Homework 2

Zahlen Zbinden

1. The scored obtained by n $= 87$ college students on the College Level Examination Program (CLEP) test for social science and history (X_1) and on the College Qualification Test (CQT) for verbal (X_2) and science (X_3) are given in the TestScores.csv file.

   a. Test the null hypothesis $H_o : \mu = \begin{pmatrix} 500 & 50 & 30 \end{pmatrix}^T$ vs the alternative $H_A : \mu \neq \begin{pmatrix} 500 & 50 & 30 \end{pmatrix}$ at significance level $\alpha = 0.05$ using simultaneous univariate t-test with Bonferonni correction. Would this multivariate hypothesis be rejected?

   Using simultaneous univariate t-test with a Bonferonnie correction I would reject the Null hypothesis that the mean vector for the population is $\begin{pmatrix} 500 & 50 & 30 \end{pmatrix}$ in favor of the alternative. Non of the following tests have a p value > 0.05, as well non have a test statistic that is less than the corrected critical value.

```
test_scores <- read_csv("D:/RepoMan/osu/data/TestScores.csv")
mu <- c(500, 50, 30)
p = ncol(test_scores)
n = nrow(test_scores)
```

Calculate Bonferroni corrected p-value for the following univariate tests.

```
alpha_star <- .05 / p
b_crit <- qt(alpha_star / 2, df = n - 1, lower.tail = F)
```

Test for X_1.

```
test_1 <- t.test(test_scores[, 1], mu = mu[1])

test_1$p.value
```

```
[1] 0.001628171
```

```
test_1$statistic
```

```
        t
3.253872
```

```
abs(test_1$statistic) > b_crit
```

```
   t
TRUE
```

Test for X_2

```
test_2 <- t.test(test_scores[, 2], mu = mu[2])

test_2$p.value
```

```
[1] 0.0001929104
```

```
test_2$statistic
```

```
        t
3.896034
```

```
abs(test_2$statistic) > b_crit
```

```
   t
TRUE
```

Test for X_3

```
test_3 <- t.test(test_scores[, 3], mu = mu[3])

test_3$p.value
```

```
[1] 5.812582e-15
```

```
test_3$statistic
```

```
        t
-9.455619
```

```
abs(test_3$statistic) > b_crit
```

```
   t
TRUE
```

1.   b. Test the null hypothesis vs the alternative at significance level 0.05 using Hotelling T^2 test. Would the multivariate hypothesis be rejected?

We can see running the Hotellings T^2 test, that we have a p-value of 0.4831 which is »0.05. I am not able to reject the null hypothesis.

```
HotellingsT2(test_scores, t(mu))
```

```
    Hotelling's two sample T2-test

data:  test_scores and t(mu)
T.2 = 0.8262, df1 = 3, df2 = 84, p-value = 0.4831
alternative hypothesis: true location difference is not equal to c(0,0,0)
```

2. A group of n $= 66$ students were given two different reading tests, Test 1 and Test 2, both before and after participating in a reading instruction program. Each student produced four test scores: pre1, pre2, post1, and post2. the pre and post-instruction scores for both tests are given in the ReadingTest.csv data file.

   a. Is a paired test appropriate to test $o : \mu_1 = \mu_2$ vs $H_A : \mu_1 \neq \mu_2$?

Yes because the same set of variables are measured under two different conditions, one of the conditions is the pre instruction test1/2 and the other is the post instruction test1/2

2.   b. Test $H_o : \mu_1 = \mu_2$ vs $H_o : \mu_1 \neq \mu_2$ using simultaneous univariate hypothesis tests with Bonferroni correction. Based on the result of these hypothesis test, would you conclude that the reading instruction produces a difference in performance on the two tests?

We can see from the bonferroni corrected p values for the following two univariate t-tests, for test 1 we have a p-value of .0027 « 0.05 and for test 2 we have a p-value of .0002 « 0.05. Because of this I reject the Null hypothesis that the true difference in means between the pairs is 0.

```
read <- read %>%
    mutate(test1_dif = PRE1 - POST1) %>%
    mutate(test2_dif = PRE2 - POST2)
```

```
test1 <- t.test(read$PRE1, read$POST1)
test2 <- t.test(read$PRE2, read$POST2)
```

```
p_adjusted_1 <- p.adjust(test1$p.value, method = "bonferroni")
p_adjusted_2 <- p.adjust(test2$p.value, method = "bonferroni")
```

```
p_adjusted_1
```

```
[1] 0.002682787
```

```
p_adjusted_2
```

```
[1] 0.0002310053
```

2. c. Perform a level 0.05 test of $H_o : \mu_1 = \mu_2$. Based on the result of this hypothesis test, would you conclude that the reading instruction produces a difference in performace of the two tests?

I would conclude that reading instruction produces a difference in performance of the two tests, as the p-value for the Hotelling Test is « 0.05.

```
HotellingsT2(read[1:2], read[3:4])
```

```
    Hotelling's two sample T2-test

data:  read[1:2] and read[3:4]
T.2 = 14.32, df1 = 2, df2 = 129, p-value = 2.418e-06
alternative hypothesis: true location difference is not equal to c(0,0)
```

2.   d. Now suppose you wanted to construct a confidence region for the difference in mean
          vectors. Would [1 1] be in that confidence region?

For construction purposes we would want to take the point estimate of the mean vector and
+/- the critical value * standard error.

For out 95% confidence interval we get a lower bound of [1.57, -1.75] and an upper bound of
[1.85, -1.75]

```
crit <- qf(1 - 0.05, 2, 129)
se_1 <- sd(read$test1_dif) / nrow(read)
se_2 <- sd(read$test2_dif) / nrow(read)
point_estimate <- colMeans(read[5:6])

point_estimate + c(se_1, se_2) * crit
```

```
test1_dif test2_dif
 1.851804 -1.463294
```

```
point_estimate - c(se_1, se_2) * crit
```

```
test1_dif test2_dif
 1.572439 -1.748827
```

The other side of this question could be asking what if we tested if the true population mean
difference vector was [1, 1] instead of [0, 0], what would that test tell us? Based on my
confidence intervale it would tell us that we would reject a null hypothesis stating that the
population mean difference was [1, 1]. We can see from the following test that the p-value
« 0.05 which gives us significant evidence against the null hypothesis. Confirming that our
confidence region and the test produced the same results.

```
HotellingsT2(read[1:2], read[3:4], mu = c(1, 1))
```

```
    Hotelling's two sample T2-test

data:  read[1:2] and read[3:4]
T.2 = 21.685, df1 = 2, df2 = 129, p-value = 7.606e-09
alternative hypothesis: true location difference is not equal to c(1,1)
```

3. Researches have suggested that a change in skull size over time is evidence of the intergreeding of a resident population with immigrant populations. Samples of 30 male egyptian skulls were obrained for five different time period. For each skull measurements of four different dimensional variables were taken.

   a. Compute and compare the covariance matrices for each of the five time periods. Do they seem approsimately similar? Would you be comforable assuming that the population covariance matrices are the same for these five different time periods?

   I would have a hard time determining if they were appropriately similar just from looking at them, my intuition tells me that they are pretty different. The numbers are not as close as I would like them to be to say "the covariance matrix" are the "same" for each time period.

```
skull <- read_csv("D:/RepoMan/osu/data/SkullData.csv")
```

```
first <- skull %>% filter(Year == -4000) %>% subset(select = -Year)
second <- skull %>% filter(Year == -3300) %>% subset(select = -Year)
third <- skull %>% filter(Year == -1850) %>% subset(select = -Year)
fourth <- skull %>% filter(Year == -200) %>% subset(select = -Year)
fifth <- skull %>% filter(Year == 150) %>% subset(select = -Year)
```

```
fst_cov <- cov(first)
sc_cov <- cov(second)
th_cov <- cov(third)
fr_cov <- cov(fourth)
ft_cov <- cov(fifth)
```

```
fst_cov
```

```
          MB         BH         BL         NH
MB 26.309195   4.1517241  0.4540230  7.2459770
BH  4.151724  19.9724138 -0.7931034  0.3931034
BL  0.454023  -0.7931034 34.6264368 -1.9195402
NH  7.245977   0.3931034 -1.9195402  7.6367816
```

```
sc_cov
```

```
          MB        BH        BL        NH
MB 23.136782  1.010345  4.7678161 1.8425287
```

```
BH  1.010345 21.596552   3.3655172 5.6241379
BL  4.767816   3.365517 18.8919540 0.1908046
NH  1.842529   5.624138  0.1908046 8.7367816
```

  th_cov

```
          MB            BH         BL           NH
MB 12.1195402  0.78620690 -0.7747126   0.89885057
BH  0.7862069 24.78620690  3.5931034 -0.08965517
BL -0.7747126  3.59310345 20.7229885  1.67011494
NH  0.8988506 -0.08965517  1.6701149 12.59885057
```

  fr_cov

```
          MB         BH         BL       NH
MB 15.362069 -5.534483 -2.172414 2.051724
BH -5.534483 26.355172  8.110345 6.148276
BL -2.172414  8.110345 21.085057 5.328736
NH  2.051724  6.148276  5.328736 7.964368
```

  ft_cov

```
          MB         BH         BL         NH
MB 28.6264368 -0.2298851 -1.8793103 -1.9942529
BH -0.2298851 24.7126437 11.7241379  2.1494253
BL -1.8793103 11.7241379 25.5689655  0.3965517
NH -1.9942529  2.1494253  0.3965517 13.8264368
```

3. b. Perform separate univariate ANOVA for each of the four variables at level $\alpha^* = \frac{\alpha}{p}$ with $\alpha = 0.05$. Are any of these univariate ANOVAs significant?

All of the results are statistically significant, each single univariate test has a p-value < 0.05, which suggests that we reject the Null hypothesis that the skull size is the same over the years in each case.

Univariate Test for MB

```
skull %>%
    select(c("Year", "MB")) %>%
    aov(MB ~ Year, data = .) %>%
    summary()
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
Year         1  491.3   491.3   23.66 2.9e-06 ***
Residuals  148 3072.6    20.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Univariate Test for BH

```
skull %>%
    select(c("Year", "BH")) %>%
    aov(BH ~ Year, data = .) %>%
    summary()
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Year         1    120  119.56   5.033 0.0263 *
Residuals  148   3516   23.75
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Univariate Test for BL

```
skull %>%
    select(c("Year", "BL")) %>%
    aov(BL ~ Year, data = .) %>%
    summary()
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
Year         1    780   780.0   32.71 5.73e-08 ***
Residuals  148   3529    23.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Univariate test for NH

```r
skull %>%
    select(c("Year", "NH")) %>%
    aov(NH ~ Year, data = .) %>%
    summary()
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Year         1   44.1   44.11   4.384  0.038 *
Residuals  148 1489.2   10.06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. c. Perform a level 0.05 test of the hypothesis that the population mean vectors for all
        of these five time periods are the same.

Based on the results of the manova test of this hypothesis I significant evidence that that the
skull size changes over the years, as the p-value is < 0.05. I do believe that the skull size has
been shows to change over the years, and that perhaps interbreeding could be a cause.

```r
response <- cbind(skull$MB, skull$BH, skull$BL, skull$NH)
result <- manova(response ~ skull$Year)
summary(result)
```

```
            Df  Pillai approx F num Df den Df    Pr(>F)
skull$Year   1 0.29569   15.219      4    145 2.06e-10 ***
Residuals  148
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```