Statistical Learning Report 2

Hu Wenjie 343312

Problem 1

1. Let $Z_1, Z_2, \ldots$ be independent copies of a $N(0,1)$ variable. Use these to define a chi-squared random variable with p degrees of freedom χ^2_p. Similarly, recall the definition of an F distribution with d_1 and d_2 degrees of freedom F_{d_1,d_2}

A chi-squared random variable X_p with p degrees of freedom is defined as:

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

An F-distribution random variable Y with d_1 and d_2 degrees of freedom is defined as:

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

2. Consider a random variable distributed according to $F_{p,n-p}$. What distribution will $F_{p,n-p}$ approximately follow for p = 4 and n = 1000 ?

$p = 4, n = 1000, n - p = 996$ The F-distribution F_4996 can be approximately described by a normal distribution because of the high degree of symmetry and low variance due to the large denominator degrees of freedom.
Mean: Approximately to 1 since $\frac{996}{996-2} \approx 1$
Variance $\approx \frac{2 \cdot 996^2 \cdot (4+996-2)}{4 \cdot 994^2 \cdot 992} \approx 0.5$
Thus, F_4996 can be approximated by a normal distribution $N(1, 0.5)$

3. Let $X_1, /\ldots, X_n \sim N_p(\mu, \Sigma)$. Show that $n(\hat{X} - \mu)^T \Sigma^{-1} (\hat{X} - \mu)$ follows a χ^2_p distribution. (Hint: $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$)

Problem 2: Multiple Testing

$X = (1.7, 1.6, 3.3, 2.7, -0.04, 0.35, -0.5, 1.0, 0.7, 0.8)$ $N(\mu, I)$

1. Which hypotheses would be rejected by the Bonferroni multiple testing procedure ?

The null hypothesis for each component i is

$$H_0 : \mu_i = 0 \text{ vs } H_1 : \mu_i \neq 0$$

Then we calculate the test statistic, since each X_i is normally distributed as $N(\mu_i, 1)$, the test statistic for each component under the null hypothesis is $Z_i = X_i$ as σ = 1. We set the sgnificance level $\alpha = 0.05$(in general),with the Bonferroni correction, the adjusted significance level for each test becomes $\alpha' = \frac{\alpha}{n} = 0.005$, then for a two-tailed test $\alpha' = 0.005$, the critical values are ± 2.807, finally we compare each $Z_i = X_i$ to the critical value 2.807 with absolute value greater than 2.807 leads to a rejection of the null hypothesis.

$$|X| = (1.7, 1.6, 3.3, 2.7, 0.04, 0.35, 0.5, 1.0, 0.7, 0.8)$$

As we see, only X_3 is greater than the critical value of 2.807. Thus, using the Bonferroni correction procedure, the null hypothesis for the third component $(\mu_3 - 0)$ would be rejected, and the rest of hypotheses would not be

1. Which hypotheses would be rejected by the BH multiple testing procedure ?

$$Z_i = X_i, \alpha = 0.05$$

First calculate the p-values for each hypothesis and then rank it such that $|X|_{(1)} \geq |X|_{(2)} \geq \ldots \geq |X|_{(p)}$, the result shows below.

X_i	p-Value	critical value
X_3	0.0009	0.0050
X_4	0.0069	0.0100
X_1	0.0891	0.0150
X_2	0.1095	0.0200
X_8	0.3173	0.0250
X_{10}	0.4237	0.0300
X_9	0.4839	0.0349
X_7	0.6170	0.0400
X_6	0.7263	0.0450
X_5	0.9680	0.0500

Then we find the largest index i the $P_i \leq$ critical values, in this case the $i = 2$, this means we reject the null hyotheses for the two corresponding components of of X which are $X_3 = 3.3$ and $X_4 = 2.7$ (X_3 and X_4 has smallest p-values when ranked).

Therefore, using the BH procedure, the null hypotheses for the X_3 and X_4 of X would be rejected, indicating significant evidence against the null hypothesis for these components.

In [ ]:

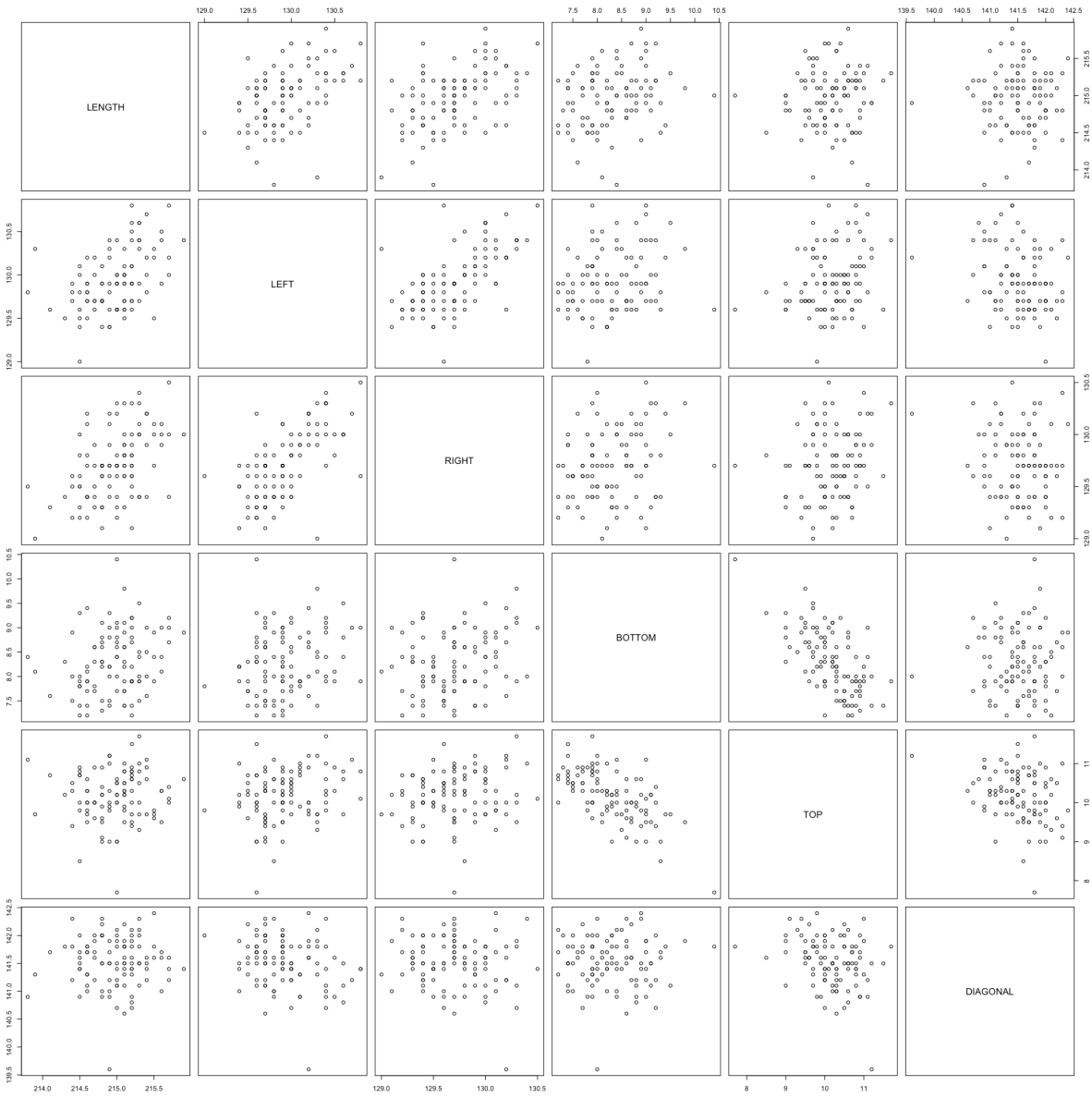
```
[1] 0.0009668483 0.0069339476 0.0891309255 0.1095985834 0.3173105079
[6] 0.4237107972 0.4839273044 0.6170750775 0.7263386976 0.9680931263
[1] 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045 0.050
[1] 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045 0.050
[1] 3 4
```

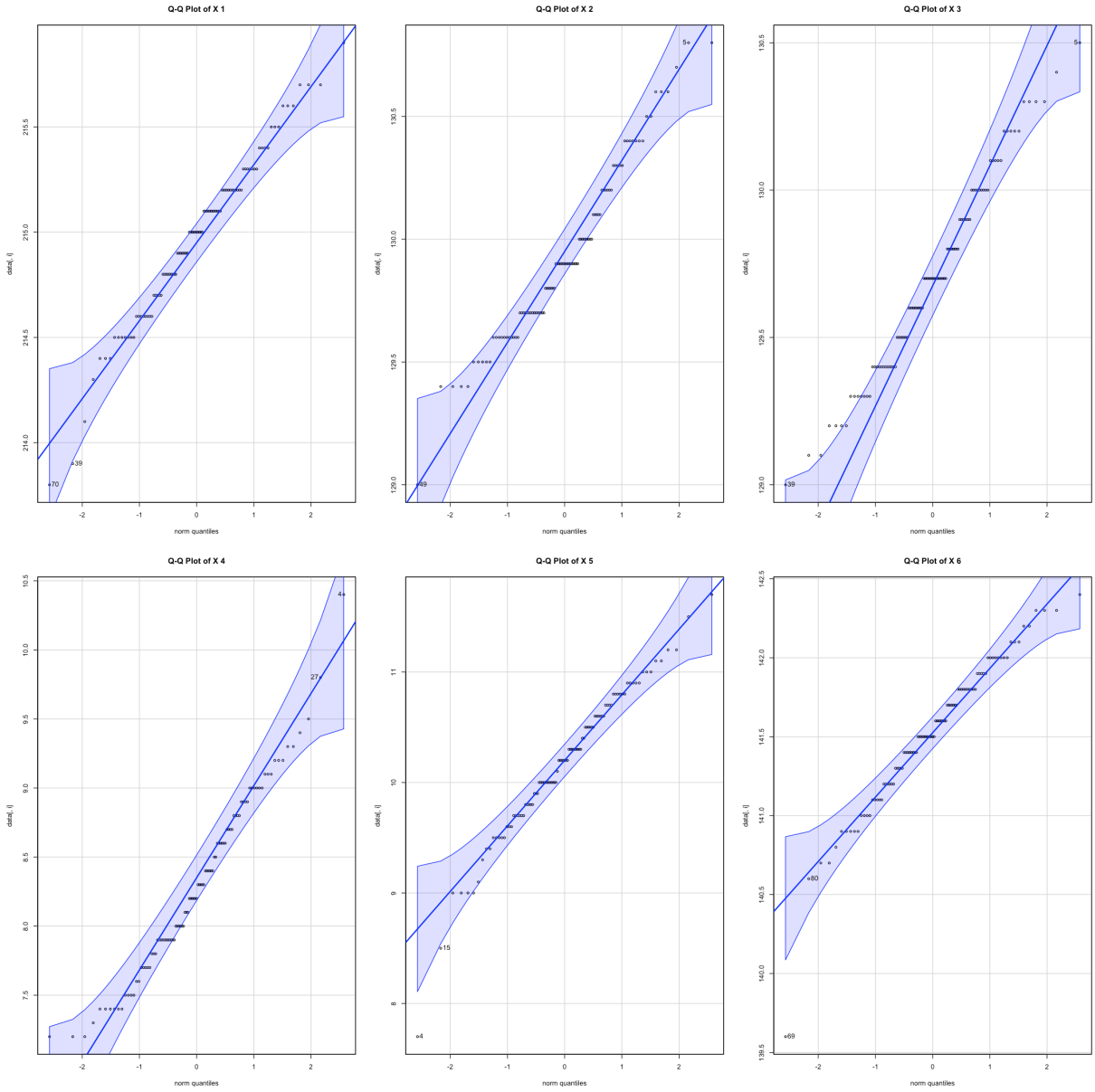
3. Assume H_0 is false. Explain what happens to T^2 as n goes to infinity. What happens to the probability of rejecting H_0?

Project

P2.1 Load the data, produce scatter plots and qq-plots of the data.

In [ ]:


LENGTH
LEFT
RIGHT
BOTTOM
TOP
DIAGONAL



P2.2 Calculate estimators of the vector of means and the covariance matrix

In [ ]:

```
   LENGTH      LEFT     RIGHT    BOTTOM       TOP  DIAGONAL 
214.97071 129.93232 129.70606   8.29798  10.17273 141.52222 
```

	LENGTH	LEFT	RIGHT	BOTTOM	TOP
LENGTH	0.151480107	0.06044630	0.06028139	0.058919810	0.01378479
LEFT	0.060446300	0.12241394	0.07174088	0.049657803	0.05466605
RIGHT	0.060281385	0.07174088	0.10792208	0.048889920	0.03761596
BOTTOM	0.058919810	0.04965780	0.04888992	0.412444857	-0.26281076
TOP	0.013784787	0.05466605	0.03761596	-0.262810761	0.42322820
DIAGONAL	0.004637188	-0.03786848	-0.01666667	0.003514739	-0.07857143

	DIAGONAL
LENGTH	0.004637188
LEFT	-0.037868481
RIGHT	-0.016666667
BOTTOM	0.003514739
TOP	-0.078571429
DIAGONAL	0.199092971

	LENGTH	LEFT	RIGHT	BOTTOM	TOP
LENGTH	0.151480107	0.06044630	0.06028139	0.058919810	0.01378479
LEFT	0.060446300	0.12241394	0.07174088	0.049657803	0.05466605
RIGHT	0.060281385	0.07174088	0.10792208	0.048889920	0.03761596
BOTTOM	0.058919810	0.04965780	0.04888992	0.412444857	-0.26281076
TOP	0.013784787	0.05466605	0.03761596	-0.262810761	0.42322820
DIAGONAL	0.004637188	-0.03786848	-0.01666667	0.003514739	-0.07857143

	DIAGONAL
LENGTH	0.004637188
LEFT	-0.037868481
RIGHT	-0.016666667
BOTTOM	0.003514739
TOP	-0.078571429
DIAGONAL	0.199092971

P2.3 The function to verify if a point lies inside of the six-dimensional ellipsoid that serves as the 95% confidence region for the mean value of banknotes.

In [ ]:
```r
is_inside_ellipsoid <- function(point, mean_vector, cov_matrix, n, p, alpha = 0.05) {
  t_square <- n * t(point - mean_vector) %*% solve(cov_matrix) %*% (point - mean_vector)
  critical_value <- p*(n-1)/(n-p)*qf(1-alpha, p, n-p) / n
  if (t_square <= critical_value) {
    return(TRUE)
  } else {
    return(FALSE)
  }
}
```

P2.4 Check if the obtained mean values are within the Hotelling's confidence region that was obtained based on the original sample of banknotes

In [ ]:
```r

```

```
[1] FALSE
```

P2.5 Check if the new mean vector falls within the Bonferroni's confidence rectangular region for the mean value of the old bank note dimensions.

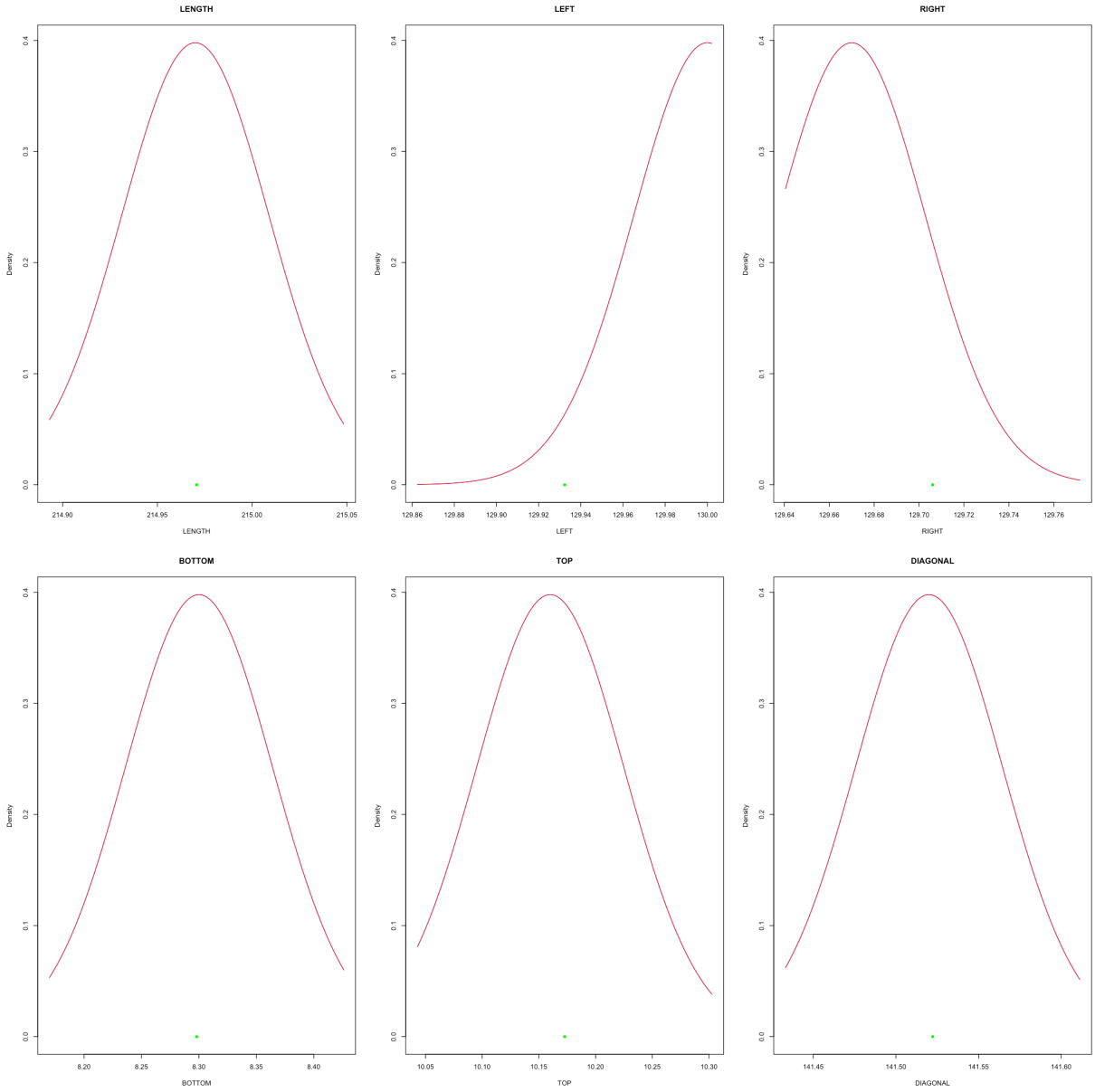
LENGTH	LEFT	RIGHT	BOTTOM	TOP	DIAGONAL
TRUE	FALSE	FALSE	TRUE	TRUE	TRUE

P2.6 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

We can observe three things from the graphs, the mean position, interval width and skewness; as we can see all the data show a central tendency, for the mean values of LENGTH, BOTTOM, TOP, DIAGONAL are located in the centre which means they have a very small bias, but for LEFT and RIGHT of the banknote were near the edge of the confidence intervals, this means skewness in the data distribution for that variable.

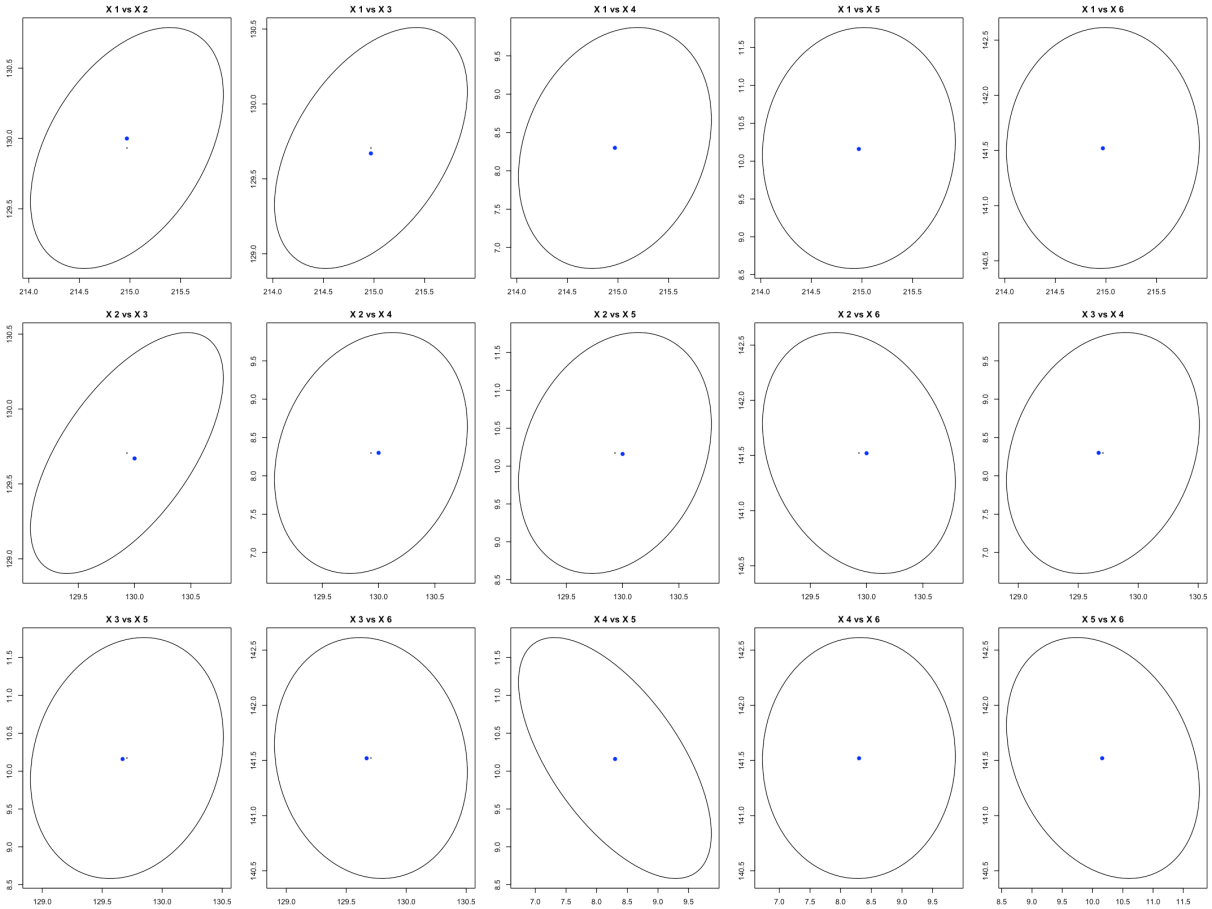
In [ ]:
```r

```



P2.7 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.

In [ ]:



P2.8 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.

The Hotelling's T^2 ellipse is shaped according to the covariance structure of the variables, it stands to reason that if the new mean vector falls outside this ellipse, it indicates that the new mean exhibits significant deviations from the variance captured within the ellipse. Since we currently don't have the correct projection (will be fixed later), we can't go any deeper into the interpretation. However, from the projection on one-dimensional space, we can see that the left and right dimensions of the new banknote have significant deviations, with the most significant change in the size of the left side of the banknote.

P2.9 Check if the new vector of means are within: a) Hotelling's confidence region; b) Bonferroni's confidence region.

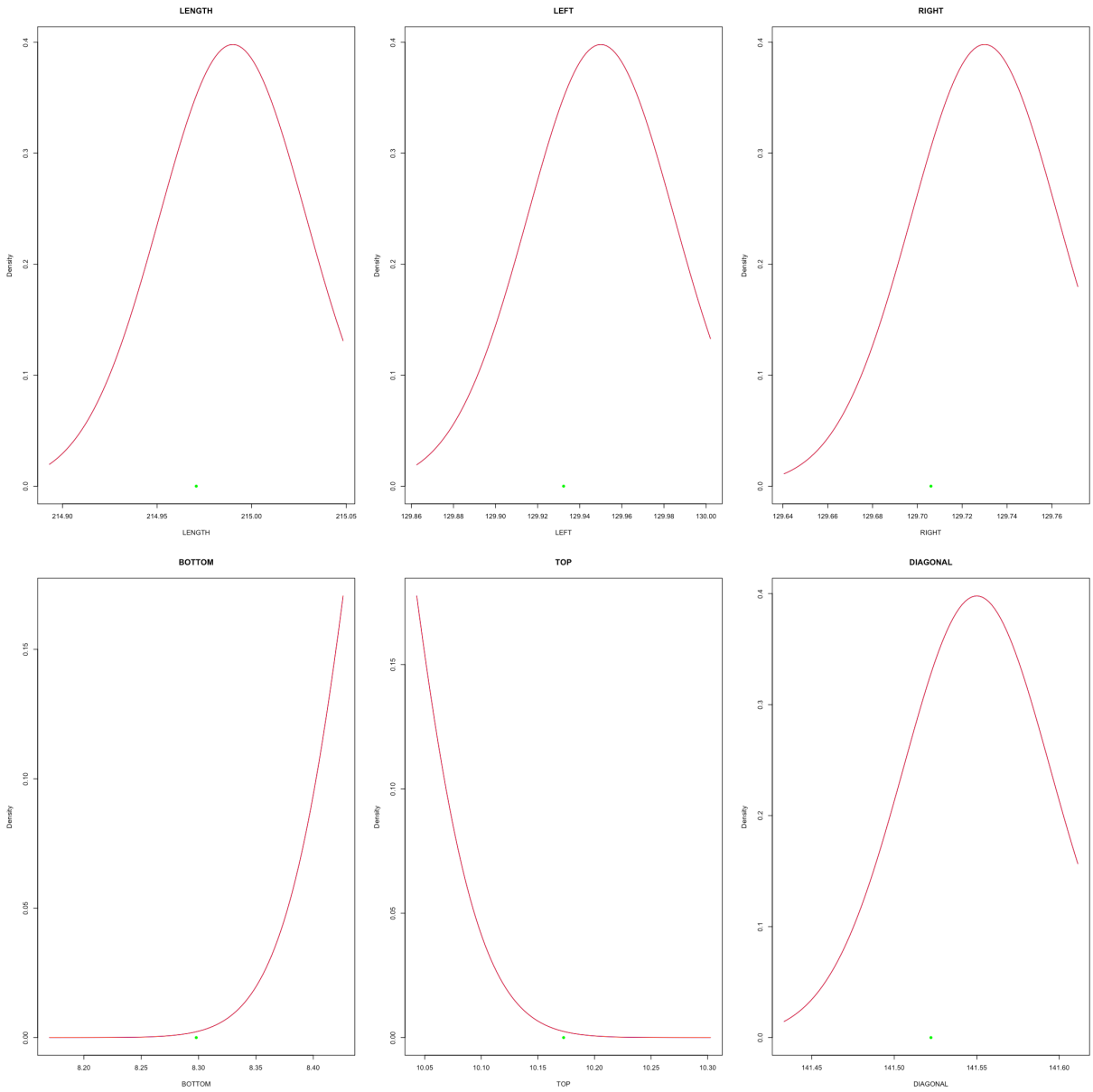
The new mean vector of bank note dimensions falls not within Hotelling's confidence region and Bonferroni's confidence intervals.

1. The new mean vector does not fall within the Hotelling's T^2 confidence region, it indicates a multivariate deviation from the expected mean vector based on the original data, this could also indicate an issue with the new production line that needs immediate attention to ensure quality and consistency.
2. The new mean vector's dimensions (LENGTH, LEFT, RIGHT, BOTTOM, TOP, DIAGONAL) falling outside their Bonferroni-adjusted confidence intervals suggests significant deviation in each dimension. As we can see from the graph provided below, the new banknote creates a larger deviation at the top and bottom.

In [ ]:

```
[1] "Inside Hotelling’s confidence region:"
[1] FALSE
[1] "Inside Bonferroni’s confidence region:"
[1] FALSE
[1] "Inside Bonferroni’s confidence region:"
  LENGTH     LEFT    RIGHT   BOTTOM      TOP DIAGONAL
   FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
```

In [ ]:



P2.10 After yet another tuning, the vector of means was m2; Is this value acceptable based on the original sample of the bank notes, or the production line still needs some tuning? Explain your answer.

The results from the latest testing round after tuning the production line still indicate that the mean vector from the new production does not fall within either the Hotelling's T^2 confidence region or the Bonferroni confidence intervals.

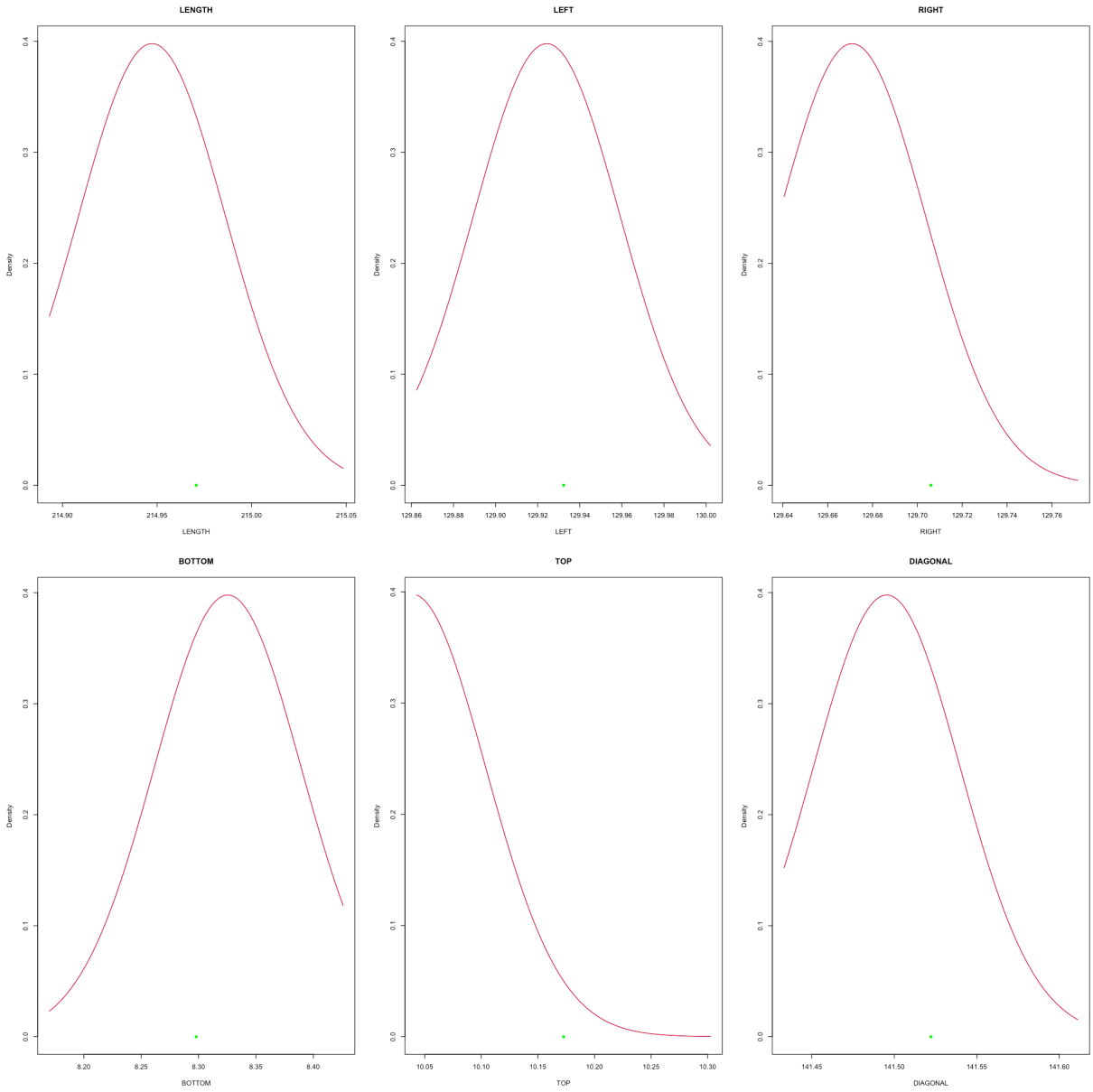
1. The persistent placement of the new mean vector outside Hotelling's confidence region, even after adjustments, suggests a consistent multivariate anomaly. This indicates that the combined variance and correlation structure of the banknote dimensions differs significantly from historical data.
2. The continuous non-conformance across all individual dimensions (LENGTH, LEFT, RIGHT, BOTTOM, TOP, DIAGONAL), as indicated by the Bonferroni tests.

In conclusion, as can be seen from the graph, the latest production of banknotes is better than the previous sample, the bottom deviation reduced. But the data shows that the new line's output is inconsistent with established patterns and expectations based on historical production data. The persistent exclusion from the statistical confidence regions necessitates a comprehensive and methodical approach to address the issues.

In [ ]:

```
[1] "Inside Hotelling's confidence region:"
[1] FALSE
[1] FALSE
[1] "Inside Bonferroni's confidence region:"
  LENGTH     LEFT    RIGHT   BOTTOM      TOP DIAGONAL
   FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
```

In [ ]:



Simulation: Multiple testing

Consider the sequence of independent random variables $X_1, \ldots, X_p$ such that $X_i \sim N(\mu_i, 1)$ and the problem of the multiple testing of the hypotheses H0i : μi = 0, for $i \in 1, \ldots, p$. For $p = 5000$ and $\alpha = 0.05$ use the simulations (at least 1000 replicates) to estimate FWER, FDR and the power of the Bonferroni and the Benjamini-Hochberg multiple testing procedures for the following setups

a) $\mu_1 = . . = \mu_{10} = \sqrt{2 \log p}, \mu_{11} = \ldots = \mu_p = 0$

b) $\mu_1 = . . = \mu_{500} = \sqrt{2 \log p}, \mu_{501} = \ldots = \mu_p = 0$

Bonferroni	FWER	FDR	Power
μ_{10}	1.000000000	0.009928717	0.388900000
μ_{500}	1.0000000000	0.0002901118	0.3859460000

Benjamini-Hochberg	FWER	FDR	Power
μ_{10}	1.00000000	0.05137105	0.54660000
μ_{500}	1.0000000	0.0453451	0.9039780

Scenario (a): First 10 means non-zero

Bonferroni Correction:

1. FWER: 1.000 - Indicates that the Bonferroni correction is very conservative, as it results in at least one Type I error in every simulation when using a strict α correction.
2. FDR: 0.00993 - Very low false discovery rate suggests that when the Bonferroni correction does lead to rejections, they are almost always correct.
3. Power: 0.389 - The power is moderate, meaning it correctly identifies about 39% of the true effects.

Benjamini-Hochberg Correction:

1. FWER: 1.000 - Similar to Bonferroni, the FWER is maximized.
2. FDR: 0.051 - Higher FDR than Bonferroni, but still controlled at a reasonable level, suggesting a good balance between identifying true positives and controlling false discoveries.
3. Power: 0.547 - Higher power than Bonferroni, indicating better ability to detect true positives.

Scenario (a): First 500 means non-zero

Bonferroni Correction:

1. FWER: 1.000 - Again shows that Bonferroni's conservative nature leads to at least one Type I error in every simulation.
2. FDR: 0.00029 - Extremely low FDR, indicating almost no false discoveries among the rejected hypotheses.
3. Power: 0.386 - Similar power to Scenario (a), despite many more true effects, highlighting Bonferroni's limitation in scenarios with many signals.

Benjamini-Hochberg Correction:

1. FWER: 1.000 - Consistently high across scenarios.
2. FDR: 0.045 - Well controlled under 5%, indicating a good balance in managing false discoveries.
3. Power: 0.904 - Very high power, demonstrating excellent ability to detect true effects, particularly in settings with a larger number of non-zero means.

In [ ]:

```
$a
$a$Bonferroni
       FWER         FDR       Power
1.000000000 0.009228403 0.386500000

$a$BH
      FWER        FDR      Power
1.00000000 0.05882353 0.54720000


$b
$b$Bonferroni
        FWER          FDR        Power
1.0000000000 0.0001975832 0.3845720000

$b$BH
      FWER        FDR      Power
1.00000000 0.04541024 0.90291400
```