Practical Exercise 3 | Statistics for Premasters DSS/CSAI

Hilal Caliskan Egilli_u927185

Part A - Cleaning a Dataset

Load the untidy_spanish_vow.csv data file. The file is included in the PE 3 zip folder. This includes the simulated data of vowel formants produced by male and female Spanish speakers. You can find more information about vowel formants here: https://www.britannica.com/science/phonetics/Vowel-formants.

```
spanishdata <- read.csv("/Users/hilalcaliskan/Documents/PM_DSS_ Assignments/CSAI/Practical Exercise 3/u
str(spanishdata)
```

```
## 'data.frame':    750 obs. of  4 variables:
##  $ label: chr  "p01-male-a" "p01-male-a" "p01-male-a" "p01-male-e" ...
##  $ rep  : int  1 2 3 1 2 3 1 2 3 1 ...
##  $ f1   : num  615 645 608 477 457 ...
##  $ f2   : num  1231 1282 1248 1612 1839 ...
```

```
summary(spanishdata)
```

```
##     label                rep          f1              f2        
##  Length:750         Min.   :1   Min.   :218.6   Min.   : 780.8  
##  Class :character   1st Qu.:1   1st Qu.:381.9   1st Qu.:1227.3  
##  Mode  :character   Median :2   Median :457.4   Median :1514.3  
##                     Mean   :2   Mean   :482.8   Mean   :1614.1  
##                     3rd Qu.:3   3rd Qu.:576.1   3rd Qu.:2046.5  
##                     Max.   :3   Max.   :813.1   Max.   :2733.3  
```

```
head(spanishdata)
```

```
##        label rep       f1       f2
## 1 p01-male-a   1 615.4477 1230.806
## 2 p01-male-a   2 644.6112 1281.965
## 3 p01-male-a   3 607.9174 1247.960
## 4 p01-male-e   1 476.9079 1612.076
## 5 p01-male-e   2 457.2205 1839.456
## 6 p01-male-e   3 444.6411 1848.639
```

```
tail(spanishdata)
```

```
##            label rep       f1       f2
## 745 p50-female-o   1 551.7140 1240.191
## 746 p50-female-o   2 577.1894 1310.138
## 747 p50-female-o   3 545.5014 1214.094
## 748 p50-female-u   1 405.7645 1491.935
## 749 p50-female-u   2 458.0345 1141.513
## 750 p50-female-u   3 457.4308 1181.657
```

```
nrow(spanishdata)
```

```
## [1] 750
```

```
#Your code here
```

Task 1a. Based on the past practical assignments and your newly acquired skills, inspect the data set and check out its variables. You can include any graphs that would allow you to get a better understanding of the data set.

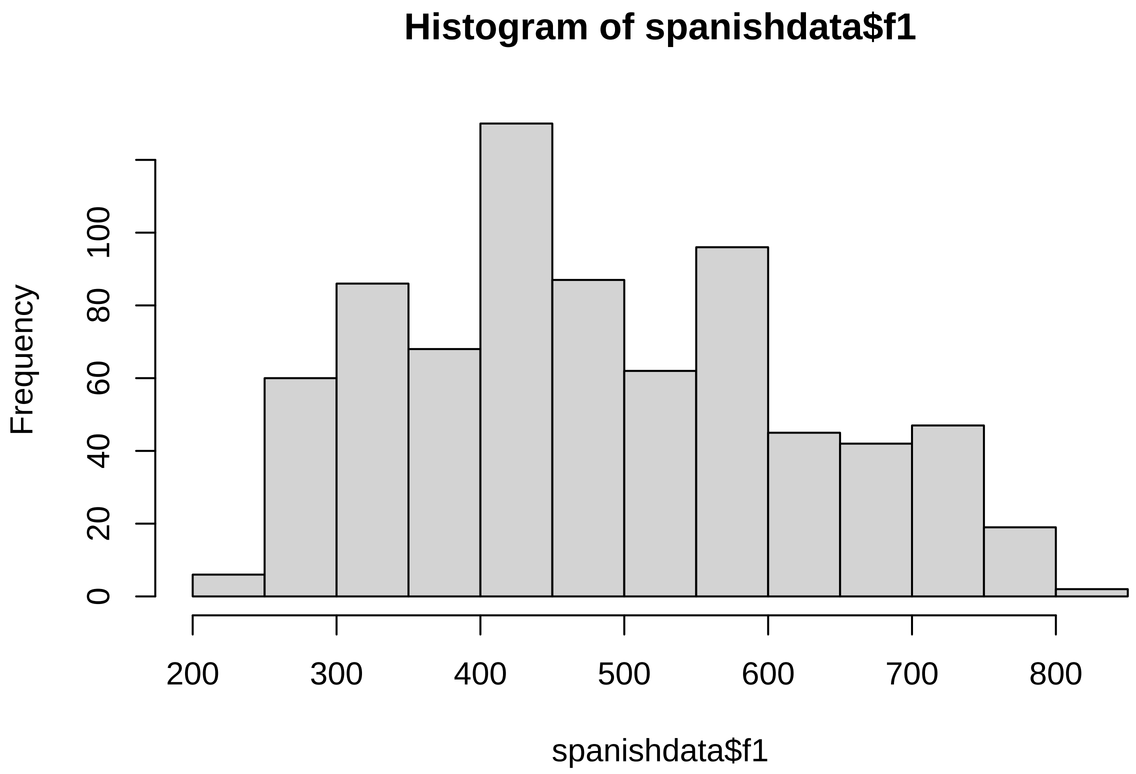
```
summary(spanishdata)
```

```
##     label                rep           f1              f2
##  Length:750         Min.   :1   Min.   :218.6   Min.   : 780.8
##  Class :character   1st Qu.:1   1st Qu.:381.9   1st Qu.:1227.3
##  Mode  :character   Median :2   Median :457.4   Median :1514.3
##                     Mean   :2   Mean   :482.8   Mean   :1614.1
##                     3rd Qu.:3   3rd Qu.:576.1   3rd Qu.:2046.5
##                     Max.   :3   Max.   :813.1   Max.   :2733.3
```

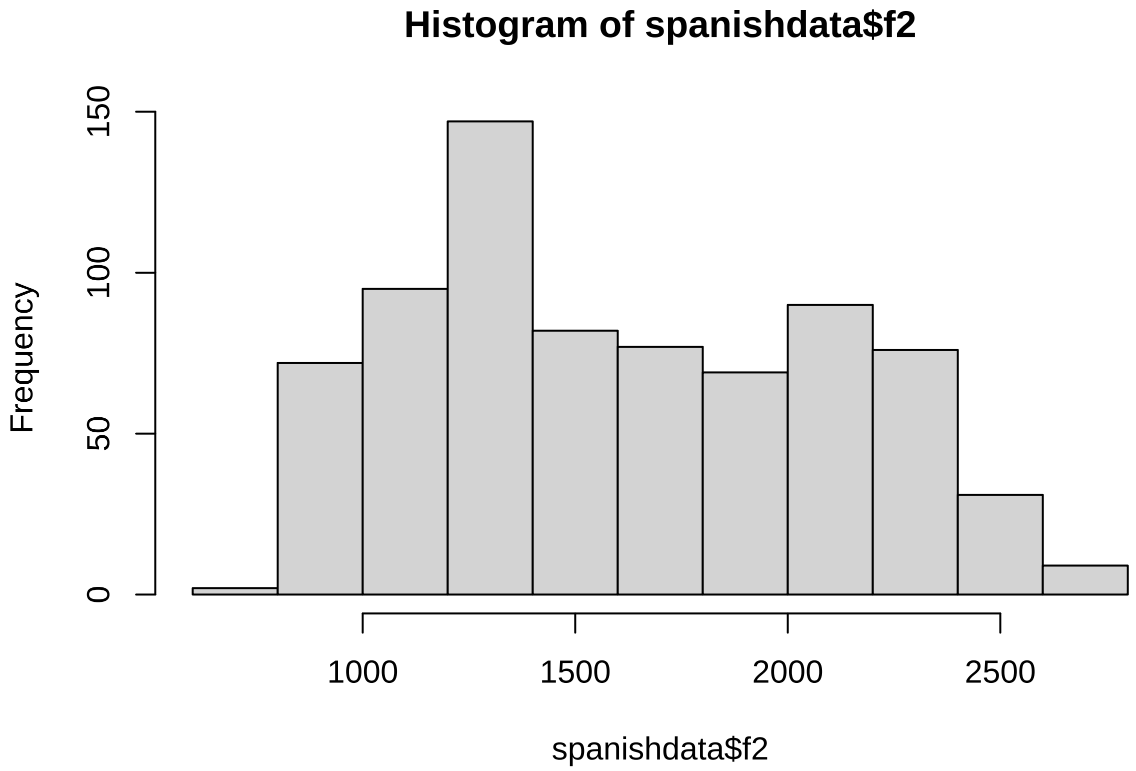
```
psych::describe(spanishdata)
```

```
##        vars   n    mean     sd  median trimmed    mad    min     max   range
## label*    1 750  125.50  72.22  125.50  125.50  92.66   1.00  250.00  249.00
## rep       2 750    2.00   0.82    2.00    2.00   1.48   1.00    3.00    2.00
## f1        3 750  482.76 137.86  457.43  476.34 156.31 218.60  813.14  594.54
## f2        4 750 1614.06 485.29 1514.30 1596.92 575.05 780.81 2733.29 1952.48
##        skew kurtosis    se
## label* 0.00    -1.20  2.64
## rep    0.00    -1.50  0.03
## f1     0.33    -0.78  5.03
## f2     0.29    -1.07 17.72
```

```
hist(spanishdata$f1)
```



```r
hist(spanishdata$f2)
```



Task 1b. Try to provide a brief description of the data set *(in your own words)*.

The “rep” column is saying how many times we did something. Then, “f1” and “f2” which are just numbers, probably measurements or similar. And names like “p01-male-a” are just labels for different groups or categories, maybe different people or situations we’re looking at. So, each row is like a combination of how many times we did it, which group it’s from, and what the measurements were. It’s probably from some kind of study or experiment.

Task 2. Based on the inspection, it seems that the “label” variable seems most problematic. Split it into three variables: id, gender, and vowel. Now examine the head() of the data and print it here.

HINT: *For this task, you can use the tidyr::separate function. The arguments you need are the data frame, the variable you want to split (label), the “into” argument which specifies the new variable names, and a “sep” argument, which tells the function where the splitting point should be (in this case ‘-’).*

```
spanishdata<-tidyr::separate(spanishdata,
                                label,
                                into=c("id","gender","vowel"),
                                sep="-")

head(spanishdata)
```

```
##    id gender vowel rep       f1       f2
## 1 p01   male     a   1 615.4477 1230.806
## 2 p01   male     a   2 644.6112 1281.965
## 3 p01   male     a   3 607.9174 1247.960
## 4 p01   male     e   1 476.9079 1612.076
## 5 p01   male     e   2 457.2205 1839.456
## 6 p01   male     e   3 444.6411 1848.639
```

Task 3. Now convert gender and vowel type into factors.

```
spanishdata$vowel <- as.factor(spanishdata$vowel)
spanishdata$gender <- as.factor(spanishdata$gender)

class(spanishdata$vowel)
```

```
## [1] "factor"
```

```
class(spanishdata$gender)
```

```
## [1] "factor"
```

Task 4. Use rbind() to convert the f1 and f2 variables to long format by combining the two columns. Add your code here.

HINT: *To use rbind(), you will need two copies of the data frame: one where you drop f1 and another where you drop f2. The columns being combined need to have the same name, so rename the remaining f column in both datasets to “frequency”. Now, you can use rbind() with the names of your two data frames.*

```
#Two data frames one with f1, and the other with f2.
f1 <- spanishdata[, -6]
f2 <- spanishdata[, -5]

#Rbind() requires the columns to have the same names,
#so both are given a generic name, such as "frequency".
```

```
colnames(f1)[5] <- "frequency"
colnames(f2)[5] <- "frequency"

#Now we can combine our two data frames into one:
spanishdata2 <- rbind(f1, f2)
```

Task 5. How many rows do you have in your data frame now?

```
nrow(spanishdata2)
```

```
## [1] 1500
```

Task 6. Create a recoded version of the variable that you converted to long format (frequency). The recoded variable should replace values of frequency with 'lower' if they are less than the mean of frequency, and 'higher' if they are greater than the mean or equal to the mean.

HINT: *After finding the mean of frequency, create a new column for the recoded frequency, where you convert values to 'lower'/'higher' based on a comparison to the mean. You can do this in two separate lines by using logical operators.*

For example, the following line would recode only values smaller than 10: data$new_column[data$old_column < 10] <- 'less_than_10'.

```
mean_span <- mean(spanishdata2$frequency)
spanishdata2$frequency_recode[spanishdata2$frequency<mean_span]<-"lower"
spanishdata2$frequency_recode[spanishdata2$frequency>=mean_span]<-"higher"
```

Task 7. How many recoded values are higher than the mean? And how many are lower than the mean?

```
table(spanishdata2$frequency_recode)
```

```
##
## higher  lower
##    651    849
```

Task 8. Imagine you want to conduct some research. Your experimental condition consists of administering the participants a placebo and then record their response times to a visual cue.

For the same experimental condition, you think about using either of the two strategies:

- Strategy A would be tak<ing a sample of 1000 participants.
- Strategy B would consist of taking a more restricted sample of only 100 participants.

In general (without considering other factors) which approach would you think would be more reliable? And what would be a good indicator of how much more reliable one approach is over the other? Explain in your own words.

Strategy A, with 1000 participants, would likely be more reliable than Strategy B with 100 participants. More participants generally mean more accurate results. Comparing the margin of error(**standard error**) between the two strategies can show how much more reliable Strategy A is.

Task 9. For the following claims, assign True or False and briefly explain your reasoning.

Given a sufficiently sized sample, which of the following are True?

Task 9a. With increasing sample size, the mean of the sampling distribution becomes a better estimate of the mean of the population.

True

Task 9b. The shape of the sampling distribution becomes less normal as the sample size increases.

False

Task 9c. The standard deviation of the sampling distribution gets bigger as the sample size increases.

False

Task 10. Imagine you run an experiment and have obtained the following values:

- Sample mean = 12.30
- The hypothesised population mean under the null hypothesis = 10.00
- Standard error = 1.5

What would be the test statistic in this case (e.g. the z-score)? Compute it and include the formula you used and offer a brief explanation of your thinking process.

$$\text{test statistic} = \frac{\text{estimate} - \text{value we hypothesise}}{\text{standard error}}$$

Part B: Z-tests

For this part we will be using a dataset which consists of 202 Australian athletes. Before you rush into the assignment, it is advised to start with functions that help you understand the data. It will make your life a lot easier if you are able to identify variables quickly. For more clarification on the variables: https://www.rdocumentation.org/packages/DAAG/versions/1.24/topics/ais

For this exercise, we are going to load the data using R. You can do so by installing and loading the package “DAAG”, and by using the function: “data(ais)”. After doing this, you will have a data frame named “ais” which you can either work with directly or rename as you wish.

```
#install.packages("DAAG")
data(ais)
```

```
## Warning in data(ais): data set ’ais’ not found
```

```
data<-(ais)
```

```
## Error in eval(expr, envir, enclos): object ’ais’ not found
```

```
head(data)
```

```
##
## 1 function (..., list = character(), package = NULL, lib.loc = NULL,
## 2     verbose = getOption("verbose"), envir = .GlobalEnv, overwrite = TRUE)
## 3 {
## 4     fileExt <- function(x) {
## 5         db <- grepl("\\\\.[^.]+\\\\.(gz|bz2|xz)$", x)
## 6         ans <- sub(".*\\\\.", "", x)
```

Task 11a. Compute the z-scores for athlete 163 (x = 163) for the following variables: body mass index, percentage body fat, lean body mass and height.

```
# The z-scores of the mentioned variables
data$bmiZ <- (data$bmi - mean(data$bmi))/ sd(data$bmi)
```

```
## Error in data$bmi: object of type 'closure' is not subsettable
```

```
data$pcBfatZ <- (data$pcBfat - mean(data$pcBfat))/ sd(data$pcBfat)
```

```
## Error in data$pcBfat: object of type 'closure' is not subsettable
```

```
data$lbmZ <- (data$lbm - mean(data$lbm))/ sd(data$lbm)
```

```
## Error in data$lbm: object of type 'closure' is not subsettable
```

```
data$htZ <- (data$ht - mean(data$ht))/ sd(data$ht)
```

```
## Error in data$ht: object of type 'closure' is not subsettable
```

```
# The z-scores for athlete
data$bmiZ[163] #4.002 - Body mass index
```

```
## Error in data$bmiZ: object of type 'closure' is not subsettable
```

```
data$pcBfatZ[163] #0.065 - Percentage of body fat
```

```
## Error in data$pcBfatZ: object of type 'closure' is not subsettable
```

```
data$lbmZ[163] #3.146 - Lean body mass
```

```
## Error in data$lbmZ: object of type 'closure' is not subsettable
```

```
data$htZ[163] #0.934 - Height
```

```
## Error in data$htZ: object of type 'closure' is not subsettable
```

Task 11b. Based on the observed z-scores what can you say about athlete 163?

Athlete 163 must be very muscular. He is quite heavy but has a low percentage of body fat.

Task 12a. Get the mean values for all numeric variables for athletes that engaged in the sport rowing.

HINT: *You can use colMeans() if you remove the non-numeric variables.*

```
#Make a subset that only includes rowers
Rowing <- data[data$sport == "Row", ]
```

```
## Error in data$sport: object of type 'closure' is not subsettable
```

```
#Remove non-numeric columns for colMeans,
#The - means we exclude the columns instead of including them
Rowing_numeric <- subset(Rowing, select = -c(sex, sport))
```

```
## Error in eval(expr, envir, enclos): object 'Rowing' not found
```

```
colMeans(Rowing_numeric)
```

```
## Error in eval(expr, envir, enclos): object 'Rowing_numeric' not found
```

Task 12b. Compare the mean values of those who engaged in the sport rowing, to those who did not. Report your findings with a 95% confidence interval (you can use lsr::ciMean() for this).

HINT: *The lsr::ciMean() function includes the 95% confidence interval by default.*

```
NonRowing <- data[data$sport != "Row", ]
```

```
## Error in data$sport: object of type 'closure' is not subsettable
```

```
lsr::ciMean(Rowing$rcc)
```

```
## Error in eval(expr, envir, enclos): object 'Rowing' not found
```

```
lsr::ciMean(NonRowing$rcc)
```

```
## Error in eval(expr, envir, enclos): object 'NonRowing' not found
```

```
lsr::ciMean(Rowing$wcc)
```

```
## Error in eval(expr, envir, enclos): object 'Rowing' not found
```

```
lsr::ciMean(NonRowing$wcc)
```

```
## Error in eval(expr, envir, enclos): object 'NonRowing' not found
```

```
lsr::ciMean(Rowing$hc)
```

```
## Error in eval(expr, envir, enclos): object 'Rowing' not found
```

```
lsr::ciMean(NonRowing$hc)
```

```
## Error in eval(expr, envir, enclos): object 'NonRowing' not found
```

Task 12c. What are the major differences between these two groups?

Look for non-overlapping regions.

Task 12d. Explain what would happen to the boundaries of the confidence interval if the number of respondents increases.

The boundaries would get smaller.

Task 12e. Explain what would happen to the boundaries of the confidence interval if variance increases.

The boundardies would get larger.

Task 13a. Compare the average BMI in the Netherlands of 25.4 (WHO, 2014) with the average BMI in the dataset. Use the standard deviation of the sample you currently have to perform a z-test.

HINT: *You can either make use of the z.test() function in the BSDA package, or use the formula for calculating a z-score.*

```
average_BMI <- 25.4
average_BMI_ais <- mean(data$bmi)
```

```
## Error in data$bmi: object of type ’closure’ is not subsettable
```

```
sd_BMI_ais <- sd(data$bmi)
```

```
## Error in data$bmi: object of type ’closure’ is not subsettable
```

```
average_BMI
```

```
## [1] 25.4
```

```
average_BMI_ais
```

```
## Error in eval(expr, envir, enclos): object ’average_BMI_ais’ not found
```

```
z <- (average_BMI_ais - average_BMI) / (sd_BMI_ais/ sqrt(nrow(ais)))
```

```
## Error in eval(expr, envir, enclos): object ’average_BMI_ais’ not found
```

```
z
```

```
## Error in eval(expr, envir, enclos): object ’z’ not found
```

Task 13b. What can you conclude based on the test statistic?

That the average BMI of Dutch people is most likely different than the BMI of the Australian athletes.

Task 13c. *Null: “There is no difference in BMI between Dutch citizens and Australian athletes.”*

How do we treat the null hypothesis mentioned above, and why?

Based on the test-statistic (-12.12), we can reject the null hypothesis because using a 95% confidence interval, the calculated z-score falls outside of the zone between the critical values of -1.96 and 1.96.