MS-C1620 Statistical Inference

Exercise 4

Homework exercise

To be solved at home before the exercise session.

---

1. a. Let $x_1, \dots, x_n$ be a random sample (iid) from some distribution F_θ with the unknown parameter θ. Which of the three one-sample tests (t-test, sign test or signed rank test) would you use (and why!) to test whether the location (expected value/median) of the data is equal to 1 if we know for certain that the distribution F_θ is
      i. an exponential distribution with unknown rate parameter θ,
      ii. a normal distribution with variance 2 and unknown expected value θ,
      iii. a Laplace distirbution with known scale parameter 5 and unknown location parameter θ,
      iv. a Poisson distirbution with unknown parameter θ?

   b. The data set `airmiles` lists the passenger miles flown by commercial airlines in the United States for each year from 1937 to 1960. To inspect whether the yearly passenger miles equal 10000 on average, a researcher performed a sign test to test the null hypothesis $med_x = 10000$ on significance level 5% with the results shown below and concluded that there is no evidence against the null hypothesis. Do you agree with the researcher's conclusion?

```
airmiles
```

```
## Time Series:
## Start = 1937 
## End = 1960 
## Frequency = 1 
##  [1]   412   480   683  1052  1385  1418  1634  2178  3362  5948  6109
## [12]  5981  6753  8003 10566 12528 14760 16769 19819 22362 25340 25343
## [23] 29269 30514
```

```
# Sign test
binom.test(sum(airmiles > 10000), length(airmiles))
```

```
## 
##  Exact binomial test
## 
## data:  sum(airmiles > 10000) and length(airmiles)
## number of successes = 10, number of trials = 24, p-value = 0.5413
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2210969 0.6335694
## sample estimates:
## probability of success 
##              0.4166667
```

Class exercise

To be solved at the exercise session.

Note: all the needed data sets are either given below or available in base ***R****.*

---

1. The data set `sleep` shows the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients. We are interested in studying whether drug 1 helps in increasing the number of hours slept compared to placebo.
   a. Extract the increases in hours of sleep of the patients who received drug 1 (`group == 1`).
   b. Visualize the data.
   c. Conduct an appropriate test to evaluate whether the location (expected value/median) of the increase in hours of sleep differs from 0 on significance level 5%.
   d. Draw conclusions.

---

2. The data set below contains the annual salaries (in dollars) of 8 American women and 8 American men (recall exercise 3.2). The observations are paired such that each woman is matched with a man having similar background (age, occupation, level of education, etc). We are interested in studying whether the locations of the salaries of women and men differ (recall that last time paired t-test concluded that the salaries differ) .
   a. Begin again by visualizing the data.
   b. Which two non-parametric tests are appropriate in studying our question of interest?
   c. State the hypotheses of the tests and conduct them on the significance level 10%.
   d. What are the conclusions of the tests?
   e. What assumptions did the test in part c make? Are they justifiable?

```
salary <- data.frame(women = c(42600, 43600, 49300, 42300, 46200, 45900, 47500, 41300),
                     men = c(46200, 44700, 48400, 41700, 48600, 49300, 48300, 44300))
```

---

3. Eight female Aalto students and twelve male Aalto students were chosen randomly and lined up based on their heights (shortest first). The sexes (female/male) in the line have the pattern shown below. Use two-sample rank test to study whether the median height of females differs from the median height of males on significance level 5%.
   a. Write down the assumptions and the hypotheses of the test.
   b. Do you think the assumptions are plausible in this case?
   c. Create two new vectors, `female` which contains the ranks of the females in the line and `male` which contains the ranks of the males in the line.
   d. Conduct the two-sample rank test and draw conclusions.

```
line <- c("F", "F", "M", "M", "F", "M", "F", "F", "M", "F", "M", "F", "M", "M", "M", "F", "M", "M", "M", "M")
```

---

4. **(Optional)** Data manipulation using just functions in base R does not always produce the most readable code. The task in 1a. can be achieved more transparently using the package `dplyr` as follows.

```
# install.packages("dplyr")
library(dplyr)
```

```
sleep_1 <- sleep %>%
  filter(group == 1) %>%
  select(extra)
```

Find out how the package and the piping operator `%>%` work by going through an online tutorial.