Problem Set 1

Dr. Young

Carson Slater

2024-06-24

5.19

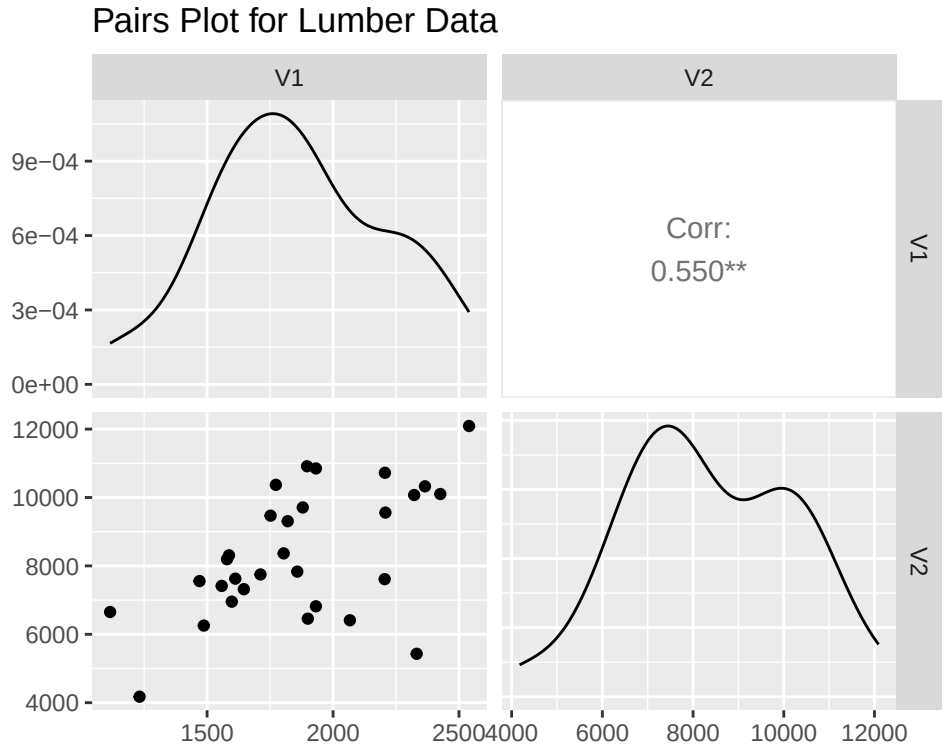
The equation

$$(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) = \frac{p(n-1)}{n(n-p)} F_{\alpha}(p,\, n-p)$$

defines a confidence ellipsoid for $\boldsymbol{\mu}$.

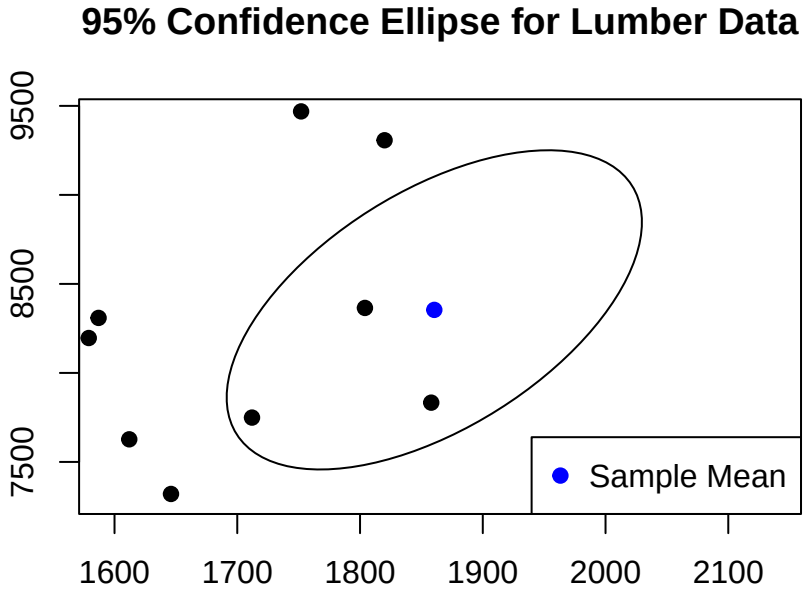
a

```
df <- read.delim("Wichern_5_19.txt", sep = '', header = FALSE)

# Observing the relationship between both variables
GGally::ggpairs(df, title = "Pairs Plot for Lumber Data")
```




```
xbar <- c(mean(df$V1), mean(df$V2))
sigma <- cov(df)
```

```r
# Getting the eigenvalues for axis lengths of the ellipse
e <- eigen(sigma)

n <- nrow(df)
p <- ncol(df)

# Hotellings T^2
t_alpha <- p*(n-1)*p/(n*(n-p))*qf(0.95, p, (n-p))

MVQuickGraphs::confidenceEllipse(
  X.mean = xbar, eig = e, n = 30,
  p = 2, axes = FALSE, x1 = c(1000, 2100),
  y1 = c(7000, 11000)
  )
points(df, pch = 19)
points(x = xbar[1], y = xbar[2], pch = 19, col = "blue")
points(x = 2000, y = 10000, pch = 19, col = "red") # does not show up
title("95% Confidence Ellipse for Lumber Data")
legend("bottomright", c("Sample Mean"), pch = c(19), col = c("blue"))
```



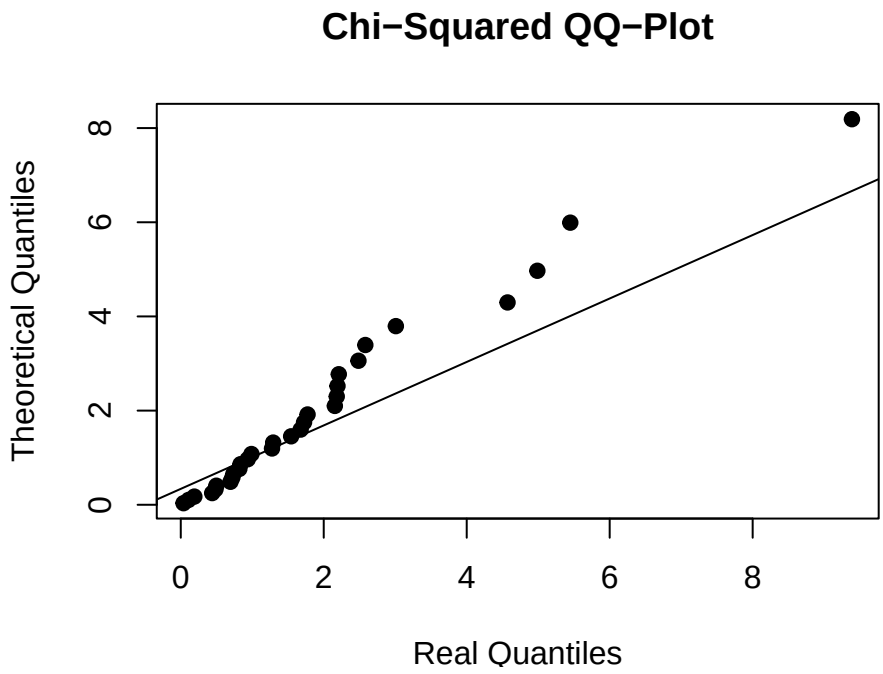
b

The proposed mean given does not appear to be consistent with those values, as it does not seem to even be near the ellipse. It does not show up on graph when plotted, as it falls outside the axes.

c

```r
# Calculating Mahalanobis Distance
dist <- mahalanobis(as.matrix(df), center = xbar, cov = sigma)

qqplot(dist, qchisq(ppoints(30), 2),
       main = "Chi-Squared QQ-Plot",
       ylab = "Theoretical Quantiles",
       xlab = "Real Quantiles",
```

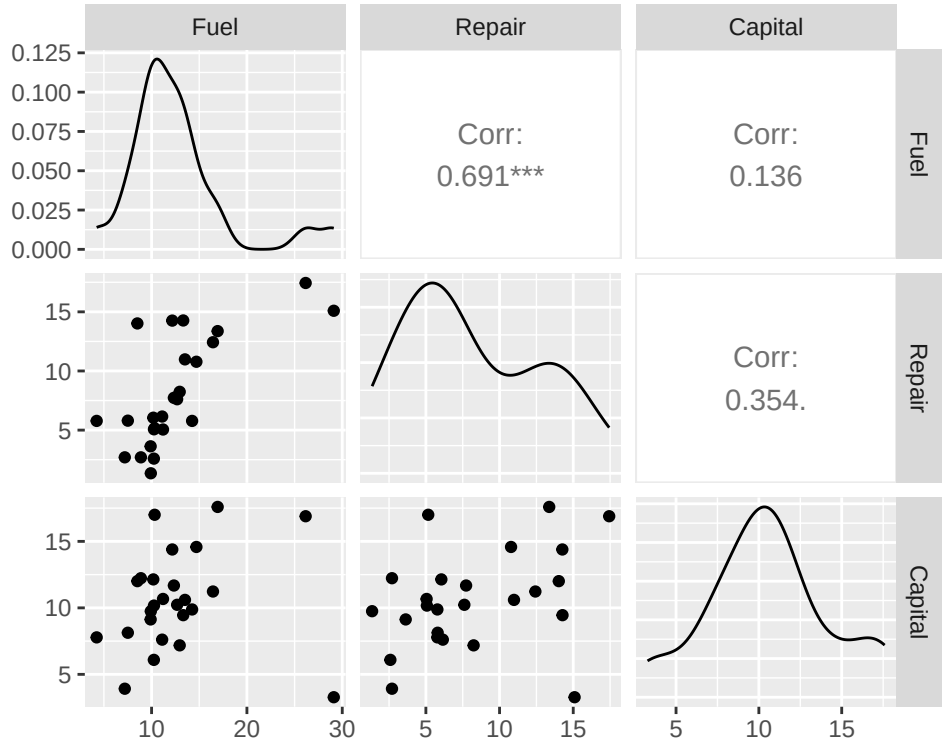
```
       pch = 19)
qqline(dist, distribution = \(prob) qchisq(prob, df = p))
```



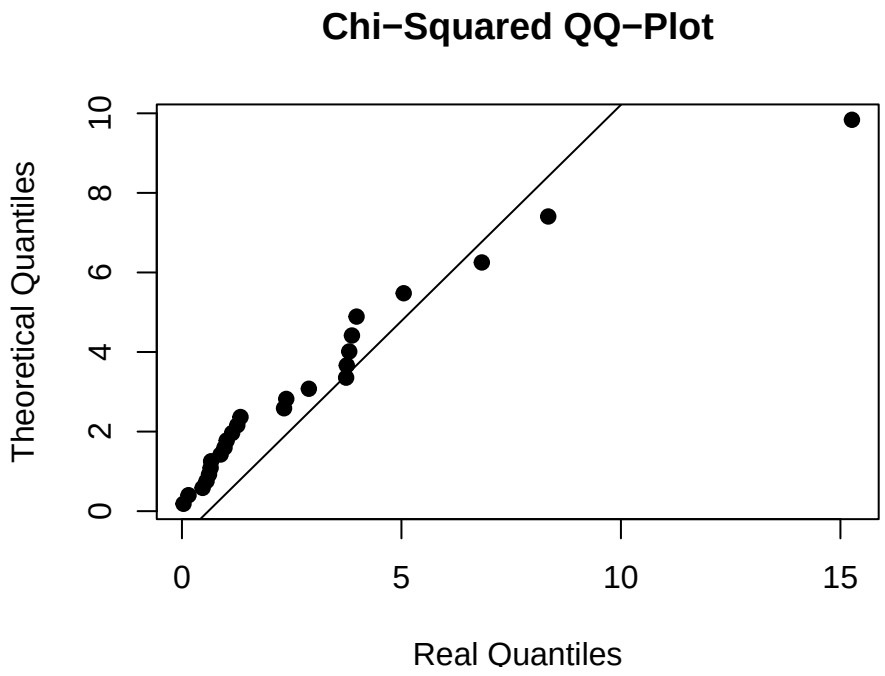
These data appear to be consistent with a multivariate normal distribution, until the tail of the distribution. This is fairly usual behavior, and so no assumptions would be eggregiously violated.

5.22

```
df <- read.table(here::here("datasets", "T5-13.dat"), sep ='', header = FALSE)
names(df) <- c("Fuel", "Repair", "Capital")

n <- nrow(df)
p <- ncol(df)
xbar <- colMeans(df)
se <- apply(df, 2, sd)/sqrt(n)
sigma <- cov(df)


GGally::ggpairs(df)
```




```
# testing Multivariate Normality
dist <- mahalanobis(as.matrix(df), center = xbar, cov = sigma)

qqplot(dist, qchisq(ppoints(n), p),
       main = "Chi-Squared QQ-Plot",
       ylab = "Theoretical Quantiles",
       xlab = "Real Quantiles",
       pch = 19)
qqline(dist, distribution = \(prob) qchisq(prob, df = p))
```



These data do not visually demonstrate large deviations from multivariate normality.

Bonferroni Intervals

```
bonferroni_alpha <- 0.05/(2*ncol(df))

t <- qt(1 - bonferroni_alpha / 2, df = (n - 1))
lower <- xbar - t * se
upper <- xbar + t * se

b_ints <- rbind(lower, upper)
rownames(b_ints) <- c("lower", "upper")

b_ints |> knitr::kable(align = 'c',
                       caption = "Bonferroni Intervals")
```

Table 1: Bonferroni Intervals

	Fuel	Repair	Capital
lower	9.465248	5.497872	8.42391
upper	15.654752	10.824528	12.66489

Hotelling's T^2 Intervals

```
sigma <- cov(df)

# Getting the eigenvalues for axis lengths of the ellipse
evals <- eigen(sigma)$values

# Hotellings T^2
t_alpha <- p*(n-1)*p/(n*(n-p))*qf(0.95, p, (n-p))
```

```
axes <- sqrt(evals*t_alpha)

t_ints <- rbind(xbar - axes/2, xbar + axes/2)
rownames(t_ints) <- c("lower", "upper")

t_ints |> knitr::kable(align = 'c',
                       caption = "Hotelling's $T^2$ Intervals") |>
  kableExtra::kable_styling()
```

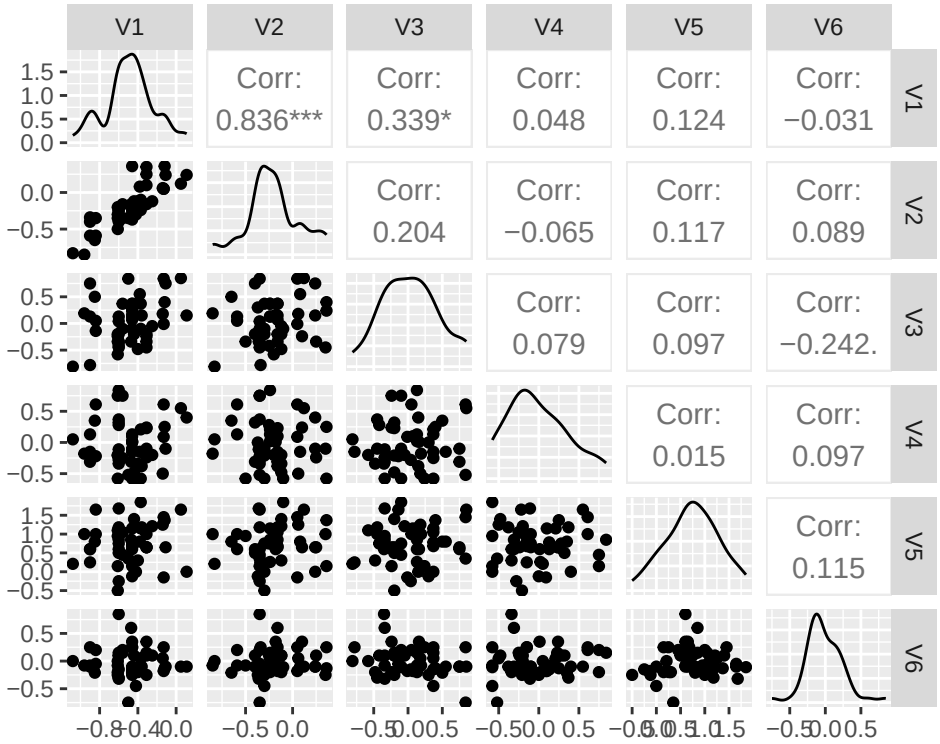
Table 2: Hotelling's T^2 Intervals

	Fuel	Repair	Capital
lower	8.931423	6.124595	9.184034
upper	16.188577	10.197805	11.904766

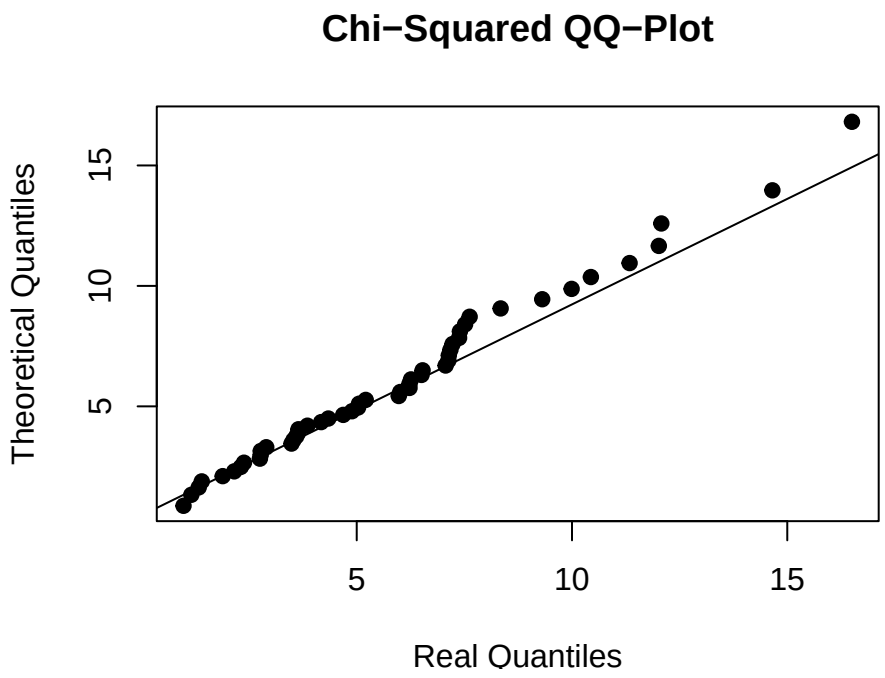
For `repair` and `capital`, the T^2 intervals were shorter, but for `fuel`, the Bonferroni interval was shorter.

5.29

```
df <- read.table(here::here("datasets", "T5-14.dat"), sep ='', header = FALSE)

xbar <- colMeans(df)
sigma <- cov(df)
inv_sigma <- solve(sigma)
n <- nrow(df)
p <- ncol(df)

GGally::ggpairs(df)
```




```
dist <- mahalanobis(as.matrix(df), center = xbar, cov = sigma)

# Looks pretty normal
qqplot(dist, qchisq(ppoints(n), p),
       main = "Chi-Squared QQ-Plot",
       ylab = "Theoretical Quantiles",
       xlab = "Real Quantiles",
       pch = 19)
qqline(dist, distribution = \(prob) qchisq(prob, df = p))
```



These data do not visually demonstrate large deviations from multivariate normality.

```
# Hotelling's T^2 test
# since mu_0 = 0, we have no mean vector in this equation
(test_stat <- t(sqrt(n) * xbar) %*% inv_sigma %*% (sqrt(n) * xbar))
```

```
##          [,1]
## [1,] 500.5834
```

```
(t_alpha <- p*(n-1)/(n-p)*qf(0.95, p, (n-p)))
```

```
## [1] 15.45681
```

```
# testing if observed test statistic is outside acceptance region
test_stat >= t_alpha
```

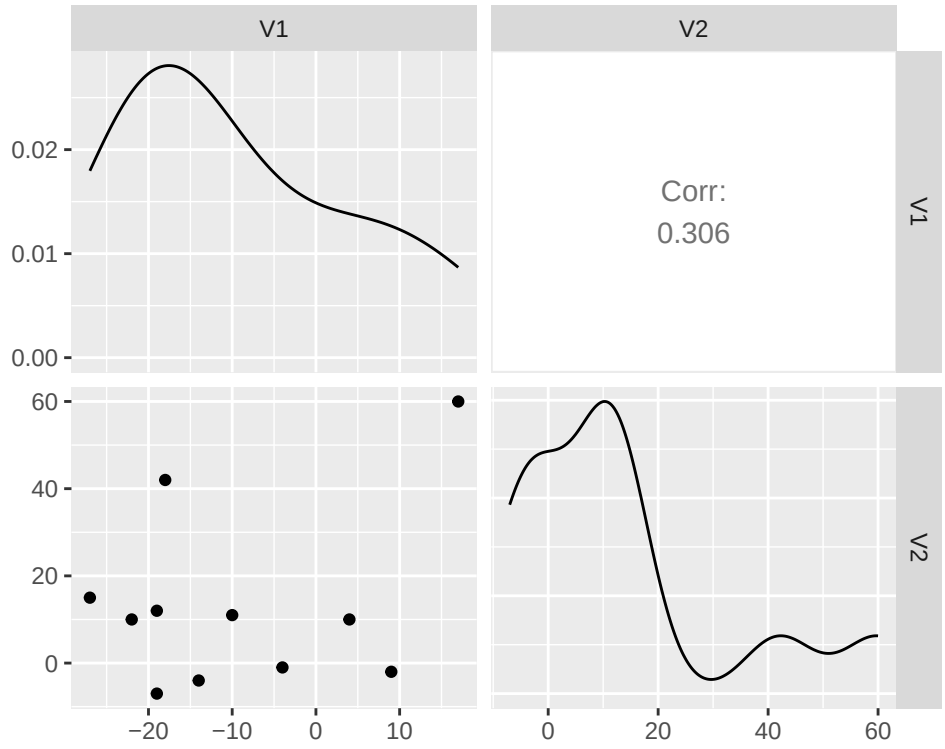
```
##      [,1]
## [1,] TRUE
```

We would reject the null hypothesis that $\boldsymbol{\mu}_0 = \mathbf{0}$ in favor of the alternative hypothesis, being $\boldsymbol{\mu}_0 \neq \mathbf{0}$.

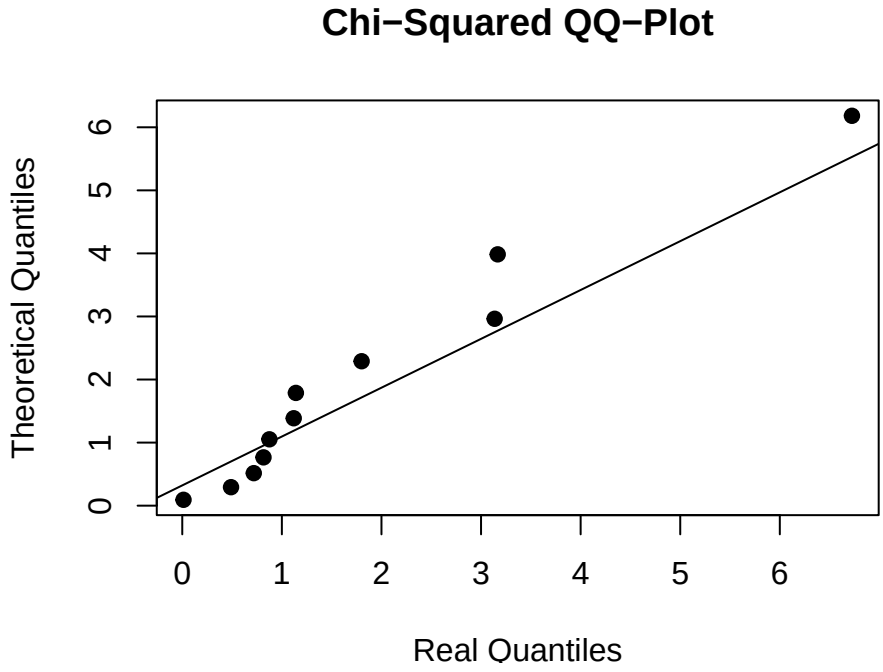
Effluent Data: Problem from the Notes

```
commercial <- matrix(c(6, 6, 18, 8, 11, 34, 28, 71, 43, 33, 20,
                       27, 23, 64, 44, 30, 75, 26, 124, 54, 30, 14),
                     nrow = 11, byrow = FALSE)
state <- matrix(c(25, 28, 36, 35, 15, 44, 42, 54, 34, 29, 39,
                  15, 13, 22, 29, 31, 64, 30, 64, 56, 20, 21),
                nrow = 11, byrow = FALSE)

diff <- commercial - state

dbar <- colMeans(diff)
dsigma <- cov(diff)
inv_sigma <- solve(dsigma)
```

```r
n <- nrow(diff)
p <- ncol(diff)

dist <- mahalanobis(diff, center = dbar, cov = dsigma)

# check pairs plot and normality assumption
GGally::ggpairs(as.data.frame(diff))
```




```r
qqplot(dist, qchisq(ppoints(n), p),
       main = "Chi-Squared QQ-Plot",
       ylab = "Theoretical Quantiles",
       xlab = "Real Quantiles",
       pch = 19)
qqline(dist, distribution = \(prob) qchisq(prob, df = p))
```



These data do not visually demonstrate large deviations from multivariate normality, though with a smaller sample size it is hard to tell.

```
(test_stat <- n*t(dbar) %*% inv_sigma %*% dbar)
```

```
##          [,1]
## [1,] 13.63931
```

```
(t_alpha <- p*(n-1)/(n-p)*qf(0.95, p, (n-p)))
```

```
## [1] 9.458877
```

```
# testing if observed test statistic is outside acceptance region
test_stat >= t_alpha
```

```
##      [,1]
## [1,] TRUE
```

We would reject the null hypothesis that $\boldsymbol{\mu}_{\mathbf{d}0} = \mathbf{0}$ in favor of the alternative hypothesis, being $\boldsymbol{\mu}_{\mathbf{d}0} \neq \mathbf{0}$.