Assignment 2

Emre Beray Boztepe

30 04 2024

Problem 1

Option 1

A chi-squared random variable X with p degrees of freedom can be defined as:

$$X_p^2 = \sum_{i=1}^{p} Z_i^2$$

A random variable Y from the F distribution with d_1 and d_2 degrees of freedom can be defined as:

$$Y = \frac{\frac{1}{d_1}\sum_{i=1}^{d_1} Z_i^2}{\frac{1}{d_2}\sum_{i=d_1+1}^{d_1+d_2} Z_i^2}$$

Option 2

$F_{p,n-p}$ will approximately follow an F-distribution, as stated, with 4 and 996 degrees of freedom. However, with n large, it also approximately follows a normal distribution due to the central limit theorem, especially since $n-p$ is much larger than p.

Option 3

$$n(\bar{X}-\mu)^T \quad \Sigma^{-1}(\bar{X}-\mu)$$

$$= n(\bar{X}-\mu)^T \left(\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}\right)^{-1} (\bar{X}-\mu)$$

let $(\bar{X}-\mu) \rightarrow Z$

$$= nZ^T \quad \left(\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}\right)^{-1} Z$$

$$= \sqrt{n}\sqrt{n}Z^T \left(\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}\right)^{-1} Z$$

$$= \left(\sqrt{n}\left(\Sigma^{\frac{1}{2}}\right)^{-1} Z\right)^T \quad \left(\sqrt{n}\left(\Sigma^{\frac{1}{2}}\right)^{-1} Z\right) \sim X_p^2$$

As, $\sqrt{n}\left(\Sigma^{\frac{1}{2}}\right)^{-1} Z \sim N_p(0,1)$

Option 4

(a)

Under H_0: $\mu = \mu_0$, T^2 follows an F-distribution with p and $n-p$ degrees of freedom:

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

The test rejects H_0 if $T^2 > F_{p,n-p,\alpha}$, the critical value at significance level α

(b)

As n increases, T^2 tends to infinity under H_1 (non-central F-distribution), reducing the probability of Type II error, thus increasing power.

(c)

If H_0 is false, T^2 tends to infinity as n increases, ensuring the probability of correctly rejecting H_0 approaches 1 (asymptotic consistency).

Problem 2

In order to calculate p-values: For each component Z_i we can calculate it as the probability of observing a value as extreme as Z_i under the standard normal distribution. This means, we can use cumulative standard distribution (CDF).

$p - value = 2 \times \left(1 - \Phi(|Z_i|)\right)$ where $\Phi =$ CDF of standard normal distribution

Option 1

The Bonferroni test statistics has the formula as follows:

$$T_{bonf} = min\{p_i\}$$

We reject the null hypothesis for small values of the test statistics:

$$\varphi_{bonf} = T_{bonf} < \frac{\alpha}{n}$$

- n: the size
- α: This is chosen as 0.05 (which is common value for α)

```
## p-values:  0.08913093 0.1095986 0.0009668483 0.006933948 0.9680931
0.7263387 0.6170751 0.3173105 0.4839273 0.4237108

##  [1] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

So, when we calculate p values from given X vector, we would only reject one hypothesis (number 3) when we apply Bonferroni procedure.

Option 2

Benjamini-Hochberg test statistics has the formula of:

Reject $H_{0,(i)}$ if:

$$\exists_{(j\geq i)} \quad p_{(j)} < \frac{j}{n}\alpha$$

- n: the size
- α: 0.05
- j: index variable that corresponds to the rank of each p-value (ordered p-values)

```
## indices of p_values after ordering:  3 4 1 2 8 10 9 7 6 5
## rejected p_values:  TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE

##  [1] FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
```

When applying the Benjamini-Hochberg procedure, we observe that we reject hypotheses until the third one after ordering, as its p-value is lower than its corresponding critical value. However, we do not reject any hypotheses beyond the third one.

Option 3

For Bonferroni: Only the third hypothesis is rejected and we assume that only the first three coordinates of μ are different from zero (from the question). So, the hypothesis rejection is correct (true positive) and there are no false positives.

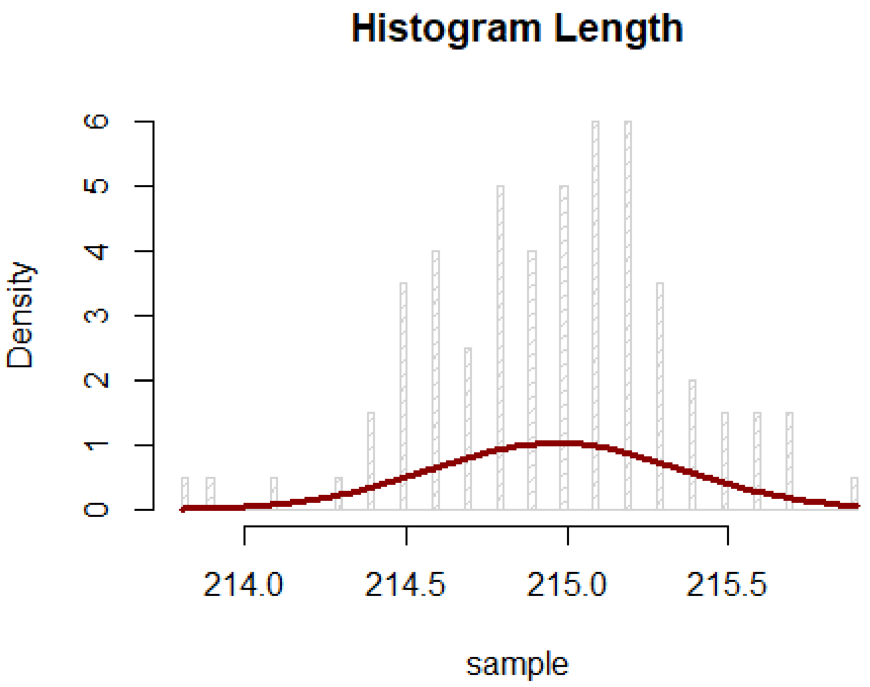
Bonferroni: FDP: $\frac{FP}{FP+TP} = \frac{0}{1+0} = 0$

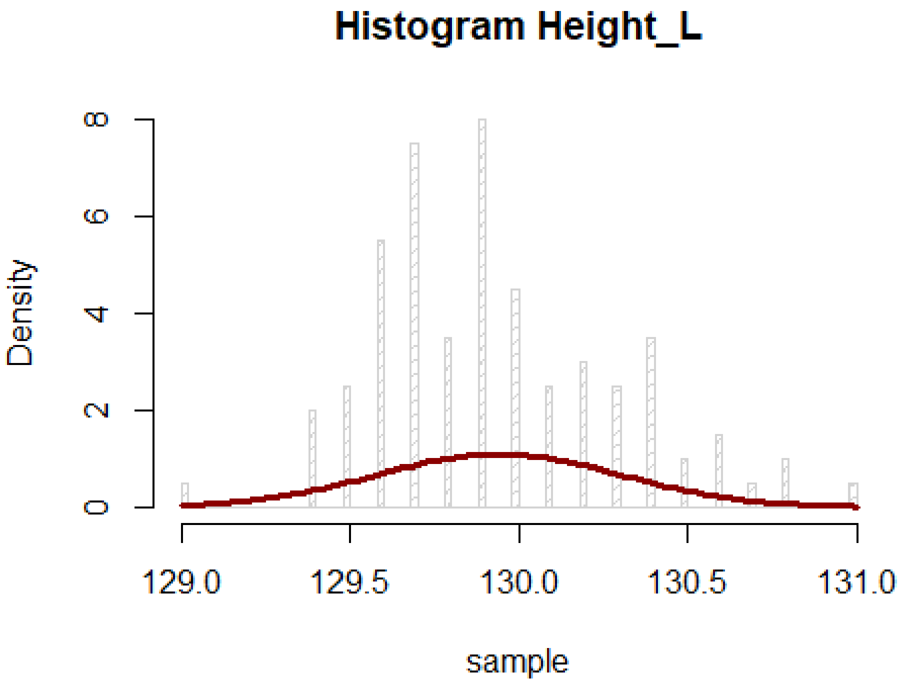
For BH: Third and fourth hypotheses are rejected. So, we have 1 true positive because 3rd hypothesis is rejected and we have 1 false positive as we reject fourth hypothesis.

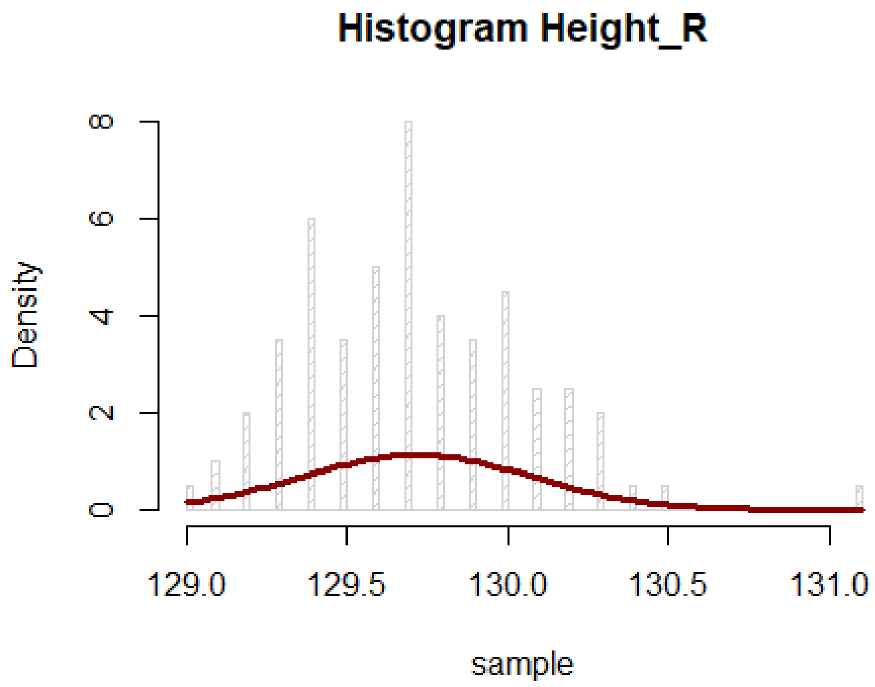
BH: FDP: $\frac{FP}{FP+TP} = \frac{1}{1+1} = \frac{1}{2}$ So, FDP for BH would be %50.

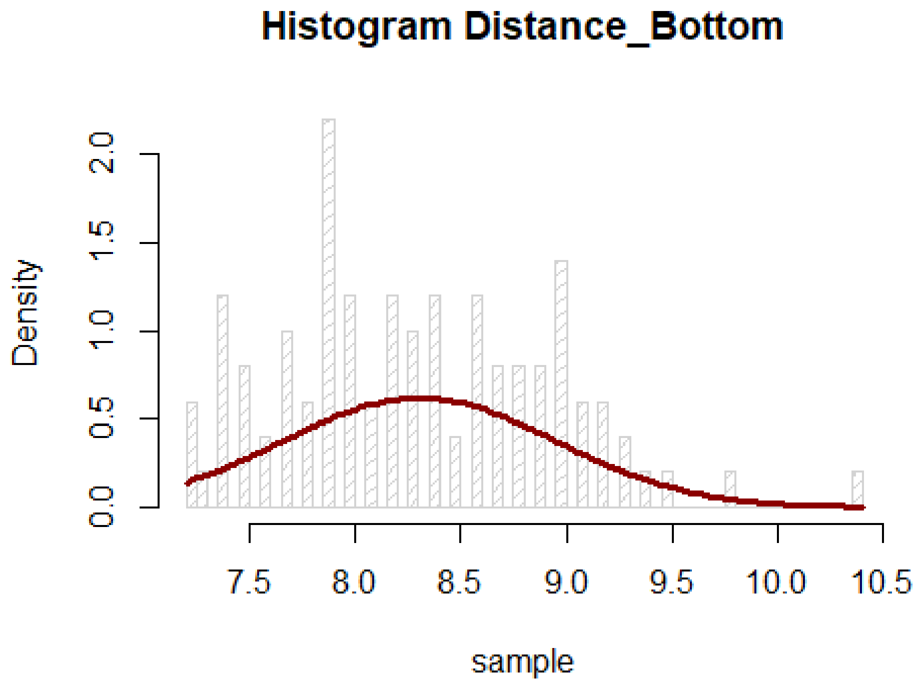
Project 2

1-) Load the data, produce scatter plots and qq-plots of the data and discuss validity of the assumption that the data are from a multivariate normal distribution.

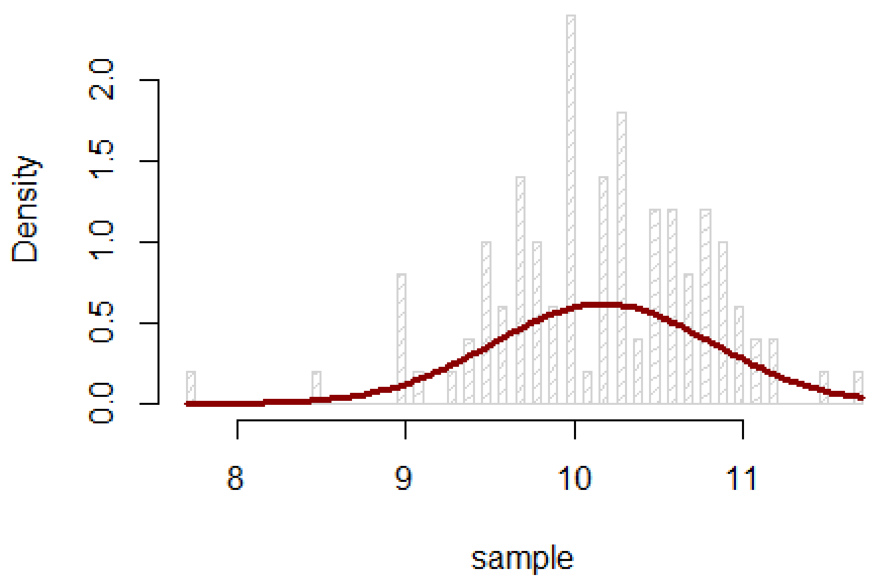

Histogram Length


Histogram Height_L

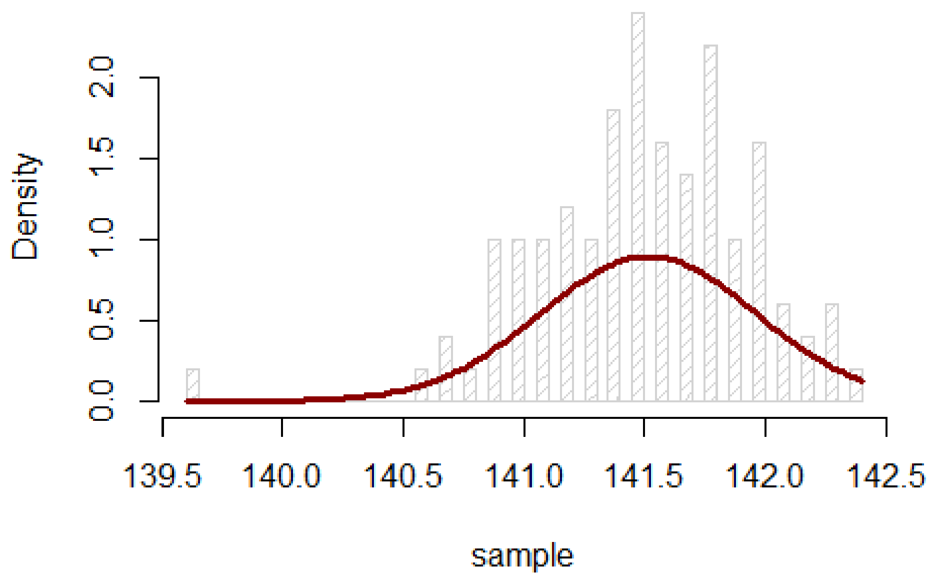

Histogram Height_R

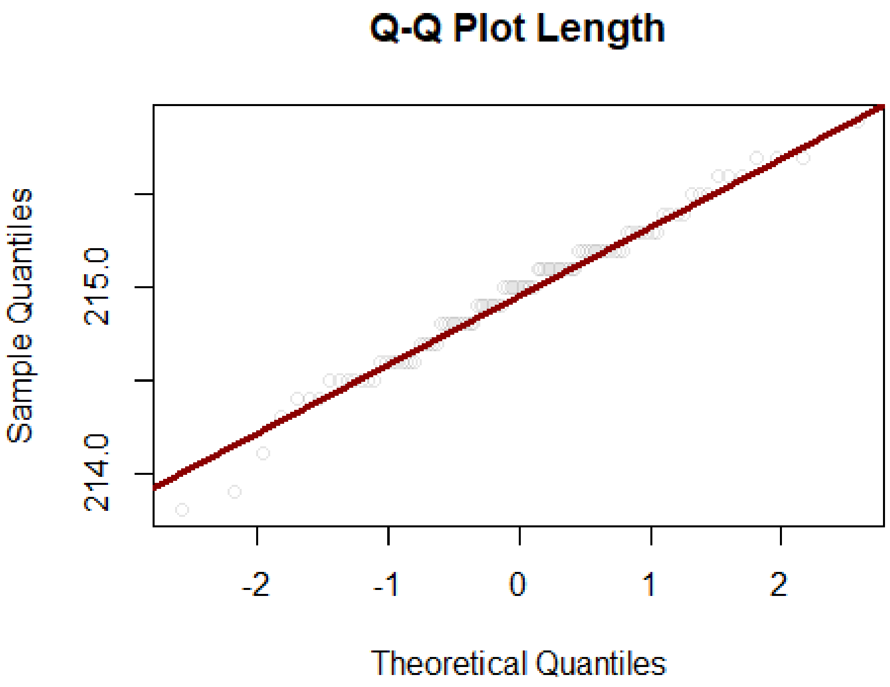

Histogram Distance_Bottom

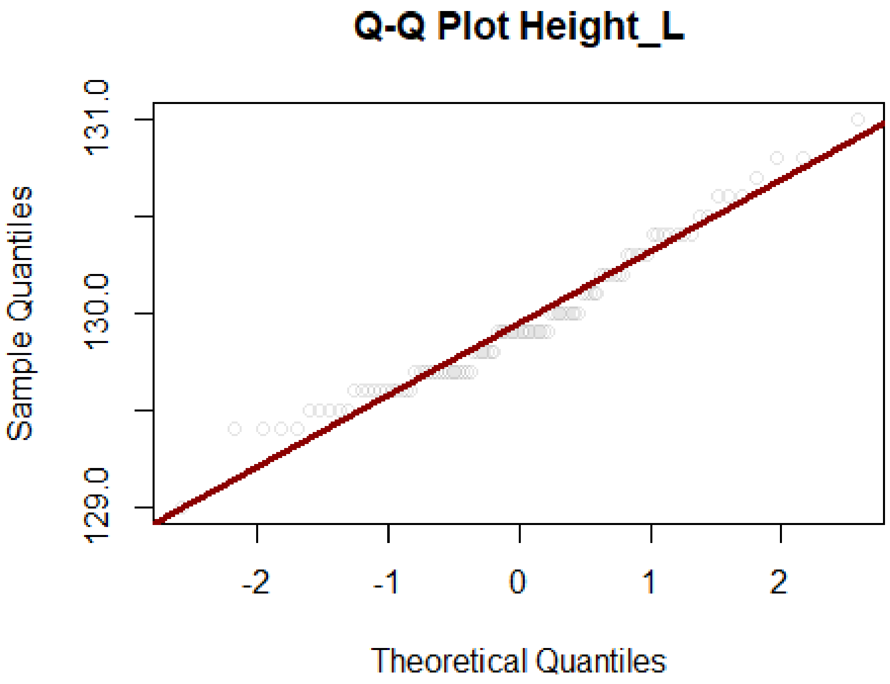
Histogram Distance_Top



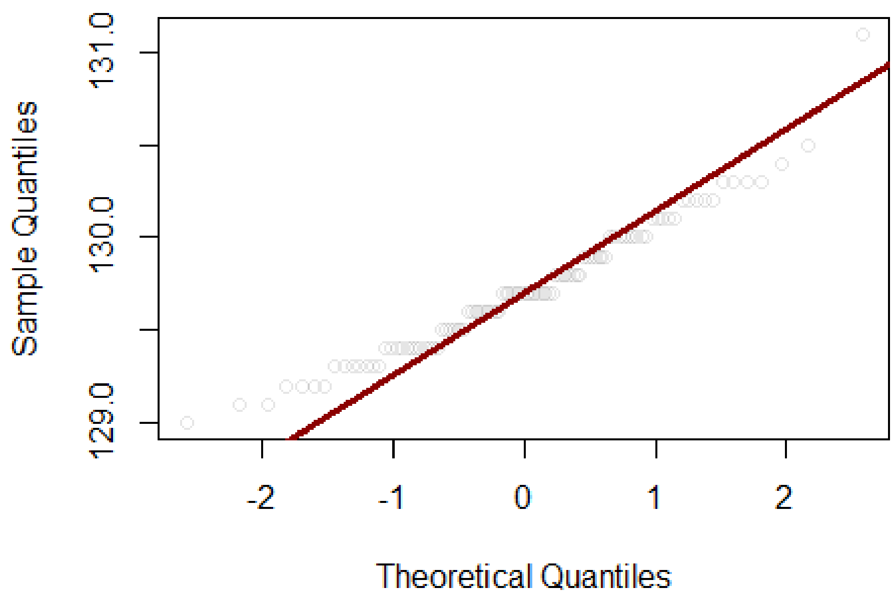
Histogram Length_Diagonal



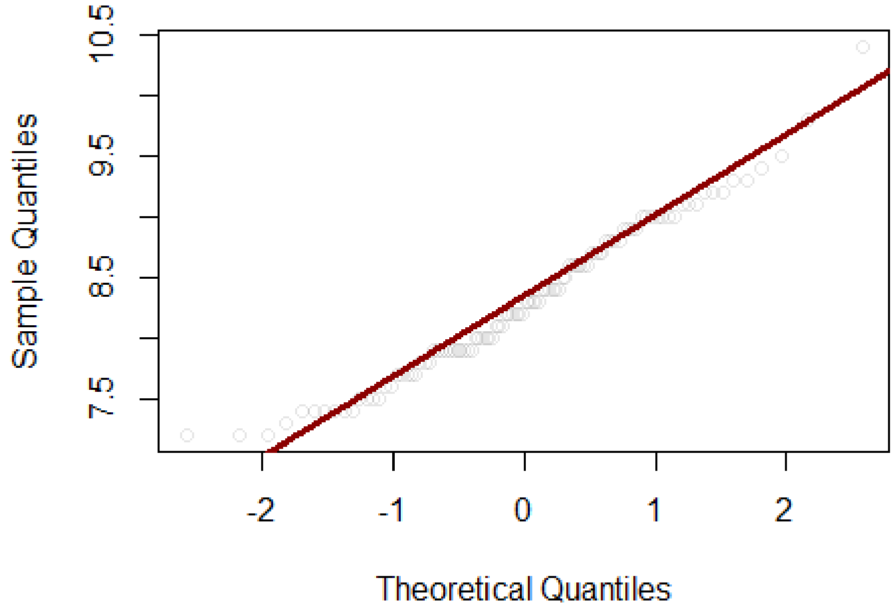

Q-Q Plot Length
Sample Quantiles
214.0
215.0
-2
-1
0
1
2
Theoretical Quantiles

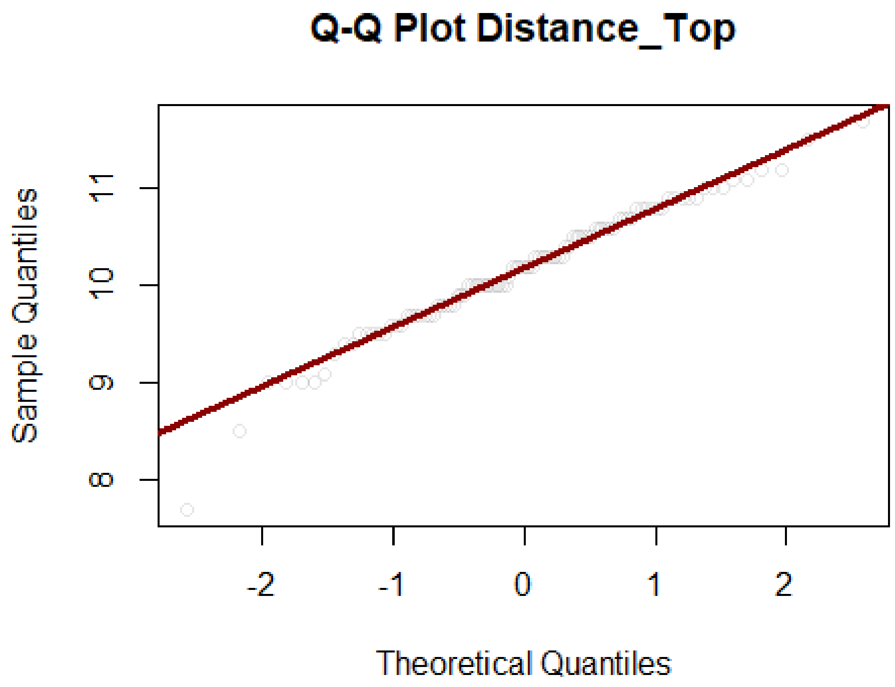

Q-Q Plot Height_L
Sample Quantiles
129.0
130.0
131.0
-2
-1
0
1
2
Theoretical Quantiles

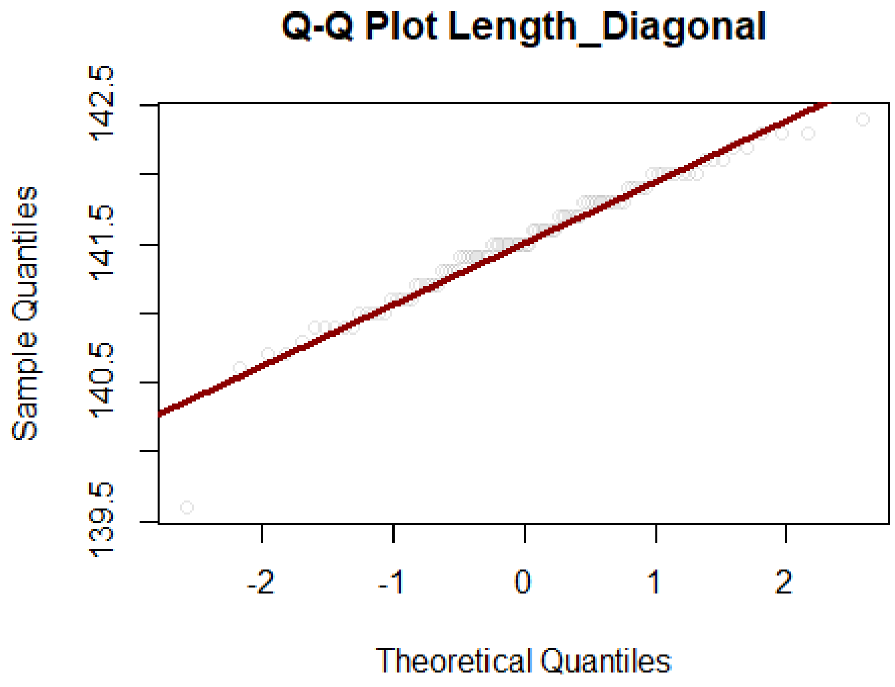
Q-Q Plot Height_R



Q-Q Plot Distance_Bottom







The histograms provided show some asymmetry and variability from the bell curve, especially for the "Height_L" and "Distance_Bottom" variables, which could suggest deviations from normality.

If the data are normally distributed, the points should lie approximately along the 45-degree reference line. In the plots we have, the points largely follow the line, but there are some deviations, especially in the tails.

2-) Evaluate estimators of the vector of means and the covariance matrix.

```
## Mean: 214.969 129.943 129.72 8.305 10.168 141.517
## Covariance Matrix:

##                      Length    Height_L    Height_R Distance_Bottom
## Length          0.150241414  0.05801313  0.05729293    0.0571262626
## Height_L        0.058013131  0.13257677  0.08589899    0.0566515152
## Height_R        0.057292929  0.08589899  0.12626263    0.0581818182
## Distance_Bottom 0.057126263  0.05665152  0.05818182    0.4132070707
## Distance_Top    0.014452525  0.04906667  0.03064646   -0.2634747475
## Length_Diagonal 0.005481818 -0.04306162 -0.02377778   -0.0001868687
##                 Distance_Top Length_Diagonal
## Length            0.01445253    0.0054818182
## Height_L          0.04906667   -0.0430616162
## Height_R          0.03064646   -0.0237777778
## Distance_Bottom  -0.26347475   -0.0001868687
## Distance_Top      0.42118788   -0.0753090909
## Length_Diagonal  -0.07530909    0.1998090909
```

3-) Write an R function that is verifying if a point lies inside of the six dimensional ellipsoid that serve as the 95% confidence region for the mean value of bank notes based on the Hotelling's T^2 statistics.

Hotteling Statistic: $n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu) \leq \frac{(n-1)p}{n-p}F_{p,n-p}(a)$

```
check_hotelling = function(sample_data, mean, cov_matrix, new_point) {
  n = nrow(sample_data)
  p = ncol(sample_data)

  T2 = n * (new_point - mean) %*% solve(cov_matrix) %*% (new_point - mean)

  F_critical = qf(0.95, p, n - p)
  T2_critical = (p * (n - 1) / (n - p)) * F_critical

  return(T2 <= T2_critical)
}
```

4-) A new production line that will be replacing the old one for printing the bank notes is tested and one of the requirements is that the average dimensions of the bank notes are comparable to these represented in the provided sample of the original bank notes. After printing a very long series of bank notes in the new production line, it was found that the mean values of the dimensions are:

m0 LENGTH LEFT RIGHT BOTTOM TOP DIAGONAL

[1,] 214.97 130 129.67 8.3 10.16 141.52

(Since the number of bank notes printed out for this purpose was very large so the error of for the obtained mean values is negligible). Check if the obtained mean values are within the Hotelling's confidence region that was obtained based on the original sample of bank notes.

```
##       [,1]
## [1,] FALSE
```

Hotelling's T^2 is a multivariate test that checks if there is a significant difference between the mean vector of a sample and a hypothesized mean vector. FALSE indicates that suggests that the new data point significantly deviates from the expected mean values of the dataset.

5-) Check if the new mean vector falls within the Bonferroni's confidence rectangular region for the mean value of the old bank note dimensions.

Bonferroni Interval: $\overline{X}(+-)t_{n-1}\left(\frac{\alpha_i}{2}\right)\sqrt{\frac{s_i^2}{n}}$

where $\sum_{i=1}^{p}\alpha_i = \alpha$

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

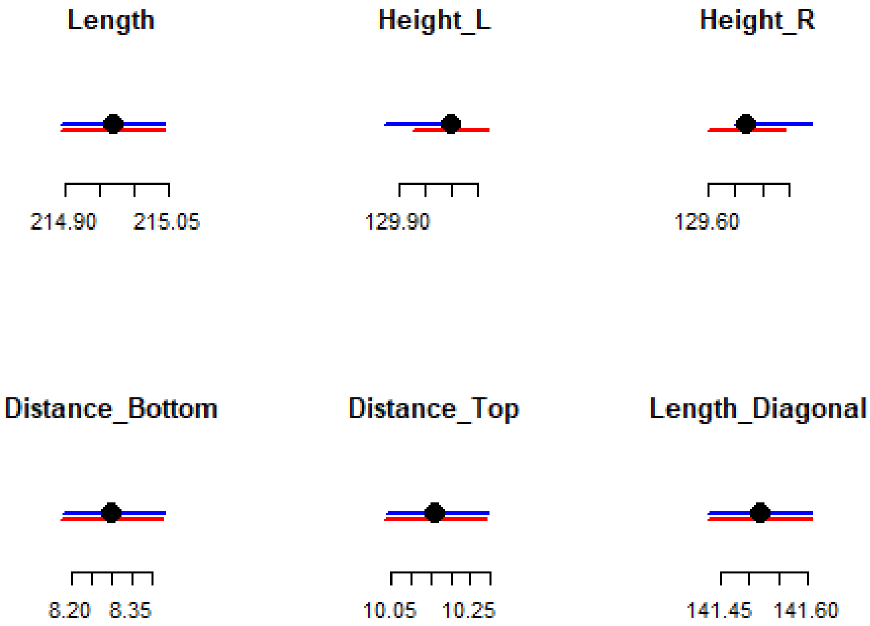
TRUE means our new point falls within its respective Bonferroni confidence interval.

```
## [1] "Bonferroni Region: "

##          Length        Height_L        Height_R Distance_Bottom 
Distance_Top 
##      215.073356      130.041030      129.815667        8.478065 
10.342728 
## Length_Diagonal 
##      141.637346 
##          Length        Height_L        Height_R Distance_Bottom 
Distance_Top 
##      214.864644      129.844970      129.624333        8.131935 
9.993272 
## Length_Diagonal 
##      141.396654
```

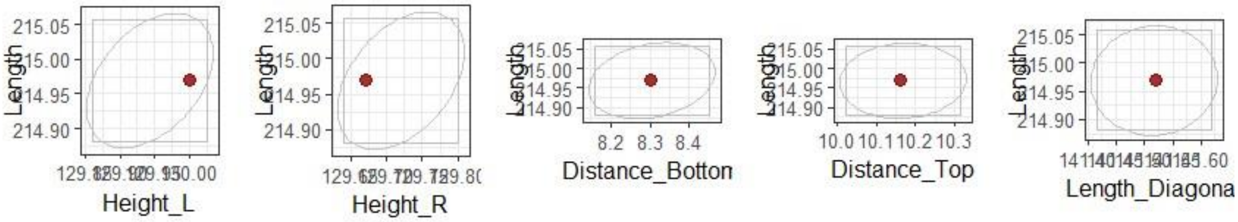
6-) Plot the projection of both confidence regions to the one-dimensional spaces marked by the axes: $X_i, i = 1, \ldots 6$. Mark the projection of the vector of means on the obtained confidence intervals. Comment what you observed.

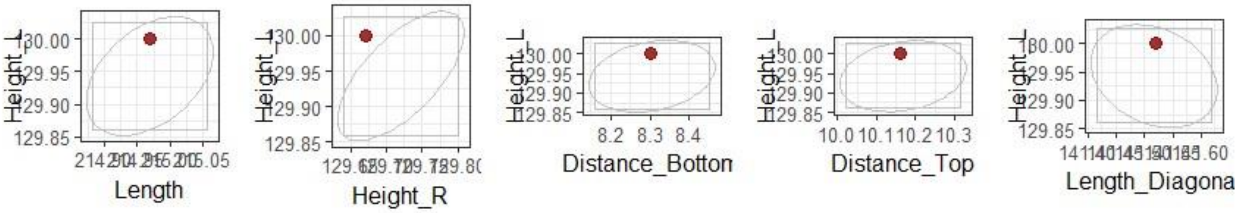
For this plot, blue line represents Bonferroni Confidence Interval, red line represents Hotteling T^2 interval and black circle is means for each variable.

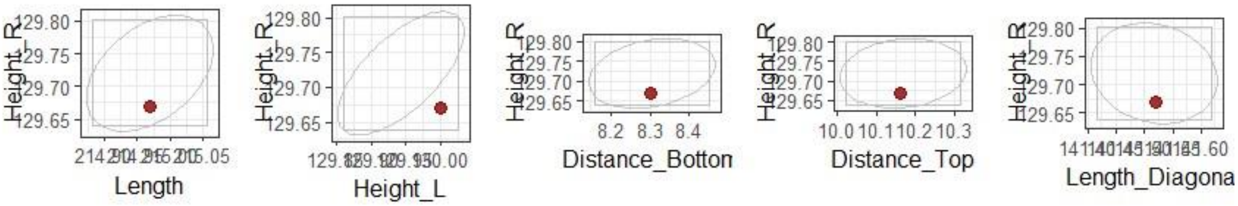


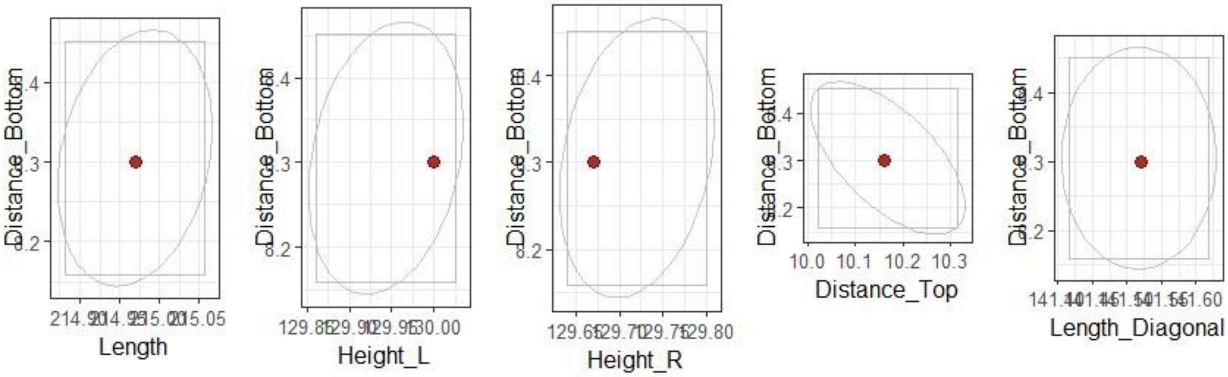
For the plot, we can say that there is a consistency in the closeness of the measured means to the center of the confidence intervals in each variable. The close proximity of the measured means to the center of both confidence intervals across all variables suggests that there are no significant deviations from expected values. Also, the confidence intervals are quite narrow across all variables, which suggests high precision in the measurements.

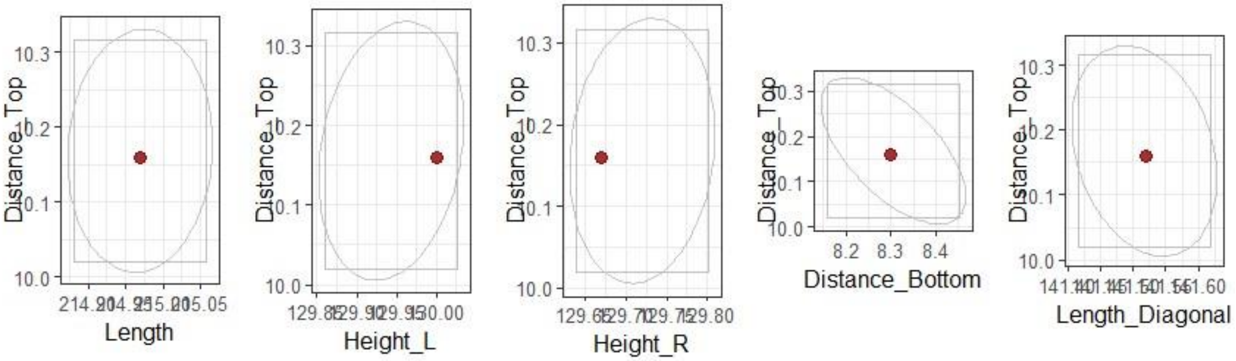
7-) Plot the projection of both confidence regions to the two-dimensional spaces marked by the pairs of axes: $X_i, X_j, i \neq j$. Mark the projection of the vector of means. Comment what you observed. Hint: Use package ellipse.

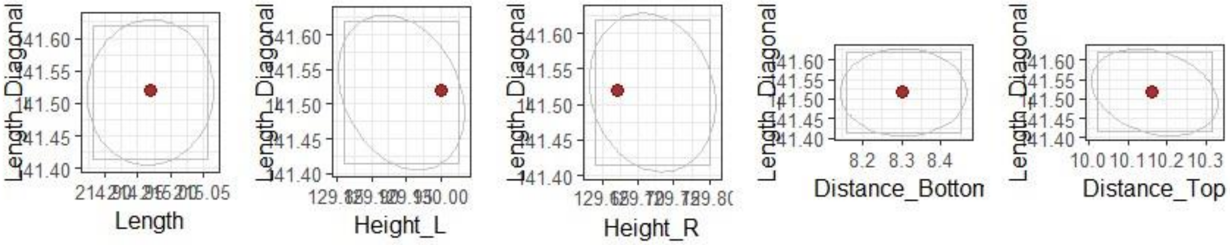

Length
215.05
215.00
214.95
214.90
129.85129.90129.95130.00
Height_L
Length
215.05
215.00
214.95
214.90
129.65129.70129.75129.80
Height_R
Length
215.05
215.00
214.95
214.90
8.2
8.3
8.4
Distance_Botton
Length
215.05
215.00
214.95
214.90
10.0
10.1
10.2
10.3
Distance_Top
Length
215.05
215.00
214.95
214.90
141.40141.45141.50141.55141.60
Length_Diagona


Height_L
130.00
129.95
129.90
129.85
214.90214.95215.00215.05
Length
Height_L
130.00
129.95
129.90
129.85
129.65129.70129.75129.80
Height_R
Height_L
130.00
129.95
129.90
129.85
8.2
8.3
8.4
Distance_Botton
Height_L
130.00
129.95
129.90
129.85
10.0
10.1
10.2
10.3
Distance_Top
Height_L
130.00
129.95
129.90
129.85
141.40141.45141.50141.55141.60
Length_Diagona


Height_R
129.80
129.75
129.70
129.65
Length
Height_L
129.85
129.90
129.95
130.00
Distance_Bottom
8.2
8.3
8.4
Distance_Top
10.0
10.1
10.2
10.3
Length_Diagonal
141.60


Distance_Bottom
8.4
8.3
8.2
Length
214.90
214.95
215.00
215.05
Height_L
129.85
129.90
129.95
130.00
Height_R
129.65
129.70
129.75
129.80
Distance_Top
10.0
10.1
10.2
10.3
Length_Diagonal
141.40
141.45
141.50
141.55
141.60





Looks like new means (red circle) lie inside both Bonferroni Confidence Region (rectangle area) and Hotelling's T^2 region (ellipse area) for almost all pairs of axes. Only, in pair Height L and Height R, we can see that the new means lie inside Bonferroni, but it is outside of Hotelling's T^2 region. Also, the same thing can be said for vice versa.

8-) Interpret geometrically the fact that the mean values of the bank note dimensions from the new production line fail to belong to the Hotelling’s confidence region. Relate to the previously created graphs.

Let's interpret some pairs of first row (pairs of Length with other attributes):

Height_L vs. Length: If the red circle is mostly centered but slightly to the upper or lower of the center within the ellipse, it indicates slight deviations in height or length, but still within the acceptable range.

Height_R vs. Length: Similar analysis as with Height_L vs. Length. The proximity of the red circle to the center suggests minimal deviation.

Distance_Botton vs. Length: If the red circle is central or slightly off-center, it indicates that the distance from the bottom measurement aligns well with the length, adhering closely to expected values.

And for Height_L and Height_R, which the pair where the new mean lies outside of Hotteling's T^2 region, geometrically, the red dot being outside this oval suggests that the mean dimensions of the banknotes from the new production line are statistically significantly different from the expected mean dimensions, assuming the confidence region was constructed based on dimensions from a standard or previous production line. This deviation can be interpreted as evidence that the new production process might not be conforming to the established specifications or that it has introduced a significant change in the dimensionality of the banknotes.9-) It has been decided that the settings of the production line needs to be tuned better to match original dimensions of banknotes. After such tuning, another test has been carried out and the resulting means were

m1 LENGTH LEFT RIGHT BOTTOM TOP DIAGONAL

[1,] 214.99 129.95 129.73 8.51 9.96 141.55

Check if the vector of means are within: a) Hotelling’s confidence region; b) Bonferroni’s confidence region. Comment your findings.

```
## Hotteling m1:  TRUE

## [1] FALSE
```

It shows differing results between the Hotelling’s T^2 test and the Bonferroni confidence regions method for our new mean vector. While the Hotelling’s T^2 test indicates no significant multivariate difference between the sample means and the population means, the Bonferroni method suggests that at least one of the individual mean estimates falls outside the adjusted confidence intervals.

10-) After yet another tuning, the vector of means was

m2 LENGTH LEFT RIGHT BOTTOM TOP DIAGONAL

[1,] 214.9473 129.9243 129.6709 8.3254 10.0389 141.4954

Is this value acceptable based on the original sample of the bank notes, or the production line still needs some tuning? Explain your answer

```
## Hotteling m2:  TRUE

## [1] TRUE
```

For Hotteling, TRUE suggests that, there is no significant multivariate difference between the two mean vectors.

Bonferroni confidence regions test: TRUE result indicates that all individual means in our mean are within their respective confidence intervals, after adjustment for multiple comparisons.

The consistency of last mean provided with the population means, as verified by both tests, implies that the production line is currently operating correctly and producing bank notes whose characteristics align well with those of the original sample. There appears to be no need for further tuning of the production line based on this statistical evidence.

Simulation 1

```
## Warning: package 'knitr' was built under R version 4.3.2
```

Simulation mu-10

	Bonferroni's	Benjamini-Hochberg's
FWER	0.048	0.307
FDR	0.012	0.055
power	0.390	0.555

Simulation mu-500

	Bonferroni's	Benjamini-Hochberg's
FWER	0.042	1.000
FDR	0.000	0.045
power	0.386	0.903

The Bonferroni method is consistently conservative, with low FWER and FDR across both scenarios. This conservatism is beneficial for controlling Type I errors but it has really low power for the test. The BH procedure controls the FDR effectively while having a much higher power, especially when the number of true alternatives is large. However, its FWER is not as well controlled as Bonferroni's. To make a choice, whether avoiding Type I errors is paramount (favoring Bonferroni) or maximizing the power to detect true effects is more important (favoring BH)