Example of Use of Hotelling's T^2: Lumber Data

Textbook problem 5.19 presents a set of data from the U.S. Forest Products Laboratory on measurements on $p = 2$ variables: x_1=stiffness, and x_2=bending strength for a sample of n=30 pieces of a particular grade of lumber. The data are provided in Table 5.11. For the remainder of this problem, both variables have been rescaled by dividing by 100.

(a) Suppose that $\mu_{10} = 20$ and $\mu_{20} = 100$ represent "typical" values for stiffness and bending strength, respectively. Test the hypothesis $H_0\colon \underline{\mu} = \underline{\mu}_0 = \begin{pmatrix} 20 \\ 100 \end{pmatrix}$ vs. $H_0\colon \underline{\mu} \neq \underline{\mu}_0$. Use $\alpha = 0.05$.

(b) Is there any evidence for departure from univariate or multivariate normality for these data?

(c) Based on (a) and (b), what do you conclude?

<u>Answer to part (a)</u>:

Here n=30 and p=2. Doing as many of the calculations as possible in ***R*** (see page 3-4 for code), we find: $\bar{\mathbf{x}} = \begin{pmatrix} 18.605 \\ 83.5414 \end{pmatrix}$ and $\mathbf{S} = \begin{pmatrix} 12.40547 & 36.16204 \\ 36.16204 & 348.63332 \end{pmatrix}$. Hand calculation of the inverse of **S**, yields $\mathbf{S}^{-1} = \begin{pmatrix} 0.115546 & -0.011985 \\ -0.011985 & 0.000411 \end{pmatrix}$. Therefore the test statistic becomes

$$T^2 = n(\underline{\bar{\mathbf{x}}} - \underline{\mu}_0)'\mathbf{S}^{-1}(\underline{\bar{\mathbf{x}}} - \underline{\mu}_0) = 30\begin{pmatrix} -1.395 & -16.459 \end{pmatrix}\begin{pmatrix} 0.115546 & -0.011985 \\ -0.011985 & 0.00411 \end{pmatrix}\begin{pmatrix} -1.395 \\ -16.459 \end{pmatrix} = 23.64.$$

Now for $\alpha = 0.05$, the critical value is $\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha) = \frac{29(2)}{28}F_{2,28}(.05) = 2.0714(3.34) = 6.92$. Using ***R*** we can also determine that the p-value is 0.00024. Therefore we have very strong evidence to reject the null hypothesis and conclude that the data are not consistent with a population mean vector of $\underline{\mu} = \underline{\mu}_0 = \begin{pmatrix} 20 \\ 100 \end{pmatrix}$.

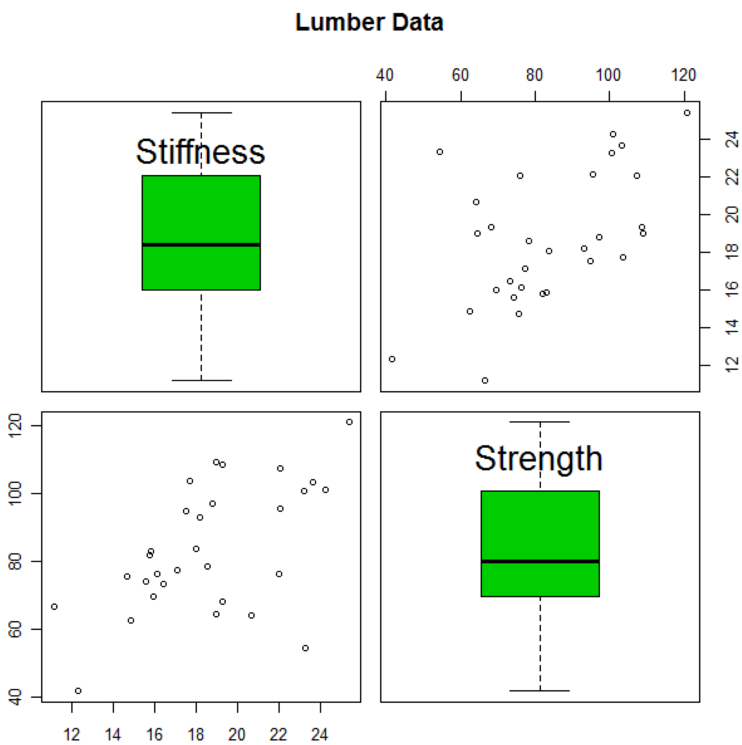
<u>Answer to part (b)</u>:

Examination of the boxplots and normal quantile plots for each variable shows that the data for each variable is not inconsistent with the assumption that we are sampling from a Normal population. Shapiro-Wilk tests corroborate this conclusion.

Examination of the multivariate scatterplot matrix of the data, and the chi-square quantile plot for the Mahalanobis distances^2, also shows that the assumption of bivariate normality is not unreasonable.

<u>Answer to part (c)</u>:

The use of Hotelling's T^2 as a method of analysis in part (a) relies on the assumption of multivariate normality of the underlying population from which this data was sampled. Based on our assessment in part (b), this assumption is justified. Thus the analysis and conclusions drawn in part (a) are justified, and we conclude that the population mean vector for stiffness and mean bending strength for pieces of lumber from which the data was sampled is <u>inconsistent</u> with the claimed value.

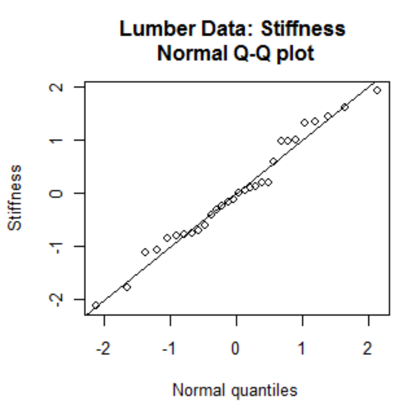


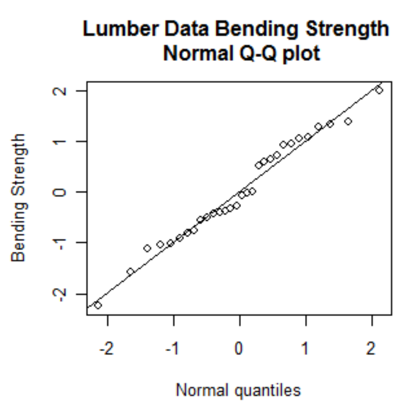
Boxplots and scatterplots of Stiffness and Bending Strength. The boxplots of each variable are not too far off from symmetric and the scatterplot, while not showing a perfectly elliptical pattern, isn't too bad.

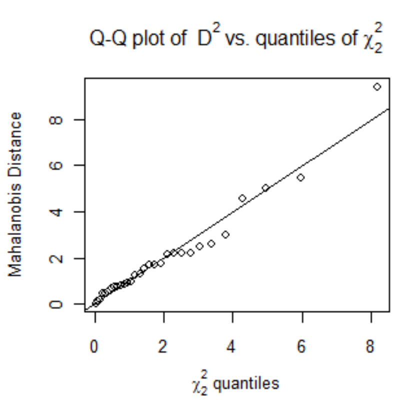
Further, the Q-Q plots below for each variable separately and for the pair of variables jointly (Mahalanobis distances), does not show any departure from the pattern we would expect for bivariate normal data.

```
Shapiro-Wilk normality test
data:  Xdata[, 1]
W = 0.9749, p-value = 0.6798
```

```
Shapiro-Wilk normality test
data:  Xdata[, 2]
W = 0.9755, p-value = 0.6979
```







(Courtesy of Dr. Roy St. Laurent)

R Commands File Analysis of Lumber Data

```
#
# Analysis of Lumber Data from Johnson & Wichern page 267, Problem 5.19
#
TX<-c(12.32, 41.75,
11.15,  66.52,
22.05,  76.12,
18.97, 109.14,
19.32, 108.50,
16.12,  76.27,
15.98,  69.54,
18.04,  83.65,
17.52,  94.69,
20.67,  64.10,
23.65, 103.27,
16.46,  73.20,
15.79,  81.96,
18.80,  97.09,
17.73, 103.70,
17.12,  77.49,
19.32,  68.18,
18.20,  93.07,
19.00,  64.57,
24.26, 101.02,
15.58,  74.14,
14.70,  75.56,
18.58,  78.33,
15.87,  83.09,
22.08,  95.59,
14.87,  62.55,
22.06, 107.23,
23.32,  54.30,
25.40, 120.90,
23.22, 100.72)
#
Xdata<-matrix(TX,ncol=2,byrow=TRUE)
Lumber <- data.frame(Stiffness=Xdata[,1],Strength=Xdata[,2])
#
# Calculate test Hotelling's T^2 for test of hypothesis
#xbar<-mean(Lumber)
xbar<-apply(Lumber, 2, mean)
print(xbar)
S<-cov(Lumber)
print(S)
library(MASS)
#
# Inverse of Covariance Matrix
Sinv<-ginv(S)
#
# Difference in observed and hypothesized mean
Xdiff <- xbar - c(20,100)
#
# Test Statistic
T2 <- 30* (Xdiff %*% Sinv %*% Xdiff)
print(T2)
#
# Critical Value for alpha = 0.05
n <- nrow(Lumber)
p <- ncol(Lumber)
((n-1)*p / (n-p) ) * qf(p=1-0.05,df1=p,df2=n-p)
#
# Calculation of p-value
Pvalue <- 1-pf(T2*(n-p)/(p*(n-1)),df1=p,df2=n-p)
print(Pvalue)
#
```

```
#  Assess Univariate Normality
#
library(car)
spm(Lumber,diagonal=list(method="boxplot"),smooth=FALSE,regLine=FALSE,main=c("Lumber Data"))
#
par(mfrow=c(2,2))
Qtiles<-qnorm(ppoints(n))
qqplot(Qtiles,(Lumber$Stiffness-mean(Lumber$Stiffness))/sd(Lumber$Stiffness),main = "Lumber
Data: Stiffness \nNormal Q-Q plot",xlab="Normal quantiles",ylab="Stiffness")
abline(0,1)
qqplot(Qtiles,(Lumber$Strength-mean(Lumber$Strength))/sd(Lumber$Strength),main = "Lumber Data
Bending Strength \n Normal Q-Q plot",xlab="Normal quantiles",ylab="Bending Strength")
abline(0,1)
shapiro.test(Xdata[,1]); shapiro.test(Xdata[,2])
#
# Assessing Multivariate Normality
#
MHX<-mahalanobis(Lumber, xbar, S)
Qtiles<-qchisq(ppoints(n),df=p)
qqplot(Qtiles, MHX,main = expression("Q-Q plot of " * ~D^2 * " vs. quantiles of" * ~
chi[2]^2),xlab=expression(chi[2]^2 * " quantiles"),ylab="Mahalanobis Distance")
abline(0, 1)
```