Practical Exercise 1 | Statistics for Premasters DSS/CSAI

Hilal Caliskan u927185

Part A - Knowledge to Discuss Statistical and Methodological Concepts

Task 1. For each of the following, identify the scale of measurement and *briefly explain your reasoning.* :

Task 1a. Temperature in degree Celsius

Temperature degree is measured on a continuous interval scale because the variable does not have a true 0 value.

Task 1b. Gender

Gender has a nominal scale of measurement because gender has more values than two and none of them are better or worse than the other.

Task 1c. The order that runners cross the finish line in a marathon race

The order that runners cross the finish line is measured on an ordinal scale because there are more than two categories but they have a logical order.

Task 1d. The amount of time Bob took to solve a calculus problem

The amount of time Bob took to solve a calculus problem should be measured with a ratio/interval scale. The numerical value is genuinely meaningful. The differences between the numbers are interpretable, and zero really means zero; therefore, it is fine to multiply and divide the values.

Task 2. Apply each of the following types of validity to the case, described above. You are encouraged to use your imagination.

Task 2a. External validity

The extent to which we can generalize the results of this study to the real population, namely young Dutch women between 20-25 years old with depression issues. When for example just using undergraduate psychology students with small mental issues as the participants, your sample has a high risk of not representative of the population. This experiment will carry a risk of lacking external validity.

Task 2b. Internal validity

Internal validity depends on the time. If the study results are examined based on the timeframe since it is assumed that the experiment is a study that measures the relationship between cause and effect, the internal validity is fine.

Task 2c. Construct validity

Construct validity is to show that the test is actually measuring what you want to be measuring. If you are trying to examine the effect of cognitive therapy on depression, the research should have the tools and the methods in order to somehow measure the concept of “depression”.

Task 2d. Ecological validity

The entire set up of the study should closely approximate the real world scenario that is being investigated. In the above mentioned case, this should mean that the therapy session that is considered to be the independent variable, should approximate “real world” therapy sessions. This also applies to the environment where the participants are questioned about the change in depression levels.

Task 3. For each of the following types of reliability, give your own example scenario that would demonstrate that reliability, and briefly explain how it does so. You do not have to limit yourself to the given case. You are allowed to use your own examples.

Task 3a. Inter-rater reliability

Tom and Edith are two judges measuring 50 speed skaters' time in a short track speed skating competition. If the results of the two judges were very similar, the results showed an excellent inter-rater reliability.

Task 3b. Test-retest reliability

Test-retest reliability is a measure of reliability obtained by administering the same test twice over a period of time to a group of individuals. The scores from Time 1 and Time 2 can then be correlated in order to evaluate the test for stability over time. For example, a group of respondents is tested for IQ scores: each respondent is tested twice – the two tests are, say, a month apart. Then, the correlation coefficient between two sets of IQ-scores is a reasonable measure of the test-retest reliability of this test.

Task 3c. Internal consistency reliability

Internal consistency reliability is a measure to assess the consistency of responses across items within a test or scale. It evaluates how well the items or questions in a measure are interrelated or correlated with each other.For example, if you have a survey about job satisfaction, internal consistency reliability ensures that all the questions about different aspects of the job, like work hours, relationships with colleagues, and opportunities for growth, are consistent and reliable in measuring overall job satisfaction.

Task 3d. Parallel forms reliability

Parallel forms reliability: An experimenter developed a large set of word memory questions (i.e., list of words). He split these questions into half, and administered them to a randomly selected half of a target sample. If the results of the two sets of questions show a high correlation, this would be one indicator that the tests have a good parallel forms reliability.

Part B: Knowledge and Skills Using R

Task 4a. Make two vectors with values: vector1 = (1,2,3,4,5,6) and vector2 = (2,2,3,3,-1,-1). After you have made these vectors, try to add these up, and explain what is happening.

```
vector1 <- c(1, 2, 3, 4, 5, 6)
vector2 <- c(2, 2, 3, 3, -1, -1)
result <- vector1 + vector2
print("Vector1:")
```

```
## [1] "Vector1:"
```

Task 4b. Make two vectors with values: vector1 = (1,2,3,4,5,6) and vector2 = (1,2). After you have made these vectors, try to add these up, and explain what is happening.

```
vector1 <- c(1,2,3,4,5,6,)
```

```
## Error in c(1, 2, 3, 4, 5, 6, ): argument 7 is empty
```

```
vector2 <- c(1,2)
vector1 + vector2
```

```
## [1] 2 4 4 6 6 8
```

Task 4c. What do you think would happen if vector1 got an extra value, for example (1,2,3,4,5,6,7). Repeat the process of Task 4b, and describe what has happened, and why.

```
vector1 <- c(1,2,3,4,5,6,7)
vector2<- c(1,2)
vector1 + vector2
```

```
## Warning in vector1 + vector2: longer object length is not a multiple of shorter
## object length
```

```
## [1] 2 4 4 6 6 8 8
```

Task 5. Create a factor with the food you ate at dinner last week; please assume you ate something more than once, thus, repeat it.

HINT: *Use factor(x, labels).*

```
foodslastweek <- factor( x= c(1,2,3,4,4,5,6),
                         labels = c('mayo', 'ketchup', 'egg', 'boerenkool', 'spaghetti', 'pizza'))

                         foodslastweek
```

```
## [1] mayo       ketchup    egg        boerenkool boerenkool spaghetti  pizza
## Levels: mayo ketchup egg boerenkool spaghetti pizza
```

Task 6. Create a simple Body-Mass-Index (BMI) calculator using R code, and put in some values for weight and height to play around with. The formula for calculating BMI is: weight in kilograms / (height in meters)^2

```
weight <- 100

height <- 185

  height_in_meters <- height * 0.01

bmi <- weight / height_in_meters ^ 2

bmi
```

```
## [1] 29.21841
```

Part C: Looking at Descriptives in R

Task 7. Load data from the following comma-separated-value (csv) file using read.csv(): Statistics_survey_PMBA_2017_PU2_reduced.csv. The file is included in the PE 1 zip folder you downloaded from Canvas. Take a quick look at the data using summary(), and try to get a basic understanding of what the data is about.

HINT: *Add "stringsAsFactors=TRUE" as a parameter to the read.csv() function and run this block again if your summary() output is not as expected*

```
dtvfile <- read.csv('/Users/hilalcaliskan/Documents/PM_DSS_ Assignments/CSAI/Practical Exercise 1/Statis
```

Task 8. Inspect the loaded file by using the str() function to show a summary of the data object's structure.

```
str(dtvfile)
```

```
## 'data.frame':    94 obs. of  7 variables:
##  $ age        : int  21 22 26 34 21 23 22 23 25 22 ...
##  $ gender     : chr  "female" "female" "female" "female" ...
##  $ bachelor   : chr  "media" "language" "other" "media" ...
##  $ nationality: chr  "dutch" "italian" "dutch" "dutch" ...
##  $ inhabitants: num  8.0e+03 1.2e+04 1.0e+01 1.5e+02 9.0e+03 8.8e+03 3.0 2.0e+07 2.0e+05 3.0e+04 ...
##  $ drinks     : num  0 5 5 3 4 30 4 3 0 12 ...
##  $ lextale    : num  82.5 70 88.8 78.8 71.2 ...
```

Task 9. Inspect the data by looking at the first few entries and the last few entries in the dataset. Use the function head() which shows the first N rows of the dataframe. Use the tail() function that shows the last N rows.

```
head(dtvfile)
```

```
##   age gender bachelor nationality inhabitants drinks lextale
## 1  21 female    media       dutch        8000      0   82.50
## 2  22 female language     italian       12000      5   70.00
## 3  26 female    other       dutch          10      5   88.75
## 4  34 female    media       dutch         150      3   78.75
## 5  21 female    media       dutch        9000      4   71.25
## 6  23 female business       dutch        8800     30   80.00
```

```
tail(dtvfile)
```

```
##    age gender bachelor nationality inhabitants drinks lextale
## 89  26   male business       other      200000      6   92.50
## 90  20   male     none       dutch       10000      5   61.25
## 91  26   male  culture       dutch          25     15   77.50
## 92  19   male     none       dutch          10     15   63.75
## 93  25   male    other       dutch      225000      8   61.25
## 94  24   male  culture       dutch          10     15   77.50
```

Task 10a. Determine the mean value for all numeric variables.

HINT: *If you run into issues with missing values, use the na.rm = TRUE parameter for mean()*

```
means <- c( mean(dtvfile$age, na.rm = TRUE ),
            mean(dtvfile$inhabitants, na.rm = TRUE ),
            mean(dtvfile$drinks, na.rm = TRUE ),
            mean(dtvfile$lextale, na.rm = TRUE))

means
```

```
## [1] 2.240426e+01 5.922995e+05 5.758065e+00 7.434840e+01
```

```
round(means,2)
```

```
## [1]     22.40 592299.49      5.76     74.35
```

Task 10b. Explain why it is possible that the mean value of some variables is not a value that was observed in the set of values?

The formula for mean is: sum of element / count of elements.

Task 10c. Determine the standard deviation of two variables of your liking.

HINT: *Removing missing values works the same way as it does for mean.*

```
sdeviation <- c(sd(dtvfile$age, na.rm = TRUE),
                sd(dtvfile$inhabitants, na.rm = TRUE))

round(sdeviation, 2)
```

```
## [1]       3.88 3156889.73
```

Task 11a. Make a subset of data where only those who drank more than fifteen beers are stored, and call this dataframe: "heavy_drinkers". After doing so, make a second subset in which only the ages of participants that are considered heavy drinkers are stored, and call this subset "heavy_drinkers_age".

```
heavy_drinkers <- dtvfile[dtvfile$drinks > 15, ]
heavy_drinkers_age <- heavy_drinkers$age

heavy_drinkers_age
```

```
## [1] 23 NA 19 21 22 27 24
```

Task 11b. What is the mean age of participants who drank more than fifteen drinks?

22.66667

Task 11c. What is the mean age difference between participants that drank more than fifteen drinks, and those who had less than or equal to fifteen drinks? Round your answer to two decimal places.

```
heavy_drinkers <- dtvfile[dtvfile$drinks > 15, ]
slow_drinkers <- dtvfile[dtvfile$drinks <= 15, ]
heavy_drinkers_age <- heavy_drinkers$age
slow_drinkers_age <- slow_drinkers$age
age_difference <- mean(heavy_drinkers_age, na.rm = TRUE) - mean(slow_drinkers_age, na.rm = TRUE)
round(age_difference, 2)
```

```
## [1] 0.29
```

Task 11d. Compare the standard deviations for those who are not heavy drinkers to the standard deviation of all numeric variables in the "heavy_drinkers" variable. What do you notice, and can you explain what has happened?

```r
heavy_drinkers <- dtvfile[dtvfile$drinks > 15, ]
slow_drinkers <- dtvfile[dtvfile$drinks <= 15, ]
sd(heavy_drinkers$age, na.rm = TRUE) - sd(slow_drinkers$age, na.rm = TRUE)
```

```
## [1] -1.241937
```

```r
sd(heavy_drinkers$inhabitants, na.rm = TRUE) - sd(slow_drinkers$inhabitants, na.rm = TRUE)
```

```
## [1] -3046579
```

```r
sd(heavy_drinkers$drinks, na.rm = TRUE) - sd(slow_drinkers$drinks, na.rm = TRUE)
```

```
## [1] -0.3064414
```

```r
sd(heavy_drinkers$lextale, na.rm = TRUE) - sd(slow_drinkers$lextale, na.rm = TRUE)
```

```
## [1] -1.660056
```

Task 12a. Use the psych::describe() function to get more descriptive statistics

```r
psych::describe(dtvfile)
```

```
##              vars  n      mean         sd  median  trimmed      mad   min
## age             1 94     22.40       3.88   22.00    21.87     1.48 18.00
## gender*         2 94      1.22       0.42    1.00     1.16     0.00  1.00
## bachelor*       3 94      3.79       1.64    4.00     3.86     1.48  1.00
## nationality*    4 94      2.23       0.72    2.00     2.11     0.00  1.00
## inhabitants     5 94 592299.49 3156889.73 9500.00 31052.91 14003.16  3.00
## drinks          6 93      5.76       6.27    4.00     4.74     5.93  0.00
## lextale         7 94     74.35      11.02   73.12    73.82    10.19 53.75
##                   max     range  skew kurtosis        se
## age                48 3.000e+01  3.42    18.61      0.40
## gender*             2 1.000e+00  1.31    -0.29      0.04
## bachelor*           6 5.000e+00 -0.49    -0.87      0.17
## nationality*        4 3.000e+00  1.63     1.89      0.07
## inhabitants  23010000 2.301e+07  6.30    39.27 325608.31
## drinks             30 3.000e+01  1.47     2.05      0.65
## lextale           100 4.625e+01  0.40    -0.57      1.14
```

Task 12b. Determine if any of the variables might be skewed. If so, indicate which variables are.

All of the variables are skewed they all have either a negative or a positive value. Therefore it is accurate to say that all of the variables have a skewness value.

Task 12c. Summarize in your own words what it means for a variable to be skewed.

The meaning of a variable being positively skewed means outliers are located on the right side of the table. A long tail to the right also shows that the variable has a positive skewness value.

Task 13a. Try to get the descriptive statistics per gender using another function from the psych package.

```r
psych::describeBy(dtvfile, group = dtvfile$gender)
```

```
##
##  Descriptive statistics by group
## group: female
##              vars  n      mean         sd  median  trimmed      mad   min
## age             1 73     22.16       4.13   22.00    21.51     1.48 18.00
## gender          2 73      1.00       0.00    1.00     1.00     0.00  1.00
## bachelor        3 73      3.86       1.56    4.00     3.95     1.48  1.00
## nationality     4 73      2.21       0.73    2.00     2.08     0.00  1.00
## inhabitants     5 73 728718.59 3574546.02 8000.00 21107.31 11638.41  3.00
## drinks          6 72      4.94       6.02    3.00     3.82     4.45  0.00
## lextale         7 73     73.42      10.85   71.25    72.75     9.27 53.75
##                   max     range  skew kurtosis        se
## age                48 3.000e+01  3.71    19.25      0.48
## gender              1 0.000e+00   NaN      NaN      0.00
## bachelor            6 5.000e+00 -0.60    -0.56      0.18
## nationality         4 3.000e+00  1.61     2.05      0.08
## inhabitants  23010000 2.301e+07  5.47    29.27 418368.97
## drinks             30 3.000e+01  1.96     4.35      0.71
## lextale           100 4.625e+01  0.54    -0.30      1.27
## ------------------------------------------------------------
## group: male
##              vars  n      mean        sd  median  trimmed      mad   min
## age             1 21     23.24      2.74    22.0    23.18     2.97 19.00
## gender          2 21      2.00      0.00     2.0     2.00     0.00  2.00
## bachelor        3 21      3.52      1.91     4.0     3.53     2.97  1.00
## nationality     4 21      2.33      0.73     2.0     2.18     0.00  2.00
## inhabitants     5 21 118080.71 197636.33 48000.0 66451.47 56338.80 10.00
## drinks          6 21      8.57      6.45     7.0     8.24     7.41  0.00
## lextale         7 21     77.56     11.25    77.5    77.57    11.12 58.75
##                   max     range  skew kurtosis       se
## age              29.0     10.00  0.31    -0.91     0.60
## gender            2.0      0.00   NaN      NaN     0.00
## bachelor          6.0      5.00 -0.12    -1.62     0.42
## nationality       4.0      2.00  1.66     0.94     0.16
## inhabitants  750000.0 749990.00  2.18     3.71 43127.78
## drinks           20.0     20.00  0.32    -1.31     1.41
## lextale          97.5     38.75 -0.05    -1.11     2.46
```

Task 13b. Which gender drank more?

Male participants drank more.

Task 13c. Describe what you notice regarding the kurtosis per group, and explain what this means.

Females seem to have higher kurtosis on average upon their variables. Which means that the values in those variables are more centered around the mean, which results in a very "pointy" histogram.

Task 14. Based on the past exercises and your newly acquired skills, fill in the blanks.

In total, 94 respondents from 3 different countries (China, Netherlands and Italy; 12 respondents were from other countries) completed the survey (21 males and 73 females). Their mean age was 22.4 years (SD = 3.87). Of these respondents, 12 did a language related bachelor, 27 did a media related bachelor, and 55 did a different bachelor (or did not do a bachelor yet). On average, respondents drink 5.76 units of alcohol

per week (SD = 6.27) and their mean score on the LexTale test of English was 74.35 (SD = 11.02). 21 respondents do not drink any alcohol. Male respondents drink more units of alcohol per week (M = 8.57, SD = 6.45) than female respondents (M = 4.94, SD = 6.02)