MS-C1620 Statistical Inference

Exercise 3

Homework exercise

To be solved at home before the exercise session.

---

1. Consider the confidence interval for the expected value of the normal distribution on page 2.9 of the lecture notes. Describe what will (most likely) happen to the *width* of the confidence interval (does it get smaller, larger or stay the same?) if we,
    a. Increase the sample size n.
    b. Decrease the confidence level $100(1-\alpha)$.
    c. Increase the variance σ^2.
    d. Decrease the expected value μ.

---

2. Consider the following four hypothesis testing scenarios. For each scenario, describe what the Type I error and Type II error mean in that particular context. Comment also on the possible consequences of the two errors in each case (which one of the errors is more "dangerous"?). For part d, come up with a typical hypothesis testing scenario from your own field of science.
    a. A suspect is brought before a judge.
        - H0: The suspect is innocent.
        - H1: The suspect is guilty.
    b. A new experimental cancer treatment is compared to placebo.
        - H0: The new treatment is no better than placebo.
        - H1: The new treatment is better than placebo.
    c. An automated security screening scans passengers at the airport.
        - H0: The passenger is not carrying dangerous items.
        - H1: The passenger is carrying dangerous items.
    d. Your own scenario here!

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 `iris` contains measurements of the sepal length and width and petal length and width for 50 flowers from each of 3 species of iris. We want to study the distribution of the ratio between sepal length and sepal width of an iris of the species `setosa`.
    a. Create a new 1-dimensional data set which contains only the ratios `Sepal.Length/Sepal.Width` for the irises of the species `setosa`.
    b. Find a suitable way to visualize the ratio.
    c. Use bootstrapping to construct a 95% confidence interval for the expected value of the ratio.
    d. Add the confidence interval end points to the plot of part b.
    e. What does the confidence interval tell us about the distribution of the ratio?
    f. What assumptions did the confidence interval in part c make?

---

2. The data set below contains the annual salaries (in dollars) of 8 American women and 8 American men. 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 expected values of the salaries of women and men differ.
    a. Find a suitable way to visualize the data.
    b. Which test is appropriate in studying our question of interest?
    c. State the hypotheses of your chosen test and conduct it on the significance level 10%.
    d. What is the conclusion of the test?
    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. Consider again the `iris` data set from exercise 1. Study whether the expected values and variances of `Petal.Length` differ between irises of the species `versicolor` and `virginica`
    a. Find a suitable way to visualize the data.
    b. Test whether the expected values differ using the two-sample t-test on a significance level 5%.
    c. Test whether the variances differ using the variance comparison test on a significance level 5%.
    d. What are the conclusions of the tests?
    e. What assumptions did the tests in parts b and c make? Are they justifiable?

---

4. **(Optional)** Writing bootstrapping code every time from scratch gets quickly repetitive and a better idea is to use the package `boot`.
    a. Find out how the function `boot` works and use it to solve exercise 1c.
    b. Use also the function `abc.ci` to compute a bootstrap confidence interval for the ratio and compare the results.

---

5. **(Optional)** Consider the data set `nhtemp` which contains the mean annual temperatures in New Haven, Connecticut, from 1912 to 1971. Does it make sense to use bootstrap to estimate confidence intervals for this kind of *time series* data? (Hint: are the bootstrap samples similar to the original sample in a meaningful way?)

---

6. **(Optional)** Let $x_1, \ldots, x_{100}$ be a random sample from the exponential distribution with the unknown *rate* parameter λ. We test the hypotheses,
    - H0: $\lambda = 1$,
    - H1: $\lambda \neq 1$,

    using $t = (\bar{x})^{-1}$ as a test statistic. Using simulations,
    a. find an approximate 95% critical region for the test,
    b. find the approximate Type II error probaility when the true value of the parameter is $\lambda = 2$ and we use the critical region from part a.