# STAT 636, Fall 2017 - Assignment 2

Due Tuesday, October 3, 11:55pm Central

1. Find the maximum likelihood estimates of the $2 \times 1$ mean vector $\boldsymbol{\mu}$ and the $2 \times 2$ covariance matrix $\boldsymbol{\Sigma}$ based on the random sample

$$\mathbf{X} = \begin{bmatrix} 3 & 6 \\ 4 & 4 \\ 5 & 7 \\ 4 & 7 \end{bmatrix}$$

2. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{60}$ be a random sample of size $n = 60$ from a $N_6(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population. Specify each of the following.

   (a) The distribution of $(\mathbf{X}_1 - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_1 - \boldsymbol{\mu})$.

   (b) The distributions of $\bar{\mathbf{X}}$ and $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})$.

   (c) The distribution of $n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$.

   (d) The approximate distribution of $n(\bar{\mathbf{X}} - \boldsymbol{\mu})\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$.

3. Consider the `used_car` data. For each of 10 used cars, we have the numeric variables `Age` (age of the car) and `Price` (sale price of car, in $1,000s).

   (a) Determine the power transformation $\hat{\lambda}_1$ that makes the $x_1$ values approximately normal. Construct a Q-Q plot for the transformed data.

   (b) Determine the power transformation $\hat{\lambda}_2$ that makes the $x_2$ values approximately normal. Construct a Q-Q plot for the transformed data.

   (c) FOR 5 POINTS EXTRA CREDIT: Determine the power transformations $\hat{\boldsymbol{\lambda}}' = [\hat{\lambda}_1, \hat{\lambda}_2]$ that make the $[x_1, x_2]$ values approximately multivariate normal. Compare the results with those from above.

4. Consider the `sweat` data. For each of 20 healthy females, we have three numeric variables that measure aspects of perspiration: `Sweat` (sweat rate), `Sodium` (sodium content), and `Potassium` (potassium content).

   (a) Construct univariate Q-Q plots for each of the three variables. Also make the three pairwise scatterplots. Does the multivariate normal assumption seem reasonable?

   (b) Determine the 95% confidence ellipsoid for $\boldsymbol{\mu}$. Where is it centered? What are its axes and corresponding half-lengths?

   (c) Compute 95% $T^2$ simultaneous confidence intervals for the three mean components.

   (d) Compute 95% Bonferroni simultaneous confidence intervals for the three mean components.

   (e) Carry out a Hotelling's $T^2$ test of the null hypothesis $H_0 : \boldsymbol{\mu}' = [4.0, 45.0, 10.0]$ at $\alpha = 0.05$. What is the test statistic, critical value, and the p-value? What is your conclusion regarding $H_0$?

(f) Is $\boldsymbol{\mu}' = [4.0, 45.0, 10.0]$ inside the 95% confidence ellipse you computed in part (b)? Is this consistent with your findings in part (e)? Hint: It should be.

(g) FOR 5 POINTS EXTRA CREDIT: Use the bootstrap to test the same null hypothesis as in part (e), now using this as your test statistic

$$\Lambda = \left( \frac{|\mathbf{S}|}{|\mathbf{S}_0|} \right)^{n/2},$$

where

$$\mathbf{S} = \frac{1}{n-1} \sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}}) (\mathbf{x}_j - \bar{\mathbf{x}})'$$

is the sample covariance matrix, and

$$\mathbf{S}_0 = \frac{1}{n-1} \sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu}_0) (\mathbf{x}_j - \boldsymbol{\mu}_0)'$$

is the sample covariance matrix computed under the assumption that $H_0$ is true. So that all of our answers match, first do `set.seed(101)`, and use $B = 500$ bootstrap iterations. What is the p-value?

5. Consider the `peanut` data. These represent a two-factor experiment on peanut crops. The two factors are (i) the geographical location of the crop (2 locations were considered) and (ii) the variety of peanut grown (three varieties were considered). We have 2 crops under each of the $2 \times 3 = 6$ factor combinations. For each crop, we have measurements of three weight variables: $X_1$ = total yield, $X_2$ = sound mature kernels, and $X_3$ = seed size. So, in terms of a two-way MANOVA model, $g = 2$, $b = 3$, and $n = 2$.

(a) Test for a location effect, a variety effect, and a location-variety interaction at $\alpha = 0.05$. Do this using the `manova` function in `R`. Overall, what do you conclude about these data?

(b) FOR 5 POINTS EXTRA CREDIT: Construct the two-way MANOVA table by computing $\text{SSP}_{\text{FAC 1}}$, $\text{SSP}_{\text{FAC 2}}$, $\text{SSP}_{\text{INT}}$, $\text{SSP}_{\text{RES}}$, and $\text{SSP}_{\text{COR}}$. Provide `R` code that matches the Wilks' statistics computed by `manova`. Note that your p-values (computed according to the notes) will not match those of `manova`, because the distributional results we have learned for two-way MANOVA are large-sample approximations. That said, how do your p-values compare to those of `manova`?