Practical Exercise 1 | Statistics for CSAI II

Burcu Ibicioglu, u986202

The goals of this exercise are to (a) to review specific statistical and methodological concepts (i.e., types of variables, reliability and validity), (b) to use R for some basic probability and sampling theory questions, and c) generate some and test some hypotheses using correlation analysis.

Part A - Knowledge to Discuss Statistical and Methodological concepts

In your own words, briefly (a few sentences) describe the difference between the following:

Task 1. Probability and statistics

Probability predicts the likelihood of the future events whereas statistics deals with the analysis of the frequency of past events.

Task 2. Sample statistics and population parameters

A parameter is measure describing an entire population whereas the sample statistics deals with a sample population drawn from the original population.

Task 3. Hypothesis and research question

A research question is a question that presents a problem to be addressed in a research study. Hypothesis, on the other hand, is a hypothetical outcome to that research and seeks to be proven or disproven.

Task 4. Null hypothesis and alternative hypothesis

Null hypothesis states that there are no statistically significant relation between two groups/variables, whereas the alternative hypothesis indicates that there are some statistically significant relation between the two.

Task 5. The alpha level and the p-value

Alpha is the is the significance level and the p-value is the probability of rejecting the null hypothesis.

Task 6. A one-sided and two-sided test

A onn-sided test will only have one critical point. The two-sided tests extend to both the sides (positive and negative) resulting in two criticial points.

Task 7. Type I and type II errors

Type I errors (also known as false positives), occur when we mistakenly reject a true null hypothesis. Type II errors, or false negatives, happen when we fail to reject a false null hypothesis.

Task 8. Correlation and causation

Correlation means the statistical relation between two variables, whereas causation indicates that changing one variable will causes changes in the other variable.

Part B: Using R for Basic Probability and Sampling Theory Questions

Task 9. I am going out to a restaurant for sushi, but the sushi I receive is totally up to the chef. He is choosing from salmon, tuna, avocado, eel, krab, or tofu. My favorite is salmon. I go crazy and eat 50 pieces of sushi. What is the probability that 30 of the pieces of sushi I receive are salmon?

```
dbinom (x = 30, size = 50, prob = 1/6)
```

```
## [1] 5.560678e-12
```

```
#your code here
```

Task 10. Create a variable that contains all the letters of the alphabet to represent your population.

```
alphabet <- letters
```

Task 10)a) Draw 20 random samples of size 10 without replacement and paste the results of your samples.

```
sampl = sample(alphabet, 20)
sampl
```

```
##  [1] "b" "m" "j" "u" "r" "v" "n" "t" "e" "c" "a" "x" "w" "g" "l" "y" "h" "i" "p"
## [20] "s"
```

Task 10)b) Add a bias to your sampling such that it only draws vowels from the alphabet and then draw 20 samples of size 10 and paste them.

```
vowels <- c("a", "e", "i", "o", "u")
sampl <- replicate ( 20, sample (vowels , 10, replace = TRUE))
sampl
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
##  [1,] "e"  "a"  "e"  "u"  "u"  "o"  "i"  "e"  "o"  "o"   "o"   "u"   "i"
##  [2,] "a"  "o"  "u"  "o"  "a"  "u"  "i"  "e"  "e"  "u"   "o"   "i"   "u"
##  [3,] "u"  "a"  "u"  "i"  "o"  "a"  "i"  "o"  "a"  "u"   "a"   "i"   "o"
##  [4,] "e"  "u"  "o"  "a"  "a"  "e"  "i"  "a"  "o"  "e"   "o"   "e"   "e"
##  [5,] "u"  "a"  "i"  "u"  "a"  "a"  "a"  "a"  "o"  "o"   "o"   "u"   "i"
##  [6,] "u"  "o"  "e"  "u"  "i"  "o"  "a"  "i"  "u"  "o"   "o"   "u"   "e"
##  [7,] "i"  "a"  "e"  "a"  "e"  "u"  "o"  "u"  "a"  "o"   "i"   "a"   "u"
##  [8,] "i"  "e"  "o"  "i"  "o"  "e"  "a"  "u"  "e"  "i"   "o"   "u"   "i"
##  [9,] "i"  "e"  "o"  "i"  "a"  "u"  "e"  "e"  "u"  "u"   "o"   "a"   "u"
## [10,] "i"  "e"  "a"  "a"  "a"  "o"  "o"  "a"  "a"  "a"   "e"   "o"   "e"
##       [,14] [,15] [,16] [,17] [,18] [,19] [,20]
##  [1,] "i"   "o"   "a"   "i"   "i"   "i"   "i"
##  [2,] "a"   "a"   "u"   "e"   "e"   "u"   "u"
##  [3,] "u"   "a"   "a"   "u"   "u"   "e"   "e"
##  [4,] "e"   "o"   "i"   "u"   "i"   "i"   "u"
##  [5,] "o"   "e"   "u"   "i"   "i"   "i"   "i"
##  [6,] "u"   "u"   "e"   "i"   "o"   "e"   "a"
##  [7,] "i"   "u"   "a"   "u"   "i"   "o"   "a"
##  [8,] "u"   "a"   "i"   "o"   "e"   "i"   "e"
##  [9,] "u"   "i"   "i"   "u"   "i"   "a"   "u"
## [10,] "i"   "i"   "e"   "o"   "a"   "o"   "o"
```

Task 11. Generate a variable that contains the mean of 10,000 samples of size 200 from a normal distribution with a mean of 30 and a standard deviation of 5. What is the mean value of your sampling distribution of the mean? Now perform the same, but change your sample size to 20. What is the mean now? Why are these different?

```
m1 =  mean(replicate (10000, mean (rnorm(200, mean=30, sd = 5))))
m1
```

```
## [1] 30.00165
```

```
m2 = mean(replicate(10000, mean(20, mean=30, sd = 5)))
m2
```

```
## [1] 20
```

- The first sample with a bigger population represent a more accurate mean since it tends to provide more accurate estimates of the population mean because they average out random fluctuations.
- The second sample, on the other, gives us a less accurate overview of the population mean since it contains more variation around the mean.

Part C: Using R for Correlation Analyses

For this part of Practical Exercise #1, Tasks indicate things that you need to complete in R/R Studio.

Task 12. Load data from the "The World Almanac and Book of Facts 1993" (OEF8_dataset.csv)".

```
data<- read.csv ('OEF8_dataset.csv')
#your code here
```

Task 13. 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 data frame. Use the tail() function that shows the last N rows. You can also open the full dataset and check it out. Be sure to take note of the variables. They include information about the participant's country, life expectancy, number of people per television, the number of people per physician, and then life expectancy for females and males.

```
head(data)
```

```
##           country life_exp ppl_television ppl_physician female_life male_life
## 1        Ethiopia     51.5            503         36660          53        50
## 2        Tanzania     52.5             NA         25229          55        50
## 3           Sudan     53.0             23         12550          54        52
## 4      Bangladesh     53.5            315          6166          53        54
## 5           Zaire     54.0             NA         23193          56        52
## 6 Myanmar (Burma)     54.5            592          3485          56        53
```

```
tail(data)
```

```
##           country life_exp ppl_television ppl_physician female_life male_life
## 35 United Kingdom     76.0            3.0           611          79        73
## 36         Canada     76.5            1.7           449          80        73
## 37         France     78.0            2.6           403          82        74
## 38          Italy     78.5            3.8           233          82        75
## 39          Spain     78.5            2.6           275          82        75
## 40          Japan     79.0            1.8           609          82        76
```

```
summary(data)
```

```
##    country             life_exp     ppl_television   ppl_physician
##  Length:40          Min.   :51.50   Min.   :  1.30   Min.   :  226.0
##  Class :character   1st Qu.:61.00   1st Qu.:  3.35   1st Qu.:  472.2
##  Mode  :character   Median :69.50   Median :  6.30   Median :  990.5
```

```
##                     Mean   :67.04   Mean   : 51.98   Mean   : 3997.7
##                     3rd Qu.:73.38   3rd Qu.: 23.00   3rd Qu.: 3193.2
##                     Max.   :79.00   Max.   :592.00   Max.   :36660.0
##                                     NA's   :2
##   female_life      male_life
##  Min.   :53.00   Min.   :50.00
##  1st Qu.:63.00   1st Qu.:59.75
##  Median :72.00   Median :66.00
##  Mean   :69.58   Mean   :64.50
##  3rd Qu.:77.25   3rd Qu.:69.50
##  Max.   :82.00   Max.   :76.00
##
```

Task 14. State a null hypothesis and an alternative hypothesis about the correlation between life expectancy and the number of televisions per person.

- N0: There is no affect of the number of televisions per person on the life expectancy.
- H1: There is a correlation between the two. Number of televisions is actually affecting the life expectancy.

Task 15. Run a correlation test in R to test your hypothesis.

```
corr <- cor.test (data$ppl_television, data$life_exp)
corr
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$ppl_television and data$life_exp
## t = -4.5691, df = 36, p-value = 5.561e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7753592 -0.3549002
## sample estimates:
##       cor
## -0.6058468
```

Task 16. Report the results here. Be sure to include the r-value, the p-value, and the 95% confidence interval. What can you conclude from your results? Do these results seem plausible to you?

A Pearson correlation analysis revealed a positive correlation between life expectancy and the number of televisions per person, r = -0.6, p = 5.5. The 95% confidence interval for the correlation was [-0.77, -0.35]

- since the p-value is a lot than the significance level 0.05, we can reject the alternative hypothesis. The outcome is likely due chance or other factors.

#to get text with subscript $text_{withsubscript}$

Task 17. State a null hypothesis and an alternative hypothesis about the correlation between life expectancy and the number of physicians per person.

N0: There is no correlation of affect between life expectancy and the number of physicians per person. H1: There is a correlation between the two. The number of physicians per person does have an effect on life expectancy.

Task 18. Run a correlation test in R to test your hypothesis.

```
corr1 = cor.test(data$ppl_television, data$ppl_physician)
corr1
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  data$ppl_television and data$ppl_physician
## t = 4.7377, df = 36, p-value = 3.34e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3741546 0.7840642
## sample estimates:
##       cor
## 0.6197134
```

Task 19. Report the results here. Be sure to include the r-value, the p-value, and the 95% confidence interval. What can you conclude from your results?

A Pearson correlation analysis revealed a positive correlation between life expectancy and the number of televisions per person, r = 0.61, p = 3.34. The 95% confidence interval for the correlation was [0.37, 0.78]

- since the p-value is more than the significance level 0.05, we can again reject the alternative hypothesis.