Homework 3 Solution

Due: 3/3/2019 before 11pm. Submit in Canvas (file upload). Rmd file and the html output file (submit both files) are strongly recommended, but not required.

1. [**45 points**]

Consider data from a multi-nomial distribution with p categories. Let $X_1, \dots, X_n$ be independent identically distributed (i.i.d.), where $X_i = (X_{i1}, \dots, X_{ip})'$ and $X_{ij} = \mathbf{1}\{\text{the ith observation is from the jth category}\}$. Namely, only the jth component of X_i is 1 and all other components are 0 if this observation is from the jth category. The probability of an observation from each category is $\pi = (\pi_1, \dots, \pi_p)'$. Then $Z = \sum_{i=1}^n X_i$ is multi-nomial distributed with number of trials being n, and success probability being π. This success probability is estimated by $\widehat{\pi} = Z/n = \sum_{i=1}^n X_i/n$.

a. Clearly, $\mathrm{E}(X_i) = \pi$. Find the $\Sigma = \mathrm{Cov}(X_i)$. [10 points]

   Solution:
$$\Sigma_{jj} = Var(X_{ij}) = E(X_{ij} - \pi_j)^2 = \pi_j(1 - \pi_j)$$
$$\Sigma_{jk} = Cov(X_{ij}, X_{ik}) = E(X_{ij} - \pi_j)(X_{ik} - \pi_k) = -\pi_j\pi_k, \forall j \neq k$$

b. Assume the sample size n is large. To test for the hypothesis that π is equal to a specific vector, say $\pi_j = 1/p$ for all $1 \leq j \leq p$, can we use the Hotelling's T^2 statistic on the data $\{X_i\}_{i=1}^n$? If not, what statistic we should construct? [10 points]

   Solution: Since the sample size n is large and X_i are independent, we can use the asymptotic Hotelling T^2 statistics without normality assumption.

c. What is the limiting distribution of the test statistic you used for question (b)? [10 points]

   Solution: Since X is linear correlated, sample variance S is non-invertible. We use the first $p-1$ columns instead (you can choose any $p-1$ columns),
$$T^2 = n(\bar{X}_{1:p-1} - (\pi_0)_{1:p-1})' S_{1:p-1}^{-1} (\bar{X}_{1:p-1} - (\pi_0)_{1:p-1}) \xrightarrow{d} \chi_{p-1}^2$$

d. We want to construct simultaneous confidence intervals for π_j for $j = 1, \dots, p$ and possibly some of their contrasts. List two ways of constructing such simultaneous confidence intervals. [10 points]

   Solution: The T^2 CI is given by
$$\bar{x}_j \pm \sqrt{\chi_{p-1}^2(1-\alpha)\frac{s_{jj}}{n}}$$
   The Bonferroni CI is given by
$$\bar{x}_j \pm N\left(1 - \frac{\alpha}{2p}\right)\sqrt{\frac{s_{jj}}{n}}$$

e. Which way in question (d) you prefer, why? [5 points]

   Solution: The Bonferroni CI is preferred since we only wish to make p comparisons, not all the possible linear combinations of π. T^2 CI is too conservative.

2. [**20 points**]

For the `Effluent Study` in our lecture `InferenceForMeans-Repeated`, the conclusion from the Hotelling's T^2 test doesn't agree with the result from the simultaneous confidence intervals for the difference of the means of the two variables. Possible reasons include outliers in the data, sample size is small and the data is not normal distributed. The data is in our Canvas data folder, named `effluent.dat`, with the code `effluent.R`.

a. Test for Multivariate normality assumption on the data; [5 points]

```
library(tidyverse)

set.seed(1)
effluent <- read.table("effluent.dat")
energy::mvnorm.test(effluent[, -1], R = 1e4)
```

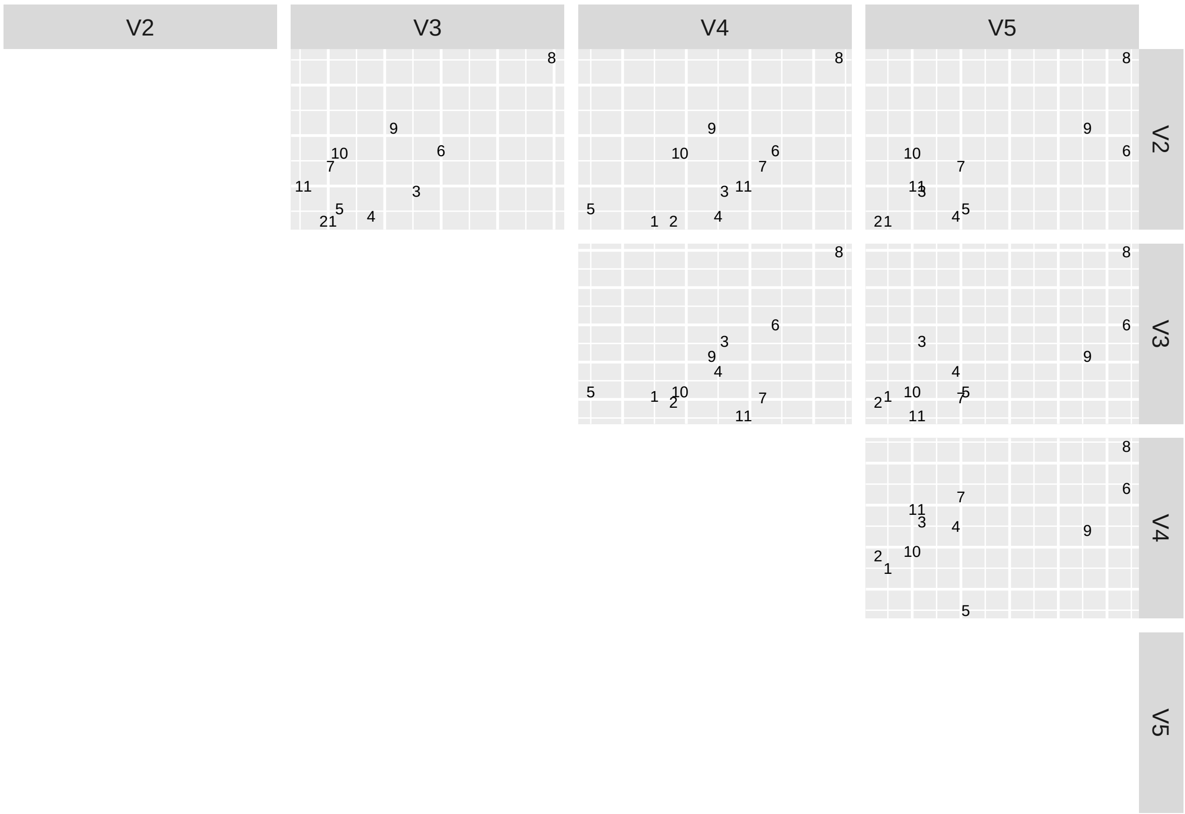
```
##
##  Energy test of multivariate normality: estimated parameters
##
## data:  x, sample size 11, dimension 4, replicates 10000
## E-statistic = 1.0537, p-value = 0.0851
```

Cannot reject the null hypothesis that the data is multivariate normal. Other normality tests are also accepted.

b. Try a more appropriate test for the hypothesis $\mu_{state} = \mu_{private}$. You can delete the possible outlier, or transform the data to normal first (no longer testing for $\mu_{state} = \mu_{private}$ on the original scale), or use permutation test (How to do permutation test for paired data?). [15 points]

Solution: Method 1: removing possible outliers.

```
GGally::ggpairs(effluent,
            mapping = aes(label = V1),
            columns = 2:5,
            upper = list(continuous = function(data, mapping, ...) ggplot(data, mapping) +
                              geom_text(size = 2, ...)),
            lower = list(continuous = "blank"),
            diag = list(continuous = "blankDiag"),
            axisLabels = "none")
```



```
# Remove the eighth point and perform the Hotelling's T2 test
Cstar <- matrix(c(1, 0, -1, 0, 0, 1, 0, -1), 2, 4, byrow = T)
ICSNP::HotellingsT2(as.matrix(effluent[-8, -1]) %*% t(Cstar), mu = c(0, 0))
```

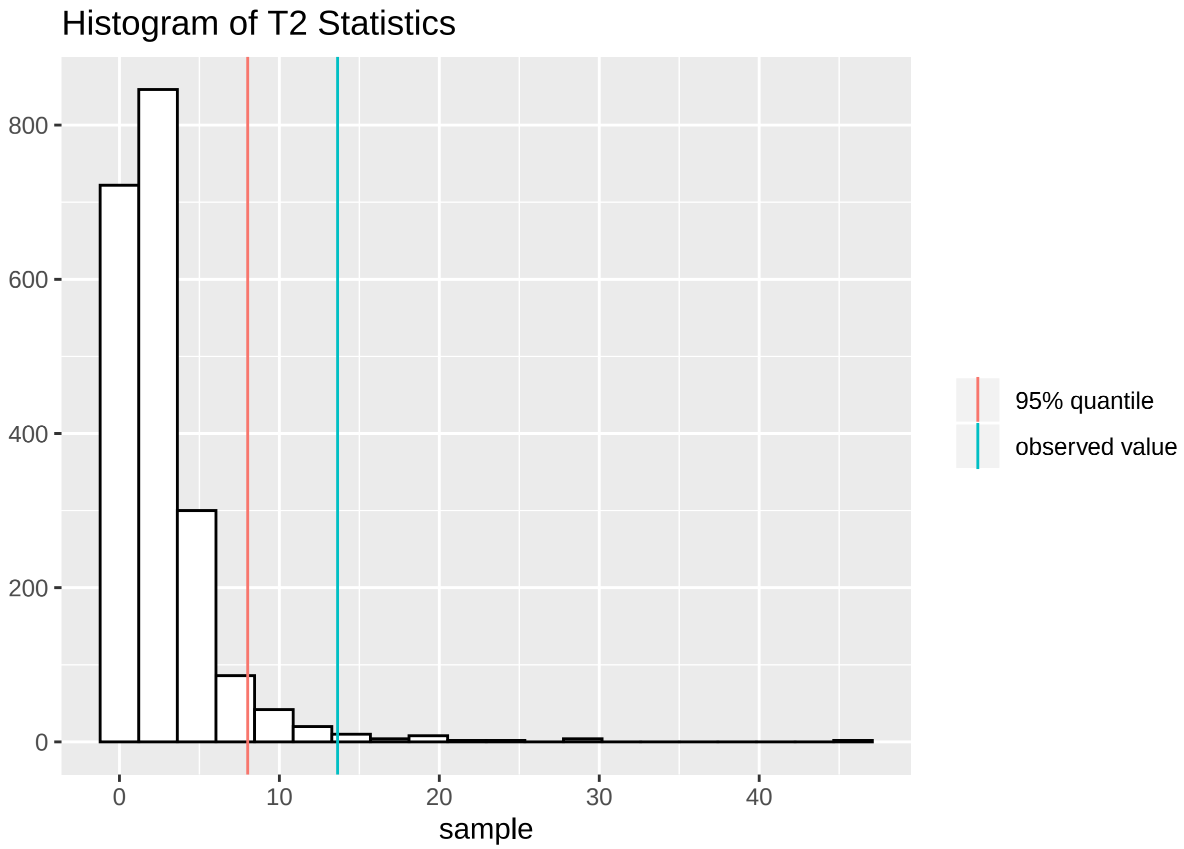
```
##
##  Hotelling's one sample T2-test
##
## data:  as.matrix(effluent[-8, -1]) %*% t(Cstar)
## T.2 = 5.0873, df1 = 2, df2 = 8, p-value = 0.03754
## alternative hypothesis: true location is not equal to c(0,0)
```

Method2: permutation test.

Use permutation test for paired data by trying all combinations of swapping the data of two groups in each pair (11 pairs in total). So there will be $2^{11} = 2048$ permutations. Then calculate Hotellings T^2 statistics for each permutation.

```
T2 <- function(x) nrow(x) * as.numeric(colMeans(x) %*% solve(var(x)) %*% colMeans(x))

diff <- as.matrix(effluent[2:3] - effluent[4:5])
perm <- expand.grid(rep(list(c(1, -1)), nrow(effluent)))
sample <- apply(as.matrix(perm), 1, function(i) T2(diff * i))
qplot(sample, fill = I("white"), color = I("black"), bins = 20) +
  geom_vline(aes(xintercept = x, color = y),
         data.frame(x = c(T2(diff), quantile(sample, .95)),
                    y = c("observed value", "95% quantile"))) +
  labs(title = "Histogram of T2 Statistics", color = NULL)
```

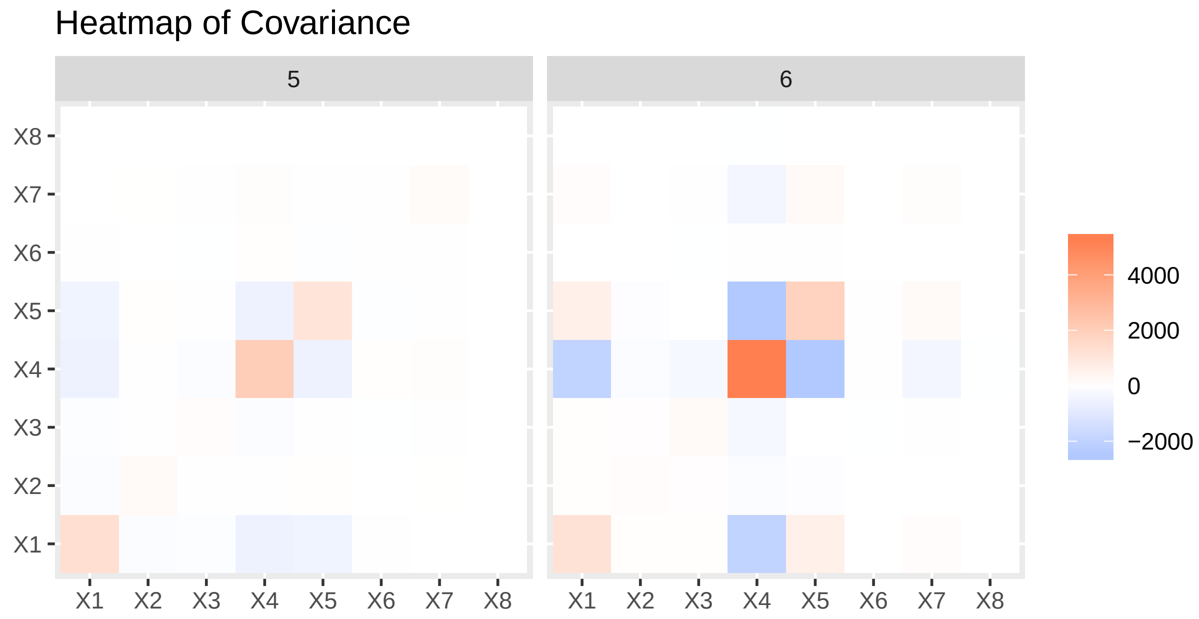


3. [35 points]

The Olive Oils dataset `olive.dat` (provided in the folder `Data and code` in Canvas) consists of measurements of eight chemical components (columns 2-8) on 572 samples of olive oil. The samples come from three different regions of Italy. The regions are further partitioned into nine areas: areas 1–4 belong to region R1, areas 5 and 6 belong to region R2, and areas 7–9 belong to region R3. The first column of the file provides the indicator for the nine regions. (Note that the file has a header.) We will now focus on Region R2. Answer the following questions:

a. Calculate the covariance matrix for the chemical composition in each of the two sub-regions. Display the covariance matrices for the two sub-regions side-by-side, and comment on possible differences. You can use the heatmap to visualize the two matrices. Be sure to use the same scale for the color. [10 points]

```r
olive <- read.table("olive.dat", header = T) %>% filter(group.id %in% 5:6)
olive %>% nest(data = -group.id) %>%
  mutate(cov = map(data, ~ reshape2::melt(var(.)))) %>%
  select(-data) %>% unnest(cols = c(cov)) %>%
  qplot(Var1, Var2, fill = value, data = ., facets = ~ group.id, geom = "tile") +
  labs(title = "Heatmap of Covariance", x = NULL, y = NULL, fill = NULL) +
  scale_fill_gradient2(low = "dodgerblue", mid = "white", high = "coral", midpoint = 0) +
  coord_fixed()
```



They have similar patterns. On average, the covariance on group 6 is larger than group 5.

b. Test for equality of the mean chemical compositions between the two groups. Specifically, report the Hotelling's T^2 statistics and its p-value. Are you comfortable to use Hotellings' T^2 test without testing for the multivariate normality of the data and the equivalence of the two covariance matrices? [10 points]

Solution: From (a), we assume two groups have same variance.

```r
n <- nrow(olive); p <- ncol(olive) - 1
T2test <- nest(olive, data = -group.id)$data %>% {ICSNP::HotellingsT2(.[[1]], .[[2]])}
T2stat <- T2test$statistic[, ] * (n - 2) * p / (n - p - 1)
T2test
```

```
##
##  Hotelling's two sample T2-test
##
## data:  .[[1]] and .[[2]]
## T.2 = 112.41, df1 = 8, df2 = 89, p-value < 2.2e-16
## alternative hypothesis: true location difference is not equal to c(0,0,0,0,0,0,0,0)
```

The statistic above is the F statistc, and the T^2 statistic is

$$T^2 = \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F = 969.973$$

We reject the null that the mean is equal. Since n_1, n_2 are large, we can use asymptotic distribution of T^2 so normal assumption is unnecessary, but we still need the covariance matrix to be homogeneous.

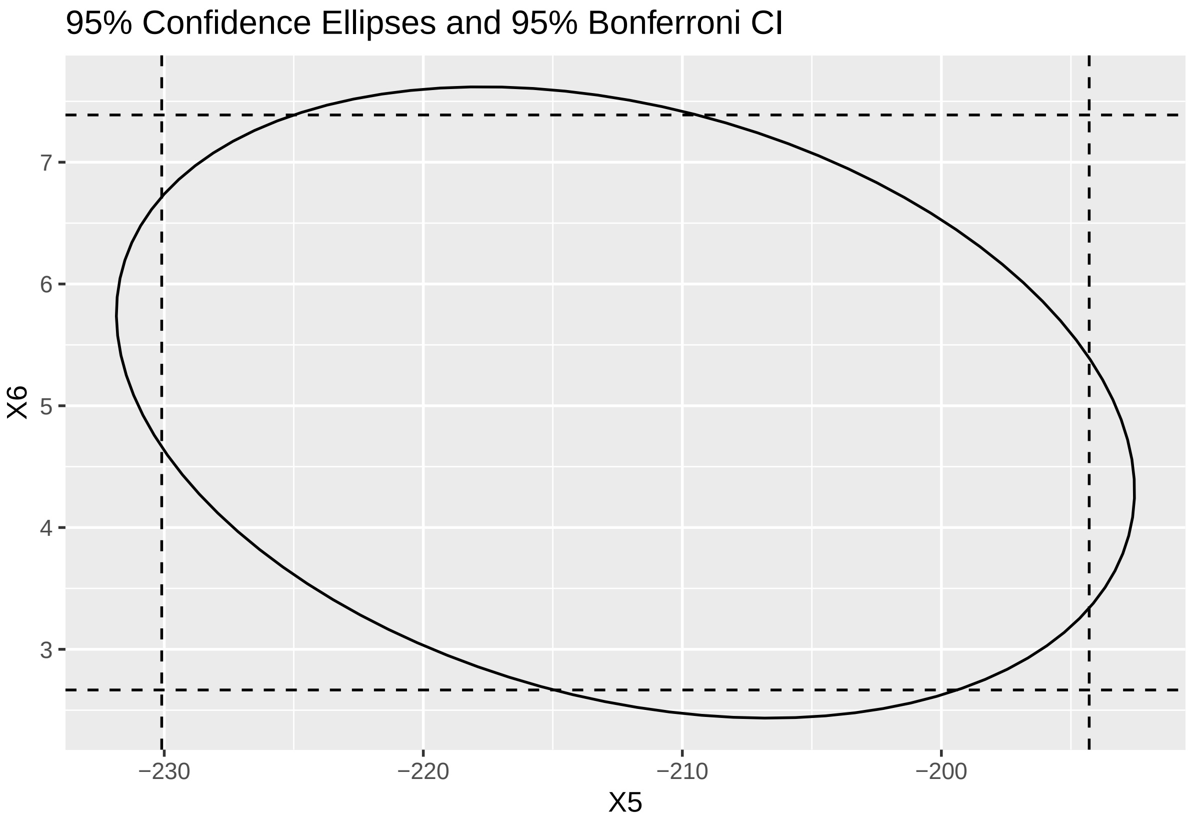
c. Restrict attention to the coordinates for the fifth and sixth chemicals. Provide individual pairwise t-tests for the differences in the composition of the two chemicals among the two groups, using Bonferroni method to adjust the 5% level of significance. Plot their Bonferroni adjusted confidence intervals, and on the same plot, draw the 95% confidence ellipses for the two groups. (You may use the ellipse package.) [15 points]

```r
olive2 <- select(olive, group.id, X5:X6)
test <- c(X5 = X5 ~ group.id, X6 = X6 ~ group.id) %>%
  lapply(function(f) t.test(f, olive2, var.equal = T, conf.level = 1 - .05 / 2))
test
```

```
## $X5
##
##  Two Sample t-test
##
## data:  X5 by group.id
## t = -26.991, df = 96, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 97.5 percent confidence interval:
##  -230.0974 -194.2942
## sample estimates:
## mean in group 5 mean in group 6
##        1125.077        1337.273
##
##
## $X6
##
##  Two Sample t-test
##
## data:  X6 by group.id
## t = 4.8483, df = 96, p-value = 4.793e-06
## alternative hypothesis: true difference in means is not equal to 0
## 97.5 percent confidence interval:
##  2.665971 7.388108
## sample estimates:
## mean in group 5 mean in group 6
##        28.78462        23.75758
```

```r
p <- ncol(olive2) - 1
data1 <- filter(olive2, group.id == 5) %>% select(-group.id)
data2 <- filter(olive2, group.id == 6) %>% select(-group.id)
n1 <- nrow(data1); n2 <- nrow(data2)
diff <- colMeans(data1) - colMeans(data2)
S <- (var(data1) * (n1 - 1) + var(data2) * (n2 - 1)) / (n1 + n2 - 2) * (1 / n1 + 1 / n2)
c2 <- (n1 + n2 - 2) * p / (n1 + n2 - p - 1) * qf(.95, p, n1 + n2 - p - 1)
bonCI <- lapply(test, function(x) x$conf)  # diff +- qt(1 - .05/4, n1 + n2 - 2) * sqrt(S)

ellipse::ellipse(S, centre = diff, t = sqrt(c2)) %>% as.data.frame() %>%
qplot(X5, X6, data = ., geom = "path", main = "95% Confidence Ellipses and 95% Bonferroni CI") +
  geom_vline(xintercept = bonCI$X5, linetype = 2) +
  geom_hline(yintercept = bonCI$X6, linetype = 2)
```


95% Confidence Ellipses and 95% Bonferroni CI
X6
7
6
5
4
3
−230
−220
−210
−200
X5