STAT 475-575 **ASSIGNMENT 2** **NAME _____**

Suggested Reading : Chapter 3 in Everitt and Hothorn.

Written Assignment: Due Friday, October 3, 2020 by 11:59pm.

1. To examine if the mean thickness of cork is the same on each side of corks trees, measurements of the thickness of cork deposits on four sides of each cork tree were taken on a random sample of n=28 trees (C. R. Rao, 1948, *Biometrika*). The data are posted on Canvas as **cork.csv**, with one line of data for each tree. The data file has five columns corresponding to the following variables

Column	Variable	
1	Tree	Tree Identification Number
2	X_1	Thickness of cork deposit on north side of the tree
3	X_2	Thickness of cork deposit on east side of the tree
4	X_3	Thickness of cork deposit on south side of the tree
5	X_4	Thickness of cork deposit on west side of the tree

(a) Report values for the vector of sample means $\underset{\sim}{\bar{X}} = \left(\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4\right)'$

(b) Present a set of box plots for the four measurements. Do the box plots suggest any differences in the distributions cork deposit thicknesses on the four different sides of the trees?

(c) Report the correlation matrix for these four measurements. Also present a scatterplot matrix for these four measurements. Summarize what you learned from this information.

(d) Check Normality of the data. Plot QQ-plot for each variable and conduct both uni-variate Shapiro test and multivariate Shapiro test. Report the pvalues and state your conclusion.

(e) Let $\mu 1, \mu 2, \mu 3, \mu 4$ denote the population means for thickness of cork on the north, east, south and west sides of trees (for the population of all trees that live in the regions from which the sample was taken). Test the null hypothesis that all the four means equal to 47. Report the Hotelling's T2 statistic and degree of freedoms of its associated F statistic. Report the pvalue and state your conclusion.

(f) Let $\mu 1, \mu 2, \mu 3, \mu 4$ denote the population means for thickness of cork on the north, east, south and west sides of trees (for the population of all trees that live in the regions from which the sample was taken). Test the null hypothesis that the means are the same on the four sides of the trees, i.e., H0: $\mu 1 = \mu 2 = \mu 3 = \mu 4$. Report the Hotelling's T2 statistic and degree of freedoms of its associated F statistic. Report the pvalue and state your conclusion.

(e) Compute a set of simultaneous Bonferroni confidence intervals for the difference from the six possible pairs of means: $\mu 1 - \mu 2$, $\mu 1 - \mu 3$, $\mu 1 - \mu 4$, $\mu 2 - \mu 3$, $\mu 2 - \mu 4$, $\mu 3 - \mu 4$. Construct the confidence intervals so that the probability that all six of the intervals contain the corresponding true difference in population means is at least 0.95. Which sides of the trees have thicker cork deposits on average, and which sides have thinner cork deposits?

2. This problem is concerned with the identification of forged bank notes. The file bnotes.csv, available from Canvas web page, has data from six measurements that roughly quantify the size and the position of the printed image on 1000-franc Swiss bank notes. Many other variables could be considered, but these six variables are easily measured with automatic scanning equipment. The file **bnotes.csv** contains data for a sample of 100 genuine 1000-franc bank notes (Flury and Riedwyl, 1987). There is one line of data for each bank note in the sample with the data arranged in the following order:

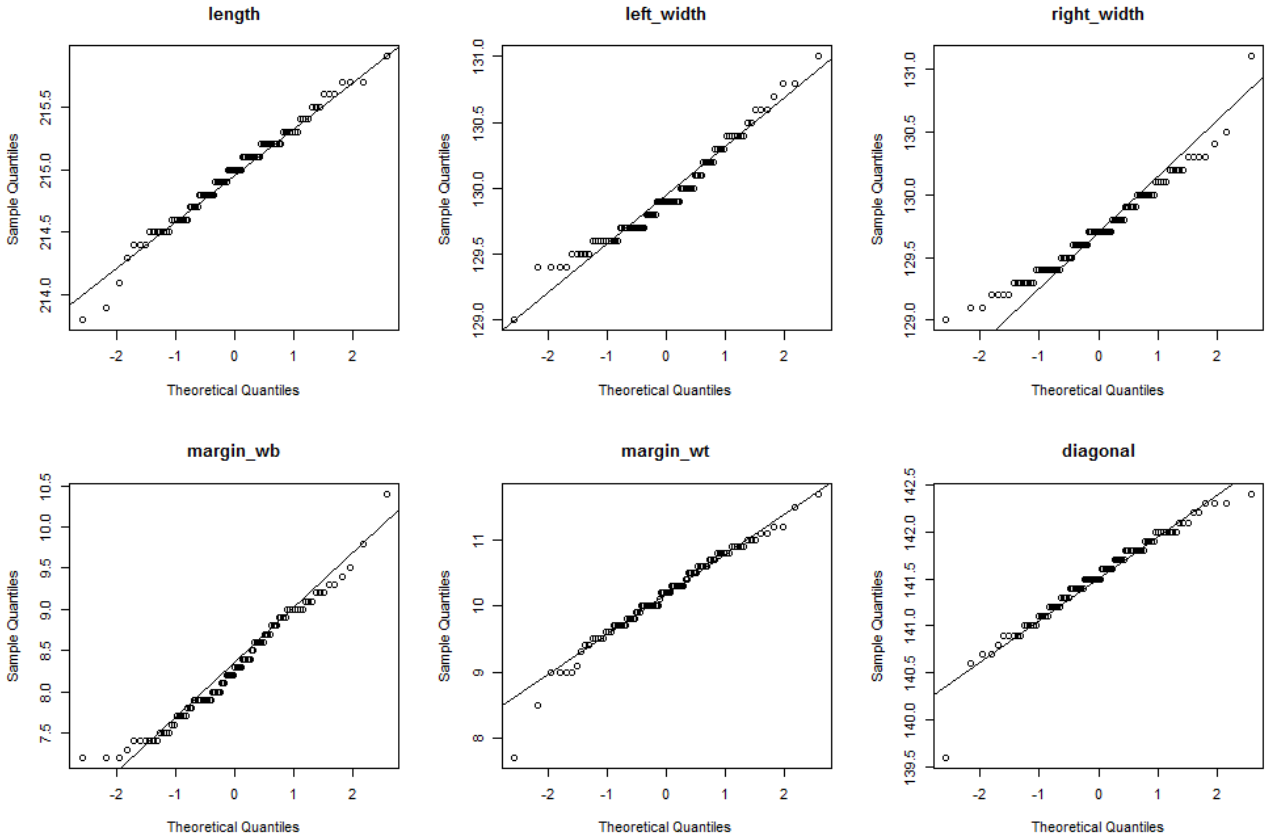
Note	Bank note identification number
X1	Length of the bill (mm)
X2	Width of the bill on the left side (mm)
X3	Width of the bill on the right side (mm)
X4	Width of the margin at the bottom of the bill (mm)
X5	Width of the margin at the top of the bill (mm)
X6	Diagonal length of the printed image (mm)

It is relatively easy to obtain a sample from the population of genuine notes, but it would be difficult to sample from the population of forged notes for obvious reasons. In this case, information is available only for the genuine notes, and no information is available for forged notes. The identification problem consists of using the data available from the sample of genuine notes to develop a procedure for deciding whether notes of uncertain origin are genuine or forged. We will first try to determine if the variation in the six measurements on genuine notes can be approximately described by a normal distribution. If so, we will use the information in the sample of 100 genuine notes to determine whether a "suspect" note was likely to come from the population of genuine notes.

(a) To check normality report the values of the Shapiro-Wilk statistic for each variable.

	X_1	X_2	X_3	X_4	X_5	X_6
Value of W	_____	_____	_____	_____	_____	_____
p-value	_____	_____	_____	_____	_____	_____

Also examine the corresponding normal q-q plots shown below. State your conclusions.



(b) Compute the p-value for the test for multivariate normality. Does it indicate that the data were not sampled from a 6-dimensional normal model?

Note: The Shapiro-Wilk tests provide some evidence of non-normality. You could investigate transformations of the variables, but it would be difficult to find a transformation for which the data points are closer to straight lines on the normal q-q plots. The horizontal stripes you see in the q-q plots occur because the measurements were rounded off in the listing of the data from which the data files was made. This contributes to the lack of normality and no transformation can fix that. If the data file contained the original unrounded measurements, the values of the Shapiro-Wilk test would be closer to one and may not have rejected the null hypothesis that the measurements were sampled from a multivariate normal distribution. Because the q-q plots are nearly straight lines, we will proceed as if the data were sampled from a multivariate normal distribution.

(c) Identification analysis: Suppose the measurements $\underset{\sim}{X}_0 = (X_{01}, X_{02}, \ldots, X_{06})'$ are made on a bank note of uncertain origin. If the new vector of measurements, $\underset{\sim}{X}_0$, is a random selection from the population of genuine bank notes, the mean vector for $\underset{\sim}{X}_0$ is the mean vector $\underset{\sim}{\mu}$ for the population of genuine bank notes. The covariance matrix for

$\underset{\sim}{X}_0$ is Σ, the covariance matrix for the population of genuine bank notes. Using $\underset{\sim}{\overline{X}}$ to represent the vector of sample means for the random sample of n genuine bank notes and S to represent the sample covariance matrix, the mean vector for $\underset{\sim}{X}_0 - \underset{\sim}{\overline{X}}$ is a vector of zeros if $\underset{\sim}{X}_0$ is a vector of measurements from a genuine bank note. The covariance matrix for $\underset{\sim}{X}_0 - \underset{\sim}{\overline{X}}$ is $\left(1+\frac{1}{n}\right)\Sigma$, and this is estimated as $\left(1+\frac{1}{n}\right)S$. Then the statistical distance from the new bank note to vector of sample means is

$$d^2 = \left(\underset{\sim}{X}_0 - \overline{X}\right)'\left[\left(1+\frac{1}{n}\right)S\right]^{-1}\left(\underset{\sim}{X}_0 - \overline{X}\right) = \frac{n}{n+1}\left(\underset{\sim}{X}_0 - \overline{X}\right)'[S]^{-1}\left(\underset{\sim}{X}_0 - \overline{X}\right).$$

If the new bank note is a random selection from the population of genuine bank notes, d^2 has a chi-square distribution with p=6 degrees of freedom. The new bill would be considered unusual enough to be a forgery if d^2 is too far out in the right tail of the chi-square distribution with p=6 degrees of freedom.

Measurements were taken on five bank notes of unknown origin. The measurements are

Bank note 1001: $\underset{\sim}{X}_0$ = (214.9, 130.5, 130.2, 8.4, 11.6, 138.4)'.

Bank note 1001: $\underset{\sim}{X}_0$ = (215.9, 129.5, 130.6, 7.9, 12.1, 140.8)'.

Bank note 1001: $\underset{\sim}{X}_0$ = (215.3, 130.7, 130.6, 8.1, 11.7, 142.4)'.

Bank note 1001: $\underset{\sim}{X}_0$ = (214.9, 130.1, 129.8, 8.7, 10.9, 141.8)'.

Bank note 1001: $\underset{\sim}{X}_0$ = (215.9, 131.3, 129.2, 8.5, 11.5, 138.4)'.

Which of these bank notes would you determine to be forgeries? Which could be genuine bank notes. Explain.