ST 557: Applied Multivariate Analysis
Fall 2019

Homework 3

Due: Friday October 25
75 Points

1. **(15 points)** The scores obtained by $n = 87$ college students on the College Level Examination Program (CLEP) test for social science and history (X_1) and on the College Qualification Test (CQT) test for verbal (X_2) and science (X_3) are given in the TestScores.csv file available on Canvas.

   (a) Test the null hypothesis $H_0 : \boldsymbol{\mu} = [500, 50, 30]^T$ vs. the alternative $H_A : \boldsymbol{\mu} \neq [500, 50, 30]^T$ at significance level $\alpha = 0.05$ using Hotelling's T^2 test. Suppose that $[500, 50, 30]^T$ is the vector of average scores for all of the students who took the test between 2000 and 2010. Is there reason to believe that the population of students this year (2011) represented by the students in the given data set is scoring differently? Explain.

   (b) Determine the lengths and directions for the axes of the 95% confidence ellipsoid for $\boldsymbol{\mu}$.

   (c) Construct quantile-quantile plots for the marginal distributions of scores for the three tests. Also, construct the three pair-wise scatter plots for these variables. Do these data appear to be normally distributed?

2. **(10 points)** Measurements of $X_1 =$ stiffness and $X_2 =$ bending strength for a sample of $n = 30$ pieces of lumber are given in the LumberData.csv file available on Canvas.

   (a) Construct and sketch/plot a 95% confidence ellipse for the mean vector $\boldsymbol{\mu} = [\mu_1, \mu_2]^T$ where $\mu_1 = E(X_1)$ and $\mu_2 = E(X_2)$.

   (b) Suppose $\mu_0 = [2000, 10000]^T$ represent "typical" values for stiffness and bending strength. Given the results of part (a), do these values seem like plausible means for the variety of lumber sampled for this data set?

3. **(10 points)** At the start of a study to determine whether exercise or dietary supplements would slow bone loss in older women, an investigator measured the mineral content of bones by photon absorptiometry. Measurements were recorded for three bones on the dominant and nondominant sides and are given in the BoneMineral.csv file available on Canvas.

   (a) Construct the 95% Bonferroni intervals for the means of the individual variables.

   (b) Construct the 95% simultaneous T^2 intervals for the means of the individual variables. Compare these intervals to those found in part (a).

4. **(20 points)** As a quality-control check on the performance of a machine that fills bags of flour, 20 bags are randomly sampled from the production line and the mass of each bag is assessed using three different scales. The data containing the measurements for the 20 bags are given in the FlourBags.csv file available on Canvas.

(a) The bags are supposed to contain 10 pounds of flour. Test whether the null hypothesis that the mean vector for the measurements of the three scales is $H_0 : \boldsymbol{\mu} = [10, 10, 10]^T$ vs. the alternative $H_A : \boldsymbol{\mu} \neq [10, 10, 10]^T$ at significance level $\alpha = 0.05$ using Hotelling's T^2 test.

(b) Test whether the null hypothesis that the mean vector for the measurements of the three scales is $H_0 : \boldsymbol{\mu} = [10, 10, 10]^T$ vs. the alternative $H_A : \boldsymbol{\mu} \neq [10, 10, 10]^T$ at significance level $\alpha = 0.05$ using the Bonferroni multiple testing method.

(c) Based on the result for part (b), can you determine (without computing) whether the T^2 simultaneous confidence intervals would contain $\mu_0 = [10, 10, 10]^T$?

(d) Perform a (univariate) level $\alpha = 0.05$ test of the null hypothesis $H_0 : \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 = 10$ vs. the alternative $H_A : \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 \neq 10$ where μ_1, μ_2, and μ_3 are the means for each of the three scales.

5. **(20 points)** Improving on Bonferroni: Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be an iid sample of $(p \times 1)$ random vectors from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

(a) If the covariance matrix is the $(p \times p)$ identity matrix ($\boldsymbol{\Sigma} = \mathbf{I}_p$), are the elements of $\bar{\mathbf{X}}$ independent? Why or why not?

(b) Suppose that you perform level α^* tests of each variable mean separately; that is, test the p hypotheses

$$\begin{array}{lcl} H_0 : \mu_1 = \mu_{01} & \text{vs.} & H_A : \mu_1 \neq \mu_{01} \\ H_0 : \mu_2 = \mu_{02} & \text{vs.} & H_A : \mu_2 \neq \mu_{02} \\ \vdots & & \vdots \\ H_0 : \mu_j = \mu_{0j} & \text{vs.} & H_A : \mu_j \neq \mu_{0j} \\ \vdots & & \vdots \\ H_0 : \mu_p = \mu_{0p} & \text{vs.} & H_A : \mu_p \neq \mu_{0p} \end{array}$$

each at level α^*. If the covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}_p$, what is the probability (in terms of α^*) that none of these individual tests will reject the null hypothesis, if all of the null hypotheses are in fact true? That is, what is

$$P_{\boldsymbol{\mu}_0}(\text{ Reject } H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0)$$

if $\boldsymbol{\mu_0} = [\mu_{01}, \mu_{02}, \ldots, \mu_{0p}]^T$, and the overall hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ is rejected if any of the individual null hypotheses are rejected?

(c) Given your answer to part (b), what should α^* be to control the overall significance level of this procedure at level α? [Note that $\boldsymbol{\Sigma} = \mathbf{I}_p$ is a worst-case scenario if the data are multivariate normal, so this value α^* will also control the overall significance level (that is, the overall level will be $\leq \alpha$) for $\boldsymbol{\Sigma} \neq \mathbf{I}_p$.]

(d) Find the ratio of the lengths of simultaneous confidence intervals constructed using the value of α^* found in part (c) to the lengths of Bonferroni simultaneous confidence intervals. Evaluate the value of this ratio for the following settings:

i. 95% confidence level (so $\alpha = 0.05$); $n = 10, p = 4$
ii. 95% confidence level (so $\alpha = 0.05$); $n = 10, p = 8$
iii. 95% confidence level (so $\alpha = 0.05$); $n = 20, p = 4$
iv. 95% confidence level (so $\alpha = 0.05$); $n = 20, p = 8$