Statistical Learning

Assignment 2

Multivariate Means and Multiple Testing

Problem 1

1. Let $Z_1, Z_2, ...$ be independent copies of a $\mathcal{N}(0,1)$ variable. Use these to define a chi-squared random variable with p degrees of freedom χ^2_p. Similarly, recall the definition of an F distribution with d_1 and d_2 degrees of freedom F_{d_1,d_2}.

2. Consider a random variable distributed according to $F_{p,n-p}$. What distribution will $F_{p,n-p}$ approximately follow for $p = 4$ and $n = 1000$?

3. Let $X_1, .., X_n \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Show that $n(\bar{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{X} - \boldsymbol{\mu})$ follows a χ^2_p distribution. (Hint: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$)

4. Let $X_1, .., X_n \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Assume we do not know either $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$. We want to test the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. Recall the Hotelling T^2 statistic:

$$T^2 := n(\bar{X} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1} (\bar{X} - \boldsymbol{\mu}_0),$$

   where $\boldsymbol{S}$ denotes the sample covariance matrix.

   (a) Assume H_0 is true. What distribution does T^2 follow? Use this to write down a test, which rejects at level α.

   (b) Explain what happens to T^2 if we gather more observations and n goes to infinity. What happens to the probability of rejecting H_0?

   (c) Assume H_0 is false. Explain what happens to T^2 as n goes to infinity. What happens to the probability of rejecting H_0?

Problem 2: Multiple Testing

Random vector

$$X = (1.7, 1.6, 3.3, 2.7, -0.04, 0.35, -0.5, 1.0, 0.7, 0.8)$$

comes from the 10 dimensional multivariate normal distribution $N(\mu, I)$.

1. Which hypotheses would be rejected by the Bonferroni multiple testing procedure ?
2. Which hypotheses would be rejected by the BH multiple testing procedure ?
3. Assume that only the first three coordinates of μ are different from zero. What is the False Discovery Proportion of the Bonferroni and the BH procedures ?

Project 2: Printing Bank Notes

The Swiss bank data consists of 100 measurements on genuine bank notes. The measurements are:

- X_1 – length of the bill
- X_2 – height of the bill (left)
- X_3 – height of the bill (right)
- X_4 – distance of the inner frame to the lower border
- X_5 – distance of the inner frame to the upper border
- X_6 – length of the diagonal of the central picture.

The data can be found in the file `BankGenuine.txt`.

The data set is used to illustrate confidence intervals for the means of the multivariate data.

1. Load the data, produce scatter plots and qq-plots of the data and discuss validity of the assumption that the data are from a multivariate normal distribution.
2. Evaluate estimators of the vector of means and the covariance matrix.
3. Write an R function that is verifying if a point lies inside of the six dimensional ellipsoid that serve as the 95% confidence region for the mean value of bank notes based on the Hotelling's T^2 statistics (see Lecture 3 Slides).
4. A new production line that will be replacing the old one for printing the bank notes is tested and one of the requirements is that the average dimensions of the bank notes are comparable to these represented in the provided sample of the original bank notes. After printing a very long series of bank notes in the new production line, it was found that the mean values of the dimensions are

```
 m0
     LENGTH LEFT  RIGHT BOTTOM   TOP DIAGONAL
[1,] 214.97  130 129.67    8.3 10.16   141.52
```

(Since the number of bank notes printed out for this purpose was very large so the error of for the obtained mean values is negligible). Check if the obtained mean values are within the Hotelling's confidence region that was obtained based on the original sample of bank notes.

5. Check if the new mean vector falls within the Bonferroni's confidence rectangular region for the mean value of the old bank note dimensions.

6. Plot the projection of both confidence regions to the one-dimensional spaces marked by the axes: X_i, $i = 1, \ldots, 6$. Mark the projection of the vector of means on the obtained confidence intervals. Comment what you observed.

7. Plot the projection of both confidence regions to the two-dimensional spaces marked by the pairs of axes: X_i, X_j, $i \neq j$. Mark the projection of the vector of means. Comment what you observed. *Hint:* Use package `ellipse`.

8. Interpret geometrically the fact that the mean values of the bank note dimensions from the new production line fail to belong to the Hotelling's confidence region. Relate to the previously created graphs.

9. It has been decided that the settings of the production line needs to be tuned better to match original dimensions of banknotes. After such tuning, another test has been carried out and the resulting means were

```
m1
     LENGTH   LEFT  RIGHT BOTTOM  TOP DIAGONAL
[1,] 214.99 129.95 129.73   8.51 9.96   141.55
```

Check if the vector of means are within: a) Hotelling's confidence region; b) Bonferroni's confidence region. Comment your findings.

10. After yet another tuning, the vector of means was

```
m2
      LENGTH     LEFT    RIGHT BOTTOM     TOP DIAGONAL
[1,] 214.9473 129.9243 129.6709 8.3254 10.0389 141.4954
```

Is this value acceptable based on the original sample of the bank notes, or the production line still needs some tuning? Explain your answer.

Simulation 1: Multiple testing

Consider the sequence of independent random variables $X_1, \ldots, X_p$ such that $X_i \sim N(\mu_i, 1)$ and the problem of the multiple testing of the hypotheses $H_{0i} : \mu_i = 0$, for $i \in \{1, \ldots, p\}$. For $p = 5000$ and $\alpha = 0.05$ use the simulations (at least 1000 replicates) to estimate FWER, FDR and the power of the Bonferroni and the Benjamini-Hochberg multiple testing procedures for the following setups

a) $\mu_1 = \ldots = \mu_{10} = \sqrt{2 \log p}, \;\; \mu_{11} = \ldots = \mu_p = 0$

b) $\mu_1 = \ldots = \mu_{500} = \sqrt{2 \log p}, \;\; \mu_{501} = \ldots = \mu_p = 0$

Summarize the results in view of the theory presented in class.