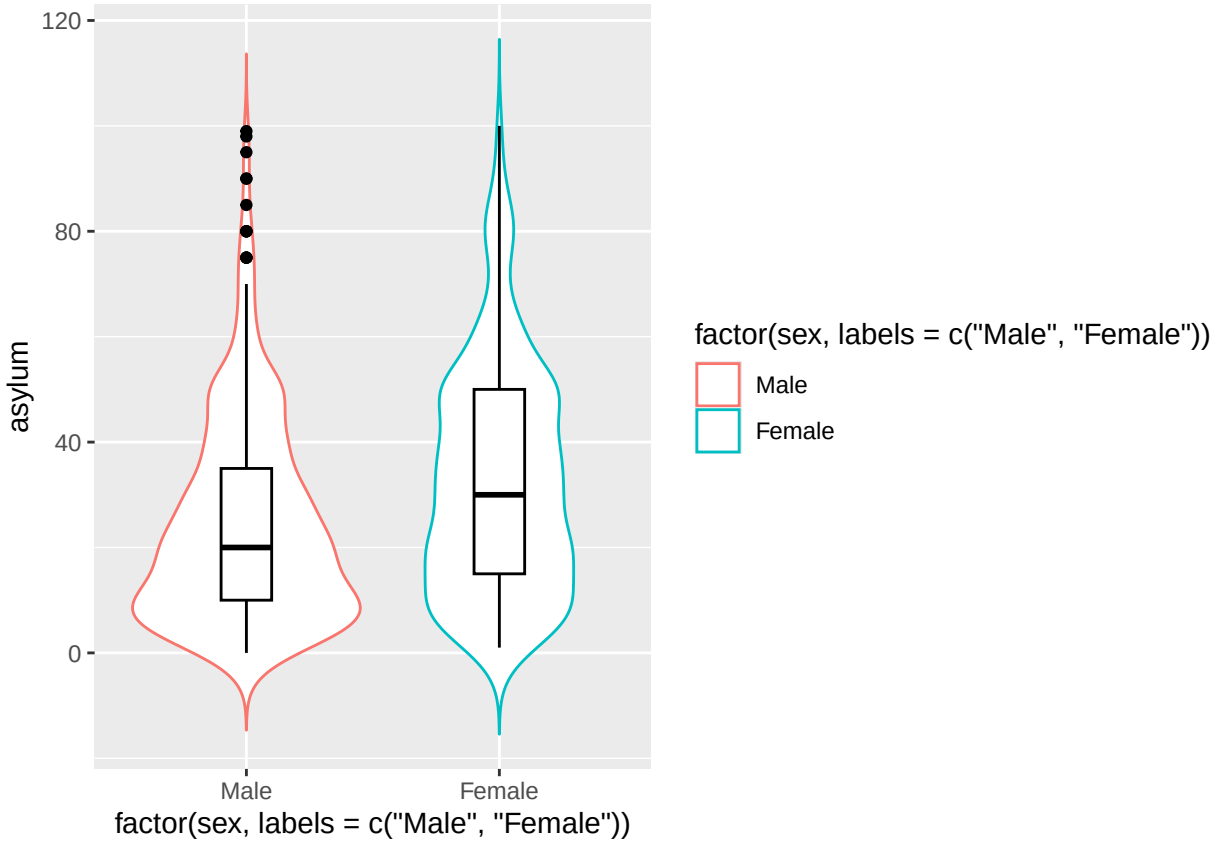
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    5.00   18.00   24.91   34.00   80.00
## ------------------------------------------------------------
## asy$partyid: Ukip
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   11.00   20.00   24.29   35.00   70.00
## ------------------------------------------------------------
## asy$partyid: BNP
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   27.50   40.00   42.19   56.25   90.00
## ------------------------------------------------------------
## asy$partyid: other
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   11.00   25.00   30.16   45.00   99.00
```

```
#Summary statistics (asylum, age, news, sex)
asy$news <- factor(asy$news, labels = c("Don't read", "Read"))
asy$sex <- factor(asy$sex, labels = c("Male", "Female"))
summary(asy)
```

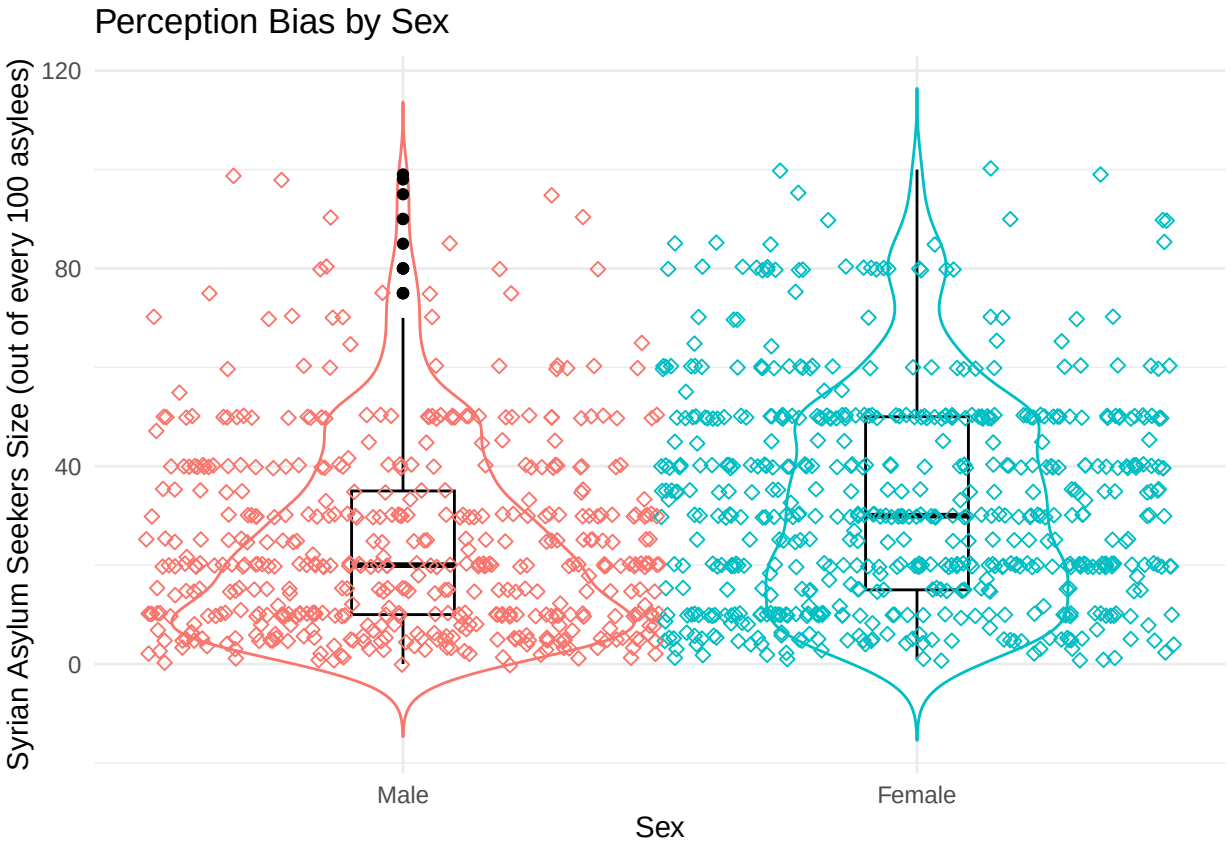
```
##      asylum           sex           age               news
##  Min.   :  0.00   Male  :478   Min.   :17.00   Don't read:573
##  1st Qu.: 10.00   Female:571   1st Qu.:36.00   Read      :476
##  Median : 25.00                Median :49.00
##  Mean   : 29.03                Mean   :49.75
##  3rd Qu.: 40.00                3rd Qu.:62.00
##  Max.   :100.00                Max.   :99.00
##
##    religious          urban          hhincome        partyid
##  Min.   :0.0000   Min.   :1.000   Min.   : 1.000   Tories:284
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.: 6.000   Labour:280
##  Median :0.0000   Median :3.000   Median : 9.000   SNP   : 16
##  Mean   :0.4929   Mean   :2.568   Mean   : 9.586   Greens: 23
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:13.000   Ukip  : 31
##  Max.   :1.0000   Max.   :4.000   Max.   :17.000   BNP   : 32
##                                                    other :383
```

```
##Explore the relationship between sex and asylum perception through a boxplot, using `ggplot`

p <- ggplot(asy,
            aes(x=factor(sex, labels = c("Male" , "Female"))
                , y=asylum, color=factor( sex , labels = c( "Male" , "Female")))) +
    geom_violin(trim=FALSE) + geom_boxplot(width=0.2, col = "black")
p
```




```
##Adding more layers (aesthetic specifications)
pt <- p + geom_jitter(shape=5,
                      position=position_jitter(0.5)) +
  labs(title="Perception Bias by Sex",x="Sex",
       y = "Syrian Asylum Seekers Size (out of every 100 asylees)") + scale_x_discrete(
         labels=c("1" = "Male", "2" = "Female")) +
  theme_minimal() + theme(legend.position = "none")
pt
```




```
##Observations:
#The distribution of perception bias for both sex groups are skewed.
#Female respondents have the highest bias in 1st/3rd quartiles and median.
#Male respondents have the lowest bias.

#Conducting T-test for Difference in Means and Hypothesis Testing
t.test(asy$asylum ~ asy$sex, mu = 0, alt = "two.sided", conf = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  asy$asylum by asy$sex
## t = -6.5034, df = 1041.8, p-value = 0.0000000001216
## alternative hypothesis: true difference in means between group Male and group Female is not equal to
## 95 percent confidence interval:
##  -10.74435  -5.76352
## sample estimates:
##   mean in group Male mean in group Female
##             24.53766             32.79159
```

```
#Interpretation
#In the following line you see the t-value, the degrees of freedom and the p-value.
#Knowing the t-value and the degrees of freedom you can check in a table on t distributions
#how likely you were to observe this data, if the null-hypothesis was true.
#The p-value gives you this probability directly. For example, a p-value of 0.01 would mean
```

```
#that the probability of seeing this data given that there is no difference in
#asylum seekers size perception between sex groups in the population, is 1%.
#In this case, our p-value is much smaller than this = 0.0000000001216!

#In the next line you see the 95% confidence interval because we specified conf=0.95.
#If you were to take 100 samples and in each you checked the means of the two groups,
#95 times the difference in means would be within the interval you see there.

#At the very bottom you see the means of the dependent variable by
#the two groups of the independent variable. These are the means that we estimated above.
```