STAT 636, Fall 2015 - Assignment 5
Due Monday, November 2, 2:55pm Central
Online Students: Submit your assignment through WebAssign.
On-Campus Students: Email your assignment to the TA.

1. For the (paired) effluent data in Table 6.1 from the textbook, let $\boldsymbol{\delta}' = [\delta_1, \delta_2]$, where δ_1 is the mean difference between the commercial lab BOD measurements and the state lab BOD measurements, and δ_2 is the mean difference between the commercial lab SS measurements and the state lab SS measurements.

   (a) Test $H_0 : \boldsymbol{\delta} = \mathbf{0}$ using the `HotellingsT2Test` function from the `DescTools` `R` package. Provide `R` code that matches the test statistic and p-value returned by `HotellingsT2Test`. What is your conclusion if testing at $\alpha = 0.01$?

   (b) Construct and plot a 99% confidence region for $\boldsymbol{\delta}$. Is $\boldsymbol{\delta} = 0$ inside this region? Is this consistent with the results of part (a)?

   (c) Construct 99% simultaneous confidence intervals for δ_1, δ_2, and $\delta_1 - \delta_2$, using both the T^2 and Bonferroni methods. Compare the lengths of the T^2 intervals to those of the Bonferroni intervals.

   (d) Sample 8 could arguably be considered an outlier. What is the p-value for testing $H_0 : \boldsymbol{\delta} = \mathbf{0}$ when this sample is excluded?

2. For the data in Exercise 6.8 from the textbook ($p = 2$, $g = 3$, $n_1 = 5$, $n_2 = 3$, $n_3 = 4$):

   (a) Consider just treatments 2 and 3, and assume that the covariance matrix is the same for both treatments.

      i. Test $H_0 : \boldsymbol{\mu}_2 - \boldsymbol{\mu}_3 = \mathbf{0}$ with a two-sample T^2 statistic, at $\alpha = 0.01$. Do this using the `hotelling.test` function from the `Hotelling` `R` package, and provide `R` code that matches its output (statistic and p-value).

      ii. Construct 99% Bonferroni simultaneous confidence intervals for the differences $\mu_{2i} - \mu_{3i}$, $i = 1, 2$.

   (b) Now consider all three treatments.

      i. Construct the one-way MANOVA table.

      ii. Use the Wilks' lambda statistic to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$ at $\alpha = 0.01$. Do this using the `manova` function in `R`, and provide `R` code that matches its output (statistic and p-value). Note that to get the Wilks' statistic and p-value from `manova`, you can do the following (you will need to fill in the "..."):

```
summary(manova(...), test = ''Wilks'')
```

3. The use of the pooled sample covariance matrix in the two-sample T^2 test can be motivated by a likelihood argument. Give the likelihood function, $L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, for two independent samples of sizes n_1 and n_2 from $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ populations, respectively. Show that this likelihood is maximized by the choices $\hat{\boldsymbol{\mu}}_1 = \bar{\mathbf{x}}_1$, $\hat{\boldsymbol{\mu}}_2 = \bar{\mathbf{x}}_2$, and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n_1 + n_2}\left[(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2\right] = \left(\frac{n_1 + n_2 - 2}{n_1 + n_2}\right)\mathbf{S}_{\text{pooled}}$$

Hint: Refer to slides 34-40 in the Topic 4 notes and / or Section 4.3 of the textbook.

4. Consider the data in Table 6.17 from the textbook. 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) Construct the two-way MANOVA table.

   (b) 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`, and provide `R` code that matches the Wilks' statistics it computes. Note that your p-values (computed according to the book / notes) will not match those of `manova`. 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`? Overall, what do you conclude about these data?