STAT 636, Fall 2015 - Assignment 4
Solutions

1. For the sweat data in Table 5.1 of the textbook:

   (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?

   See Figures 1 and 2. While there is some relatively minor deviation from linearity in all three of the Q-Q plots, they look reasonable overall. The pairwise scatterplots similarly look reasonably elliptical, and there are no obvious outliers. I would say multivariate normality is a reasonable assumption here.

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

   The 95% confidence ellipsoid consists of all points $\boldsymbol{\mu}$ such that

$$n\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right) \leq \frac{p(n-1)}{(n-p)}F_{p,n-p}(0.05)$$

   Here, $n=20$, $p=3$,

$$\bar{\mathbf{x}} = \begin{bmatrix} 4.640 \\ 45.400 \\ 9.965 \end{bmatrix} \quad \text{and} \quad \mathbf{S} = \begin{bmatrix} 2.8794 & 10.0100 & -1.8091 \\ 10.0100 & 199.7884 & -5.6400 \\ -1.8091 & -5.6400 & 3.6277 \end{bmatrix}$$

   We have

$$\frac{p(n-1)}{(n-p)}F_{p,n-p}(0.05) = \frac{3(19)}{(17)}F_{3,17}(0.05) = 10.7186$$

   The eigenvalues of the sample covariance matrix $\mathbf{S}$ are $\lambda_1 = 200.4645$, $\lambda_2 = 4.5316$, and $\lambda_3 = 1.3014$, with corresponding eigenvectors

$$\mathbf{e}_1 = \begin{bmatrix} -0.0508 \\ -0.9983 \\ 0.0291 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} -0.5737 \\ 0.0530 \\ 0.8173 \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 0.8175 \\ -0.0249 \\ 0.5754 \end{bmatrix}$$

   The confidence ellipsoid is centered at $\bar{\mathbf{x}}$ and has axes equal to the above eigenvectors, with half-lengths

$$\sqrt{\left(\frac{\lambda_i}{n}\right)\left(\frac{p(n-1)}{(n-p)}\right)F_{p,n-p}(0.05)}$$

   $i = 1,2,3$. These compute to half-lengths of 10.3650, 1.5584, and 0.8351, respectively.

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

   The 95% T^2 intervals are given by

$$\bar{x}_i \pm \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(0.05)}\sqrt{\frac{s_{ii}}{n}}$$

THESE COMPUTE TO

$$\text{SWEAT}: [3.3978, 5.8822], \quad \text{SODIUM}: [35.0524, 55.7476], \quad \text{POTASSIUM}: [8.5707, 11.3593]$$

WE ARE "95% CONFIDENT" THAT THE MEAN COMPONENTS ARE SIMULTANEOUSLY INSIDE THEIR RESPECTIVE INTERVALS.

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

THE 95% BONFERRONI INTERVALS ARE GIVEN BY

$$\bar{x}_i \pm t_{n-1}\left(\frac{0.05}{2(3)}\right)\sqrt{\frac{s_{ii}}{n}}$$

THESE COMPUTE TO

$$\text{SWEAT}: [3.6440, 5.6360], \quad \text{SODIUM}: [37.1031, 53.6970], \quad \text{POTASSIUM}: [8.8470, 11.0830]$$

AS WITH THE T^2 INTERVALS, WE ARE "95% CONFIDENT" THAT THE MEAN COMPONENTS ARE SIMULTANEOUSLY INSIDE THEIR RESPECTIVE INTERVALS. HOWEVER, THE BONFERRONI INTERVALS ARE NARROWER, SINCE WE ARE RESTRICTING ATTENTION TO JUST THESE THREE CONFIDENCE STATEMENTS.

(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?

THE T^2 TEST STATISTIC IS

$$T^2 = n\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right)' \mathbf{S}^{-1}\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right) = 4.3746$$

AND THE SAMPLING DISTRIBUTION UNDER H_0 IS

$$\frac{p(n-1)}{(n-p)} F_{p,n-p}$$

THE CRITICAL VALUE IS 10.7186 (WE COMPUTED IT IN PART (B)). TO COMPUTE THE P-VALUE, WE NOTE THAT THE NULL SAMPLING DISTRIBUTION OF

$$\frac{(n-p)}{p(n-1)} T^2$$

IS $F_{p,n-p}$. WE HAVE

$$\frac{(n-p)}{p(n-1)} T^2 = 1.3047$$

AND THE P-VALUE IS 0.3053. WHETHER WE COMPARE THE VALUE OF $T^2 = 4.3746$ TO THE CRITICAL VALUE OR COMPARE THE P-VALUE TO $\alpha = 0.05$, WE REACH THE SAME CONCLUSION: FAIL TO REJECT 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.

SINCE, FROM PART (E), $T^2 = 4.3746 < 10.7186$, $\boldsymbol{\mu}' = [4.0, 45.0, 10.0]$ IS INSIDE THE 95% CONFIDENCE ELLIPSE. THIS IS NECESSARILY TRUE, GIVEN OUR FINDINGS IN PART (E), SINCE THE HOTELLING'S T^2 TEST OF $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ AT SIGNIFICANCE LEVEL α IS EQUIVALENT TO CHECKING WHETHER $\boldsymbol{\mu}_0$ IS INSIDE THE $(1-\alpha)100\%$ CONFIDENCE ELLIPSE.

(g) 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?

WE HAVE $\Lambda = 0.1259$. TO CARRY OUT THE BOOTSTRAP TEST, WE FIRST TRANSFORM THE DATA HAVE SAMPLE MEAN $\bar{\mathbf{x}} = \boldsymbol{\mu}_0$ (FORCING H_0 TO BE TRUE). WE THEN SAMPLE INDIVIDUALS (ROWS) WITH REPLACEMENT FROM THE TRANSFORMED DATA, GIVING US A "PSEUDOSAMPLE" UNDER H_0. THEN COMPUTE THE TEST STATISTIC, AND REPEAT. IF H_0 IS TRUE, THEN WE WOULD EXPECT $|\mathbf{S}|$ AND $|\mathbf{S}_0|$ TO BE OF SIMILAR MAGNITUDE, RESULTING IN A TEST STATISTIC AROUND 1. IF H_0 IS NOT TRUE, THEN WE WOULD EXPECT $|\mathbf{S}| < |\mathbf{S}_0|$, RESULTING IN A SMALL TEST STATISTIC VALUE. THUS, THE P-VALUE IS THE PROPORTION OF BOOTSTRAPPED TEST STATISTICS THAT ARE LESS THAN OR EQUAL TO THAT COMPUTED ON THE ORIGINAL DATA. THE P-VALUE COMES OUT TO 0.344, SO WE AGAIN FAIL TO REJECT H_0 AT $\alpha = 0.05$.

```
####
#### Perspiration from 20 healthy females was analyzed. Three components, X_1 = sweat
#### rate, X_2 = sodium content, and X_3 = potassium content, were measured.
####

## Load data.
X <- read.table("T5-1.DAT", header = FALSE)
colnames(X) <- c("Sweat", "Sodium", "Potassium")
attach(X)

n <- nrow(X)
p <- 3

## Summary statistics.
x_bar <- colMeans(X)
S <- var(X)

##
## Check normality.
##

pdf("figures/sweat_QQ.pdf", width = 7, height = 3.5)
par(mfrow = c(1, 3))
qqnorm(Sweat, main = "Sweat"); qqline(Sweat)
qqnorm(Sodium, main = "Sodium"); qqline(Sodium)
qqnorm(Potassium, main = "Potassium"); qqline(Potassium)
dev.off()

pdf("figures/sweat_pairs.pdf")
pairs(X)
dev.off()

##
## 95% confidence ellipse.
##

## The ellipse is centered at the sample mean vector, with axes equal to the eigenvectors
## of S.
ee <- eigen(S)
lambda <- ee$values
ee <- ee$vectors

## The scaled F percentile that defines the desired squared distance.
scaled_F <- ((p * (n - 1)) / (n - p)) * qf(0.95, p, n - p)
```

```
## The half-lengths of the ellipse.
sqrt((lambda / n) * scaled_F)

##
## 95% T^2 simultaneous confidence intervals for the mean components.
##

x_bar[1] + c(-1, 1) * sqrt(scaled_F) * sqrt(S[1, 1] / n)
x_bar[2] + c(-1, 1) * sqrt(scaled_F) * sqrt(S[2, 2] / n)
x_bar[3] + c(-1, 1) * sqrt(scaled_F) * sqrt(S[3, 3] / n)

##
## 95% Bonferroni simultaneous confidence intervals for the mean components.
##

x_bar[1] + c(-1, 1) * qt(1 - 0.05 / (2 * p), n - 1) * sqrt(S[1, 1] / n)
x_bar[2] + c(-1, 1) * qt(1 - 0.05 / (2 * p), n - 1) * sqrt(S[2, 2] / n)
x_bar[3] + c(-1, 1) * qt(1 - 0.05 / (2 * p), n - 1) * sqrt(S[3, 3] / n)

##
## Hotelling's T^2 test of H_0: mu' = [3.5, 54.5, 8.8]. This is equivalent to checking
## whether mu_0 is inside the 95% confidence ellipse from above. It is, so, as expected,
## we fail to reject H_0.
##

mu_0 <- c(4.0, 45.0, 10.0)
T2 <- n * t(x_bar - mu_0) %*% solve(S) %*% (x_bar - mu_0)
p_value <- 1 - pf(((n - p) / (p * (n - 1))) * T2, p, n - p)

##
## Bootstrap test, using Lambda = (|S| / |S_0|)^{n/2} as the test statistic. Under the
## assumption of multivariate normality, this equals the likelihood ratio statistic. If
## n were "big", and assuming we were happy assuming multivariate normality, we could
## compute a p-value via the large-sample chi-square result for likelihood ratio tests.
## However, we can *always* use the bootstrap, even if n is small and normality does not
## hold.
##

set.seed(101)

## Our observed value of Lambda.
S_0_f <- function(X, mu_0) {
  matrix(rowSums(apply(X, 1, function(x) { (x - mu_0) %*% t(x - mu_0) })), nrow = p) /
    (n - 1)
}
```

```
Lambda <- (det(S) / det(S_0_f(X, mu_0))) ^ (n / 2)

B <- 500
Lambda_b <- rep(NA, B)
X_0 <- scale(X, center = TRUE, scale = FALSE) + matrix(rep(mu_0, each = n), nrow = n)
for(b in 1:B) {
  cat(".")

  ## Create bootstrap sample.
  X_b <- X_0[sample(1:n, replace = TRUE), ]

  ## Compute Lambda.
  S_b <- var(X_b)
  S_0_b <- S_0_f(X_b, mu_0)
  Lambda_b[b] <- (det(S_b) / det(S_0_b)) ^ (n / 2)
}

## We again fail to reject H_0.
p_value_boot <- mean(Lambda_b <= Lambda)
```

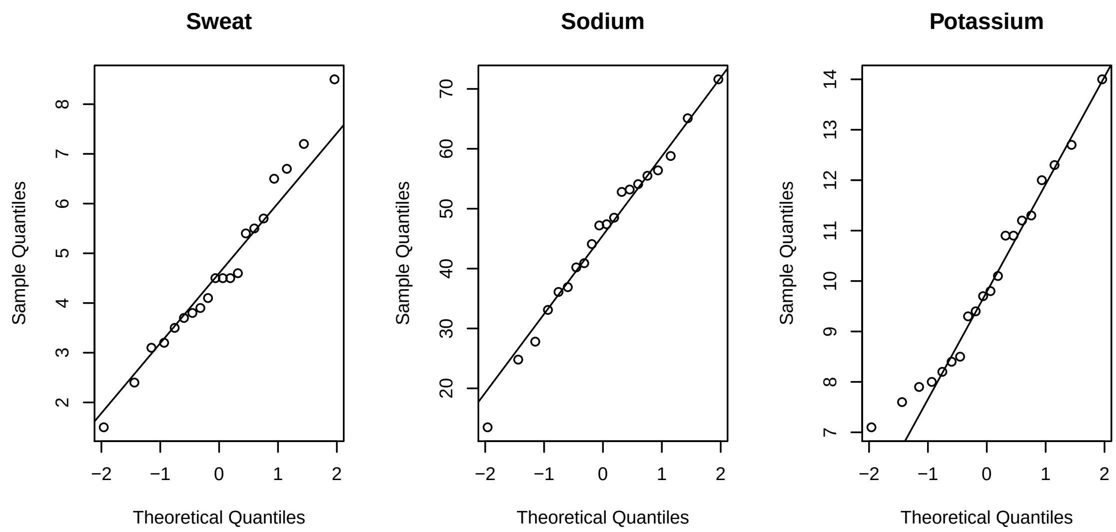

Figure 1: Q-Q plots for the sweat data variables. There is some relatively minor deviation from linearity for all three variables, but it looks pretty good overall.

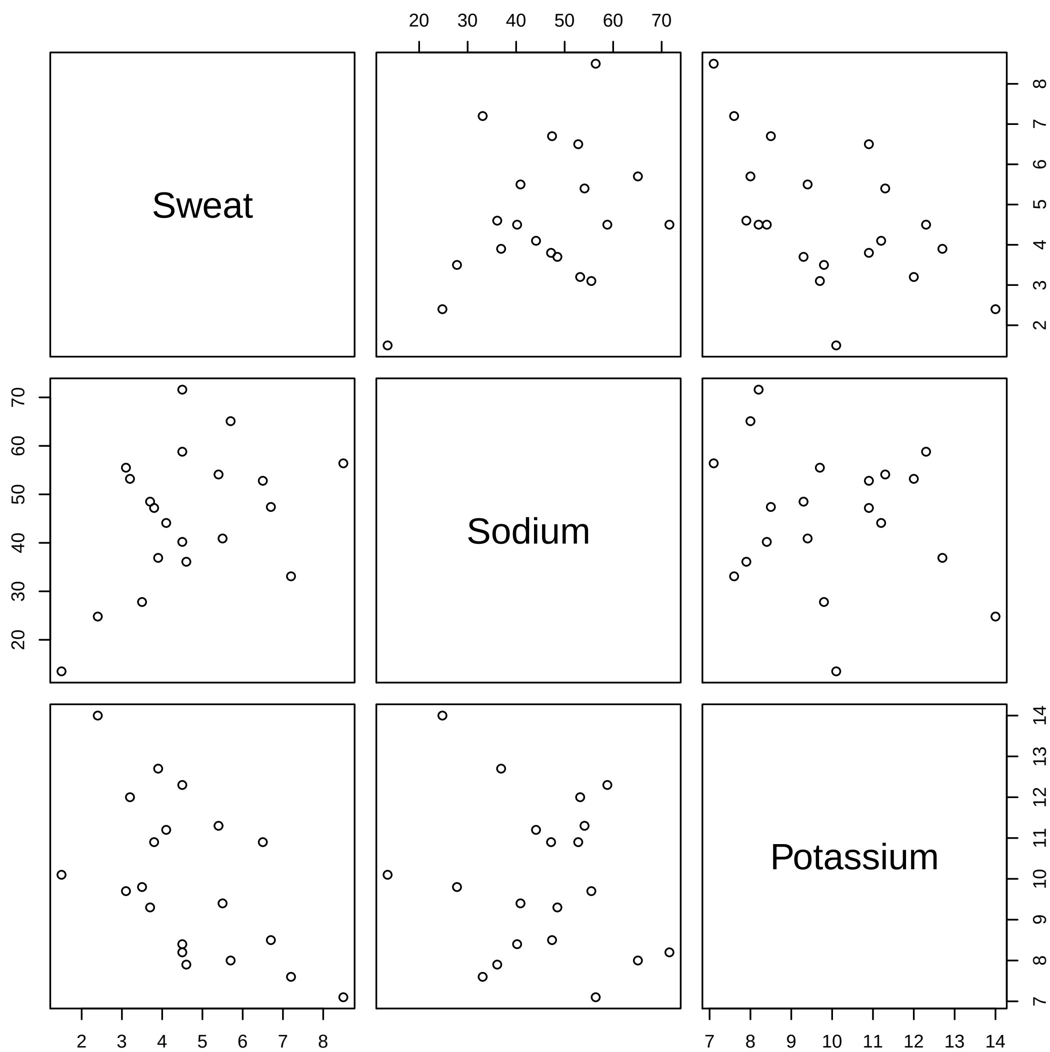

Figure 2: Pairs plots for the sweat data variables. The scatterplots look reasonably elliptical, and there are no obvious outliers.