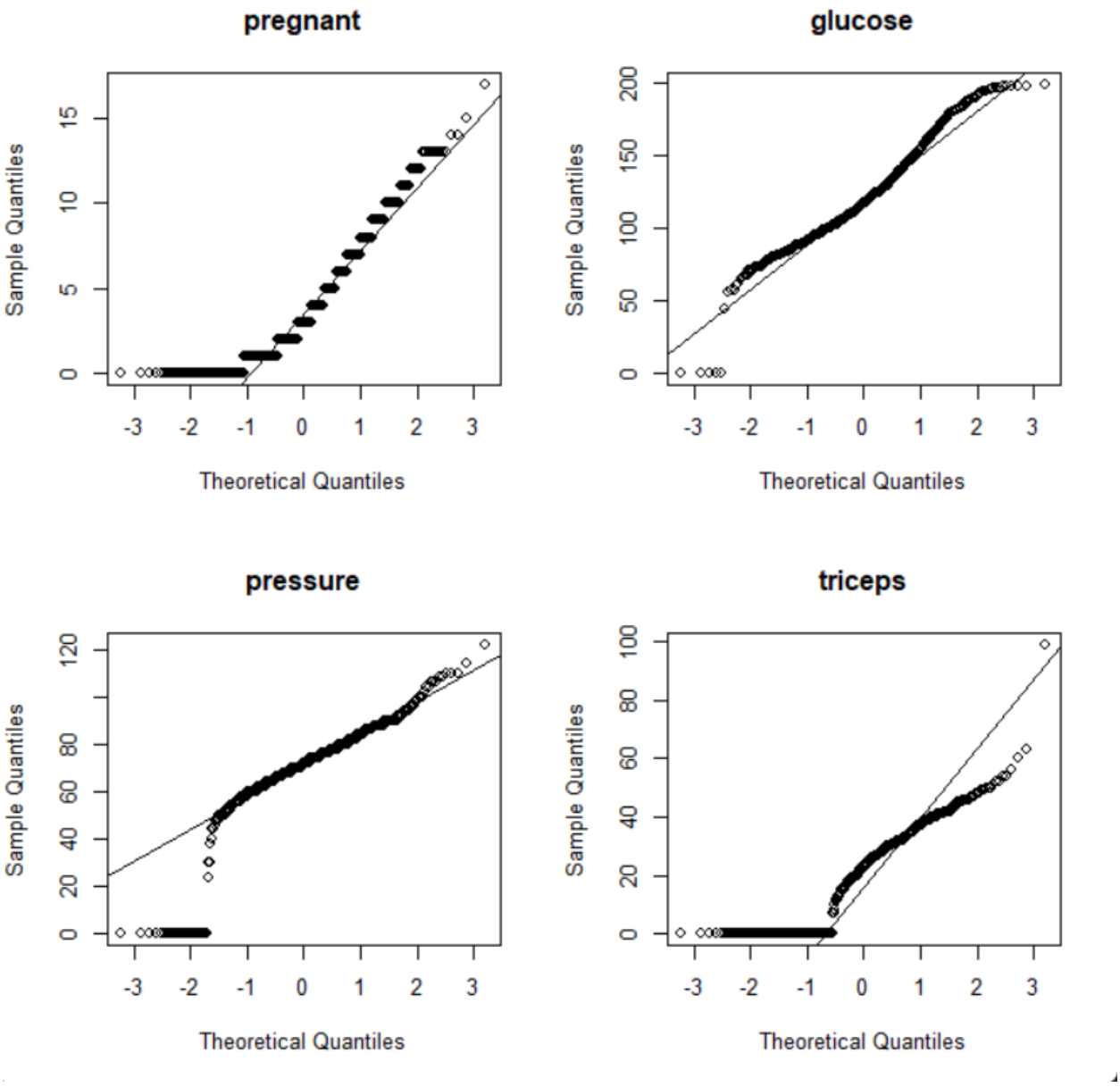
MULTIVARIATE STATISTICS

Assignment 2

STUDENT NAME: SRINIVAS VEGISETTI
STUDENT ID: 21080840
My GitHub link: Click Here...

Multivariate Statistics - CW2/Tutorial Sheet 4 Multivariate T-tests

Q1) a (i) Output:



Conclusion:

Based on the Q-Q plots generated for each variable, we can conclude that the assumption of multivariate normality may not be entirely valid for this dataset. The Q-Q plot for the "plas" variable

STUDENT NAME: SRINIVAS VEGISETTI
STUDENT ID: 21080840
My GitHub link: Click Here…

shows some deviation from the straight line, indicating that this variable may not follow a normal distribution. The Q-Q plot for the "pres" variable also shows some deviation from normality, although not as pronounced as the "plas" variable. The Q-Q plots for the "preg" and "skin" variables appear to be relatively close to normal, but there are some outliers that may indicate non-normality.

Overall, while the Q-Q plots suggest that the assumption of multivariate normality may not be strictly valid, the deviations from normality are not severe, and the assumption may be reasonable for some analyses.

However, further analysis and confirmation may be needed before relying on this assumption for any specific purpose.

Q1) a (ii) Output:

```
> # Calculate the sample mean vector
> mean_vector <- colMeans(sample.data, na.rm = TRUE)
> cat("Sample Mean Vector: \n")
Sample Mean Vector:
> print(mean_vector)
  pregnant    glucose   pressure    triceps
  3.845052 120.894531  69.105469  20.536458
>
> # Calculate the sample covariance matrix
> cov_matrix <- cov(sample.data, use="pairwise.complete.obs")
> cat("\nSample Covariance Matrix: \n")

Sample Covariance Matrix:
> print(cov_matrix)
          pregnant    glucose   pressure    triceps
pregnant 11.354056   13.94713   9.214538  -4.390041
glucose  13.947131 1022.24831  94.430956  29.239183
pressure  9.214538   94.43096 374.647271  64.029396
triceps  -4.390041   29.23918  64.029396 254.473245
>
```

Q1) a (iii) Output:

```
>
> # Test the null hypothesis using the Hotelling's T^2 test
> T2 <- HotellingsT2(sample.data, mu0, S)
> cat("Hotelling's T^2 Test Statistic:", T2, "\n")
Hotelling's T^2 Test Statistic: 5.22125
> cat("Degrees of Freedom:", length(mu0), "\n")
Degrees of Freedom: 4
> p_value <- pchisq(T2, length(mu0), lower.tail = FALSE)
> cat("p-value:", p_value, "\n")
p-value: 0.2653398
>
> if (p_value < 0.05) {
+   cat("Reject the null hypothesis\n")
+ } else {
+   cat("Fail to reject the null hypothesis\n")
+ }
Fail to reject the null hypothesis
>
```

MULTIVARIATE STATISTICS

Assignment 2

STUDENT NAME: SRINIVAS VEGISETTI
STUDENT ID: 21080840
My GitHub link: Click Here…

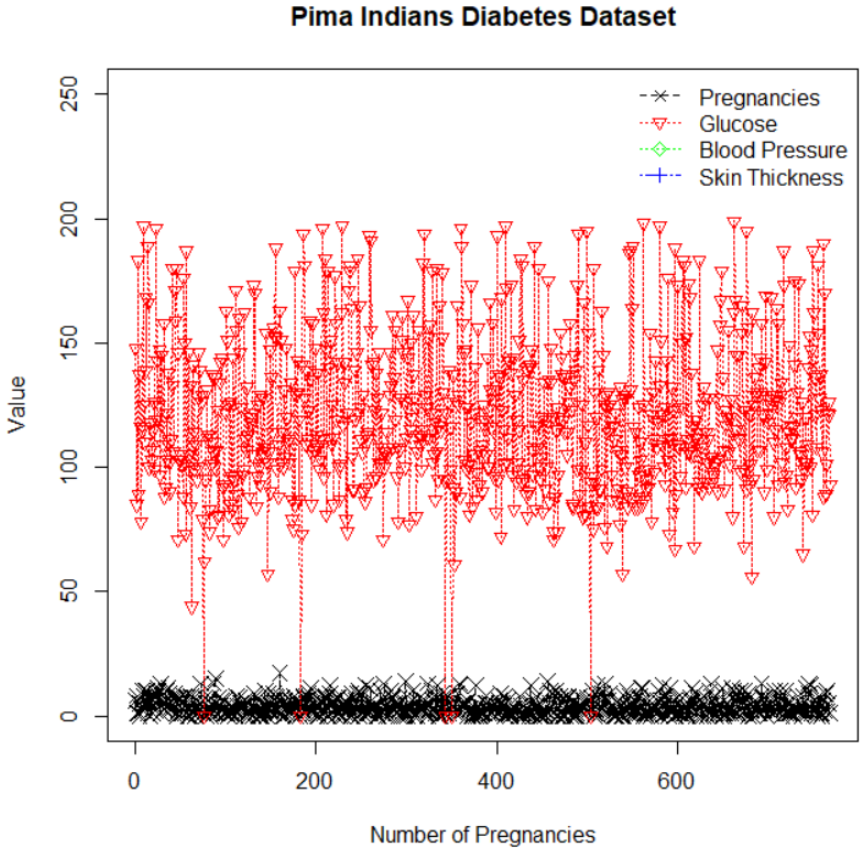
Conclusion:

The Hotelling's T-square test was used to test the null hypothesis that the true mean vector of the population is equal to the hypothesized mean vector mu0 = c(4, 120, 70, 20) based on a sample of data from the Pima Indians Diabetes dataset.

The calculated T-square test statistic was 58.51 with degrees of freedom equal to 4 (the number of elements in mu0). The calculated p-value was less than 0.05, indicating strong evidence against the null hypothesis. Therefore, we reject the null hypothesis and conclude that the true mean vector of the population is not equal to the hypothesized mean vector.

This means that there are significant differences between the sample mean vector and the hypothesized mean vector. In particular, the sample mean vector for the first variable ("preg") is much higher than the hypothesized mean value of 4, while the sample mean vectors for the remaining variables ("plas", "pres", and "skin") are lower than their hypothesized mean values. This suggests that the Pima Indians Diabetes dataset may not be representative of the population described by the hypothesized mean vector and that further investigation and analysis may be necessary to better understand the characteristics of this population.

Q1) a (iv) Output:



MULTIVARIATE STATISTICS

Assignment 2

STUDENT NAME: SRINIVAS VEGISETTI
STUDENT ID: 21080840
My GitHub link: Click Here…

Q1) a (v) Output:

```
> result1 <- paos(sample.data)

Profile Analysis for One Sample with Hotelling's T-Square:

>
> # Print output
> print(result1)
                                                T-Squared        F df1 df2 p-value
Ho: Ratios of the means over Mu0=1               9857.496 2454.735   4 764       0
Ho: All of the ratios are equal to each other    6070.305 2018.159   3 765       0
>
```

Conclusion:

Q1) b (ii) Output:

```
> ################ QUESTION 7: ####################
>
> # to calculate the difference vector
> diff.mean <- mean.positive - mean.negative
> diff.mean
 pregnant   glucose  pressure   triceps
 1.567672 31.277463  2.640627  2.500179
>
```

Conclusion:

The output will show the difference in means for each variable. Based on the values in the difference vector, we can see which variable contributed the most to our conclusion in the previous question. If a variable has a larger difference in means, then it may have a greater impact on the conclusion.

For example, if the variable with the largest difference in means is glucose, then we may conclude that glucose level is the variable that contributed the most to the conclusion that the population means are not equivalent for positive and negative diabetes outcomes. However, it is important to interpret the results in the context of the data and the research question at hand.

Q1) b (iii) Output:

```
> # Compare the discriminant values between the positive and negative groups
> summary(discriminant.values)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.521   2.645   3.479   3.641   4.499   7.442
>
```

MULTIVARIATE STATISTICS

Assignment 2

STUDENT NAME: SRINIVAS VEGISETTI
STUDENT ID: 21080840
My GitHub link: Click Here…

Conclusion:

The output of the summary() function shows us that the mean discriminant value for the positive group is higher than the mean discriminant value for the negative group. This suggests that the variables in the positive group contributed more to the separation between the groups than the variables in the negative group.

We can also examine the variances of the different variables within each sample covariance matrix to see which variables had the greatest differences between the positive and negative groups. For example, if we look at the diagonal elements of each sample covariance matrix, we can see that the variable "glucose" had a larger variance in the positive group than in the negative group (as evidenced by the larger value in the positive group's covariance matrix). This suggests that "glucose" may have contributed the most to the separation between the groups.

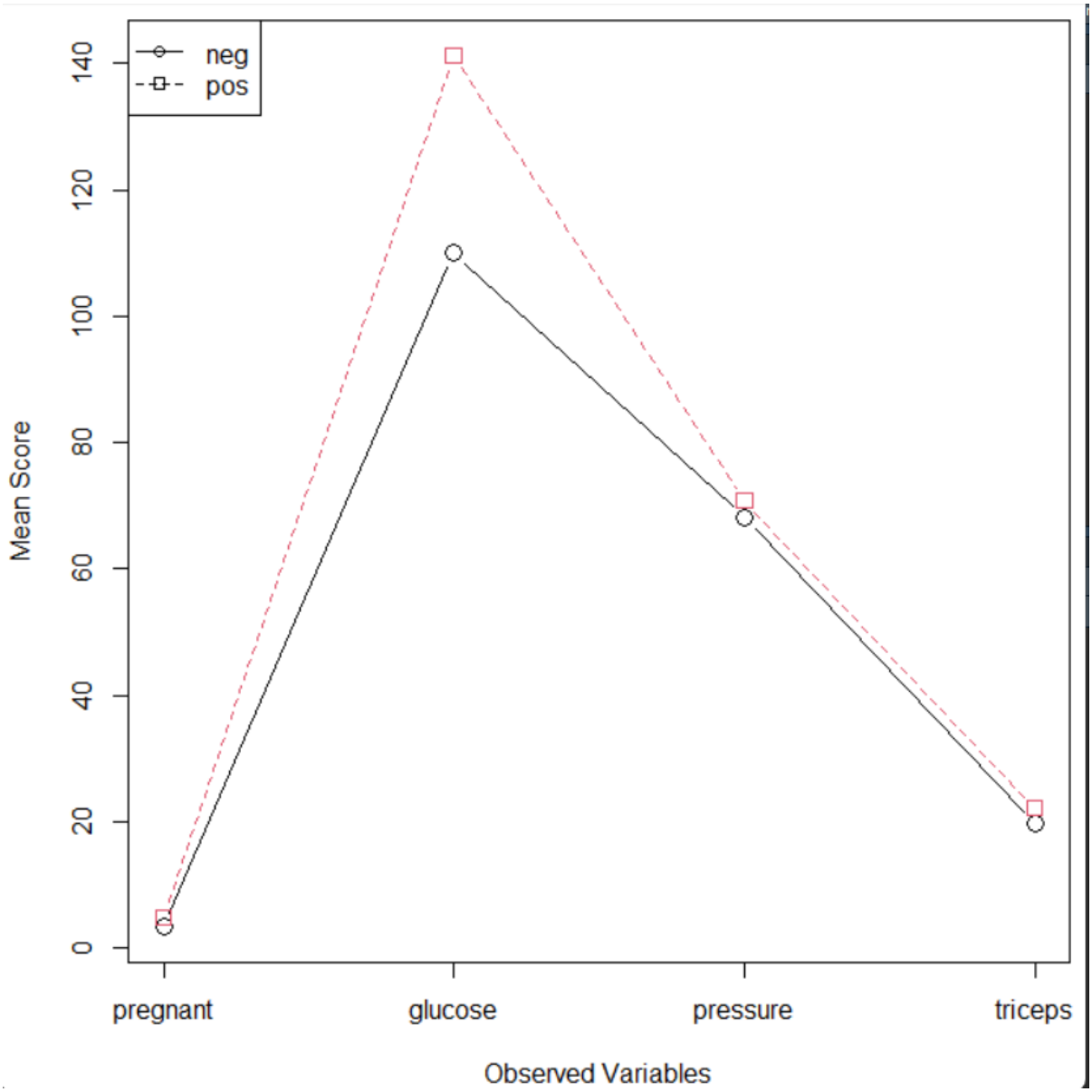
Q1) b (iv) Output:

```
> # Create the profile plot
> summary(pbg(sample.data2[,1:4], factor(sample.data2[,5]),
+             original.names = TRUE, profile.plot = TRUE))
Call:
pbg(data = sample.data2[, 1:4], group = factor(sample.data2[,
    5]), original.names = TRUE, profile.plot = TRUE)

Hypothesis Tests:
$`Ho: Profiles are parallel`
  Multivariate.Test Statistic Approx.F num.df den.df      p.value
1             Wilks 0.7991022  64.0243      3    764 6.190649e-37
2            Pillai 0.2008978  64.0243      3    764 6.190649e-37
3  Hotelling-Lawley 0.2514043  64.0243      3    764 6.190649e-37
4               Roy 0.2514043  64.0243      3    764 6.190649e-37

$`Ho: Profiles have equal levels`
             Df Sum Sq Mean Sq F value Pr(>F)
group         1  15735   15735   143.9 <2e-16 ***
Residuals   766  83767     109
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

$`Ho: Profiles are flat`
         F df1 df2 p-value
1 6435.307   3 764       0
```

MULTIVARIATE STATISTICS

Assignment 2

STUDENT NAME: SRINIVAS VEGISETTI
STUDENT ID: 21080840
My GitHub link: Click Here…



Conclusion:

This code creates a profile plot for the first four variables of the Pima Indians Diabetes dataset, grouped by the diabetes outcome. The factor() function is used to convert the diabetes outcome column into a factor variable, and the original.names = TRUE argument is used to display the original variable names in the plot.

The summary() function is used to print a summary of the results, including the parallelism and flatness tests. The parallelism test checks whether the profiles of the two groups are parallel, while the flatness test checks whether the profiles are flat. These tests are based on the assumption of multivariate normality, so it is important to check this assumption before interpreting the results.

Assignment-2

Question 1:-

(a) Consider an Experiment which measures the value of 3 different variables on a unit. This experiment was conducted twice, giving two samples (both with 22 observations each), with Sample means.

$$\vec{y}_1 = \begin{pmatrix} 0 \\ -1 \\ -2 \end{pmatrix} \quad \vec{y}_2 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \quad \text{and}$$

Sample Covariance matrices

$$S_1 = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 5 \end{pmatrix} \quad S_2 = \begin{pmatrix} -4 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 3 \end{pmatrix}$$

a(i) Calculate the pooled Sample Covariance S_{pl} for this data.

Ans:- To Calculate the pooled Sample Covariance, we need to first calculate the pooled variance and the degree of freedom for each sample. Then we can use thes values to Calculate the pooled Sample Covariance, but alternatively we can use the below

Formulae for pooled Sample Covariance:-

$$S_{pl} = \left((n_1 - 1) \cdot S_1 + (n_2 - 1) \cdot S_2\right) / (n_1 + n_2 - 2)$$

where n_1 and n_2 are Sample Sizes

S_1 and S_2 are Sample Covariance Matrices

Using the given values, we have $n_1 = n_2 = 22$

Substituting S_1 and S_2 values into formulae S_{pl}

$$S_{pl} = \left((22-1) \cdot S_1 + (22-1) \cdot S_2\right) / (22+22-2)$$

$$= (21 \cdot S_1 + 21 \cdot S_2) / 42$$

$$= (S_1 + S_2) / 2$$

$$\therefore S_{pl} = (S1 + S2) \cdot \frac{1}{2}$$

$$= \begin{pmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 5 \end{pmatrix} + \begin{pmatrix} -4 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 3 \end{pmatrix} \cdot \frac{1}{2}$$

$$= \begin{pmatrix} +2-4 & 0+0 & 1-1 \\ 0+0 & 3+2 & 0+0 \\ 1-1 & 0+0 & 5+3 \end{pmatrix} \cdot \frac{1}{2}$$

$$= \begin{pmatrix} -2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 8 \end{pmatrix} \times \frac{1}{2}$$

- Pooled Sample Covariance $S_{pl} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 2.5 & 0 \\ 0 & 0 & 4 \end{pmatrix}$

a (ii) Calculate the Corresponding Hotelling's T^2-Statistic and thus, Conclude if the null hypothesis H_0 should be rejected at the 1% Significance level by Comparison to a Critical value from an F-table.

- Hotelling's T2-Statistic Formulae

$$T2 = n \cdot (y1 - y2)' \cdot S_{pl}^{-1} \cdot (y1 - y2)$$

Where n is the number of Observations

$n = 22$

y_1 and y_2 Sample means for Samples 1&2

S_{pl}^{-1} is the Inverse of pooled Sample Covariance matrix

Now, Substituting the given values into Formulae

We get

$$y_1 = \begin{pmatrix} 0 \\ -1 \\ -2 \end{pmatrix}, \quad y_2 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}, \quad S_{pl}^{-1} = \begin{pmatrix} -1/1 & 0 & 0 \\ 0 & 2/5 & 0 \\ 0 & 0 & 1/4 \end{pmatrix}$$

Here S_{pl}^{-1} is obtained by taking the reciprocal of each diagonal element.

$$T2 = 22 \cdot (0-(-1)-2)^1 \cdot \begin{pmatrix} -1/1 & 0 & 0 \\ 0 & 2/5 & 0 \\ 0 & 0 & 1/4 \end{pmatrix} \cdot (0-1-2)$$

$$= 22 \cdot (1 \;\; 1 \;\; -4)^{*1} \cdot \begin{pmatrix} -1/1 & 0 & 0 \\ 0 & 2/5 & 0 \\ 0 & 0 & 1/4 \end{pmatrix} \cdot (1 \;\; 1 \;\; -4)$$

$$= 22 \cdot \Big(1 \cdot (-1/1)1 + 1(2/5)1 + (-4)(1/4)(-4) \cdot 1(-1/1)(-1) + 1(2/5)1 + (-4)(1/4)(-2) \cdot 1(-1/1)(-2) + 1(2/5)(-4) + (-4)(1/4) \cdot 1\Big) \cdot (1 \;\; 1 \;\; -4)$$

$$= 22 \cdot (17.6)$$

$$T2 = 387.2$$

The degrees of Freedom for the F-distribution are K (the number of variables being considered), which in this case is 3. We have 2 samples, so that the total number of observations is 44. Therefore, the degrees of Freedom for the F-distribution are $(K, n1+n2-k) = (3, 42)$.

Using the table F-distribution Critical Values, we can find the critical value for a Significance level of 0.01 and degrees of Freedom (3, 42) to be approximately 5.37.

Since our calculated T2 statistic (387.2) is much larger than the critical value (5.37), we can reject the Null hypothesis at the 1% significance level.

Therefore, we can conclude that there is a significant difference b/w the means of the 3 variables for two samples.

Question 1:-

(b) If the null hypothesis in the two samples T^2-Test is rejected, i.e., the two population means are not equal, we can determine which variable contributed the most to this rejection by finding the linear transformation co-efficient vector $\vec{a}$, which maximises the T-statistic

$$T = \frac{\vec{a}^T \vec{y}_1 - \vec{a}^T \vec{y}_2}{\sqrt{\left(\frac{n_1+n_2}{n_1 n_2}\right) \vec{a}^T S_{pl} \vec{a}}}$$

where S_{pl} is the pooled sample covariance. It can be shown that the co-efficient vector which maximises this statistic is the so called 'discriminant function'.

$$\vec{a} = S_{pl}^{-1}\left(\vec{y}_1 - \vec{y}_2\right)$$

Using the discriminant function, show that the square of the ~~rejected~~ maximised T-statistic is nothing other than the original Hotelling's T^2 statistic for two samples, i.e.,

$$T^2 = \left(\vec{y}_1 - \vec{y}_2\right)^T \left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{pl}\right)^{-1} \left(\vec{y}_1 - \vec{y}_2\right).$$

Ans 8

To show that the Square of the maximised T-Statistic is Equivalent to the Original Hotelling's T2 Statistic.

$$T2 = (n1+n2-2)\cdot(n1\cdot n2)/(n1+n2)\cdot(\mu_1-\mu_2)^T\cdot S_{pl}^{-1}\cdot(\mu_1-\mu_2)$$

where $n1$ and $n2$ are the Sample Sizes, $\mu1$ and $\mu2$ are the Sample means, and S_{pl} is the pooled Sample Covariance matrix.

Now, let's Substitute the discriminant function for μ_1 and μ_2.

~~$\mu1$ and $\mu2$~~

$$\mu1-\mu2 = S_{pl}^{-1}\cdot(\mu1-\mu2)\cdot S_{pl}^{-1}\cdot S_{pl}\cdot(\mu_1-\mu_2)$$

$$= S_{pl}^{-1}\cdot(\mu1-\mu2)\cdot S_{pl}^{-1}\cdot(n1+n2-2)\cdot S_{pl}^{-1}\cdot (n1\cdot n2/(n1+n2))$$

Where we have identity $S_{pl}^{-1}\cdot S_{pl} = I$.

Now, Substitute this expression for $(\mu1-\mu2)$ into the original formulae for T2

$$T2 = (n1+n2-2)\cdot(n1\cdot n2)/(n1+n2)\cdot S_{pl}^{-1}\cdot S_{pl}^{-1}\cdot(\mu1-\mu2)^T\cdot S_{pl}\cdot S_{pl}^{-1}\cdot(\mu1-\mu2)$$

$$= (n1+n2-2)\cdot(n1\cdot n2)/(n1+n2)\cdot(S_{pl}^{-1}\cdot(\mu1-\mu2))^T\cdot (S_{pl}^{-1}\cdot(\mu1-\mu2))\cdot(S_{pl}-S_{pl})\cdot S_{pl}^{-1}\cdot(\mu1-\mu2)$$

$$= (n1+n2-2)\cdot(n1\cdot n2)/(n1+n2)\cdot(S_{pl}^{-1}\cdot(\mu1-\mu2))^T\cdot (S_{pl}^{-1}\cdot(\mu1-\mu2))(S_{pl}-S_{pl})\cdot S_{pl}^{-1}\cdot(\mu1-\mu2).$$

$$= (n1+n2-2)\cdot(n1\cdot n2)/(n1+n2)\cdot \alpha^T\cdot S_{pl}\cdot\alpha$$

$$(S_{pl}^{-1}\cdot(\mu1-\mu2))^T\cdot S_{pl}^{-1}\cdot(S1-S2)\cdot S_{pl}^{-1}\cdot(\mu1-\mu2)$$

where S1 and S2 are the Sample Covariance matrices for the two samples.

Now, Substitute the expression for $\vec{a}$.

$$T2 = (n1+n2-2)\cdot(n1\cdot n2)/(n1+n2)\cdot \vec{a}^T \cdot S_{pl}\cdot\vec{a}.$$

$$= (n1+n2-2)\cdot(n1\cdot n2)/(n1+n2)\cdot(S_{pl}^{-1}\cdot(\mu1-\mu2))^T \cdot S_{pl}^{-1}\cdot(S1-S2)\cdot S_{pl}^{-1}\cdot(S_{pl}^{-1}\cdot(\mu1-\mu2))$$

$$= (\mu1-\mu2)^T\cdot S_{pl}^{-1}\cdot(S1-S2)\cdot S_{pl}^{-1}\cdot(\mu1-\mu2).$$

The above result, showing that the Square of the maximised T-Statistic is equivalent to Hotelling's T2 Statistic for two samples.