MATH437/537 **Fall, 2022**

Homework 3. Due October 17

1. Recall that the Hotelling T^2-statistic based on $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ (which are assumed iid $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ invertible) is defined as

$$T^2(\boldsymbol{X}, \boldsymbol{\mu}) = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\boldsymbol{X})(\overline{\boldsymbol{X}} - \boldsymbol{\mu}),$$

where $\boldsymbol{S}(\boldsymbol{X})$ is the (unbiased) sample covariance matrix based on $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$.

   (i) What is $T^2(\boldsymbol{X}, \boldsymbol{\mu})$ in the case $p = 1$ and what is its distribution?

   (ii) Write the formula for $T^2(\boldsymbol{X}, \boldsymbol{\mu})$ in terms of the maximum likelihood estimate $\widehat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ instead of $\boldsymbol{S}$ and specify its distribution.

   (iii) Suppose the data are transformed to $\boldsymbol{Y}_k = \boldsymbol{A}\boldsymbol{X}_k + \boldsymbol{a}$, where $\boldsymbol{A}$ is a fixed invertible matrix and $\boldsymbol{a}$ is a fixed vector.

   (a) What is the mean $\boldsymbol{\mu}_Y$ and covariance matrix $\boldsymbol{\Sigma}_Y$ of $\boldsymbol{Y}_1$?

   (b) Determine the sample covariance matrix $\boldsymbol{S}(\boldsymbol{Y})$ in terms of $\boldsymbol{S}(\boldsymbol{X})$ and write $T^2(\boldsymbol{Y}, \boldsymbol{\mu}_Y)$ in terms of $T^2(\boldsymbol{X}, \boldsymbol{\mu})$.

2. (Warm-up for question #3 from STATS 101) Suppose $X_1, \dots, X_n$ are iid $N(\mu, \sigma^2)$. Find the coverage of the confidence interval $\overline{X} \pm k\, S/\sqrt{n}$ for μ as a function of k. (The answer involves the CDF of the standard Guassian that we denote as ϕ. So, If $Z \sim N(0, 1)$, then $\phi(z) = \mathbb{P}(Z \leq z)$.)

3. Suppose $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are iid $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ invertible and let $\boldsymbol{a}$ be a fixed vector in $\mathbb{R}^p$. We have seen two types of confidence intervals for $\boldsymbol{a}^\top\boldsymbol{\mu}$: a t-interval and a T^2-interval. The T^2-intervals are conservative because they have simultaneous coverage over all linear functionals of $\boldsymbol{\mu}$. To see this we can compute the actual coverage of the 95% T^2-interval for $\boldsymbol{a}^\top\boldsymbol{\mu}$ for a fixed $\boldsymbol{a} \neq \boldsymbol{0}$. That is, the probability that $\boldsymbol{a}^\top\boldsymbol{\mu}$ belongs to the 95% T^2-interval for the fixed vector $\boldsymbol{a} \neq \boldsymbol{0}$. Compute this coverage for the cases $n = 25$, $p = 3$ and $n = 25$, $p = 2$. (Note that you do not need to know what $\boldsymbol{a}$ is.)

4. Return to the Lizard data from Homework 1, where we have samples of three-dimensional vectors $\boldsymbol{X}_i = (X_i, Y_i, Z_i)$ with mean $\boldsymbol{\mu} = (\mu_X, \mu_Y, \mu_Z)$ for each of 25 lizards. Assume that the assumption $\boldsymbol{X}_1, \dots, \boldsymbol{X}_{25}$ iid $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is reasonable.

(a) Determine the sample mean $\overline{\boldsymbol{X}}$ and sample covariance matrix $\boldsymbol{S}$ based on the full data set.

(b) What is the distribution of $T^2(\boldsymbol{X}, \mu)$?

(c) Determine the value of $T^2(\boldsymbol{X}, \boldsymbol{\mu}_0)$ for $\boldsymbol{\mu}_0 = (10,\ 70,\ 140)$.

(d) What is the p-value of the test $H_0 : \boldsymbol{\mu} = (10,\ 70,\ 140)$. What can you conclude?

(e) Determine the sample covariance matrix $\boldsymbol{S}_{2,3}$ for the variables Y (SVL) and Z (HLS).

(f) Draw a scatterplot of the data and add 95% and 99% confidence ellipsoids for (μ_Y, μ_Z). Is the point (70, 140) in any of these ellipsoids? To draw the ellipsoids you may use the following R code:

```
source("bivCI.R")
biv = data with last two columns of lizard data
plot(biv, col = "red", pch = 16, cex.lab = 1.5)
lines(bivCI(s = var(biv), xbar = colMeans(biv), n = dim(biv)[1],alpha = .01,
  m = 1000),type = "l", col = "blue")
lines(bivCI(s = var(biv), xbar = colMeans(biv), n = dim(biv)[1],alpha = .05,
  m = 1000),type = "l", col = "red", lwd = 1)
# Add ‘‘+’’ sign
lines(colMeans(biv)[1], colMeans(biv)[2], pch = 3, cex = .8,type = "p", lwd = 1)
```

(g) Determine a 95% confidence t-interval for $\mu_Z - \mu_Y$ and compare it to the 95% T^2-confidence interval of $\mu_Z - \mu_Y$. What is the actual coverage of the latter?