Faculty of Economics and Political Science, Cairo University

Major Statistics

Theory of Statistics III

Submitted to Dr Niveen El Zayat

Due to 1/1/2024

Submitted by:

Nour Mohamed Nashat Eletreby 5200917

Haneen Ahmed Abdelaziz Elgabry 5200199

Sara Emad Tawfek Nasim 5200298

Introduction

Utilizing a given set of population mean vectors and a variance-covariance matrix, we aim to generate a sample that contains 30 observations from a tri-variate normal distribution. Subsequently, we will plot the joint probability density function (pdf) for each bivariate normal distribution formed by every pair of variables. Histograms will be created to visualize the distribution of each variable within the simulated data, followed by a scatter matrix plot illustrating the relationships among all pairs of the three random variables.

The statistical significance of the population means for each of the three random variables will be assessed. Subsequently, the entire process will be replicated with an increased sample size of 1000, and the impact of this larger sample size on the previous results will be discussed.

Our analysis will be done through R software.

1- Use R software to simulate a sample of size 30 observations from tri-variate normal distribution (given a specific vector of population means and a matrix of variance-covariance). (Hint: you can use R to determine a positive-definite matrix to be used as the variance covariance matrix.)

- Starting off with generating a random sample of size 30 observations from a trivariate normal distribution with $\mu^T = [\,2, 5, 10\,]^T$,

$\boldsymbol{\Sigma}$ =

	[,1]	[,2]	[,3]
[1,]	0.19788241	0.05652169	0.19320531
[2,]	0.05652169	0.15380521	0.07817222
[3,]	0.19320531	0.07817222	0.19247703

- <u>Some descriptive measurements for our sample:</u>

	X_1	X_2	X_3
Minimum	1.101	4.138	9.119
1st Quartile	1.794	4.741	9.787
Median	2.080	4.969	10.103

Mean	2.018	4.997	10.016
3rd Quartile	2.282	5.230	10.299
Maximum	2.641	5.755	10.721

Table (1)

From table 1 we can conclude that:

- For $\boldsymbol{X_1}$ **:** The minimum value equals 1.1, while the 1st quartile equals 1.794 which means that 25% of the sample data is less than 1.794, with median and mean equal to 2.08 & 2.01 respectively. The 3rd quartile equals 2.282 which means that 75% of the sample data is less than 2.282 and, the maximum value equals 2.64.
- For $\boldsymbol{X_2}$ **:** The minimum value equals 4.138, while the 1st quartile equals 4.74 which means that 25% of the sample data is less than 4.74, with median and mean equal to 4.96 & 4.99 respectively. The 3rd quartile equals 5.23 which means that 75% of the sample data is less than 5.23 and, the maximum value equals 5.755.
- For $\boldsymbol{X_3}$ **:** The minimum value equals 9.11, while the 1st quartile equals 9.78 which means that 25% of the sample data is less than 9.78, with median and mean equal to 10.10& 10.01 respectively. The 3rd quartile equals 10.29 which means that 75% of the sample data is less than 10.29 and, the maximum value equals 10.72.
- <u>Scatter Plot for the Trivariate Distribution:</u>

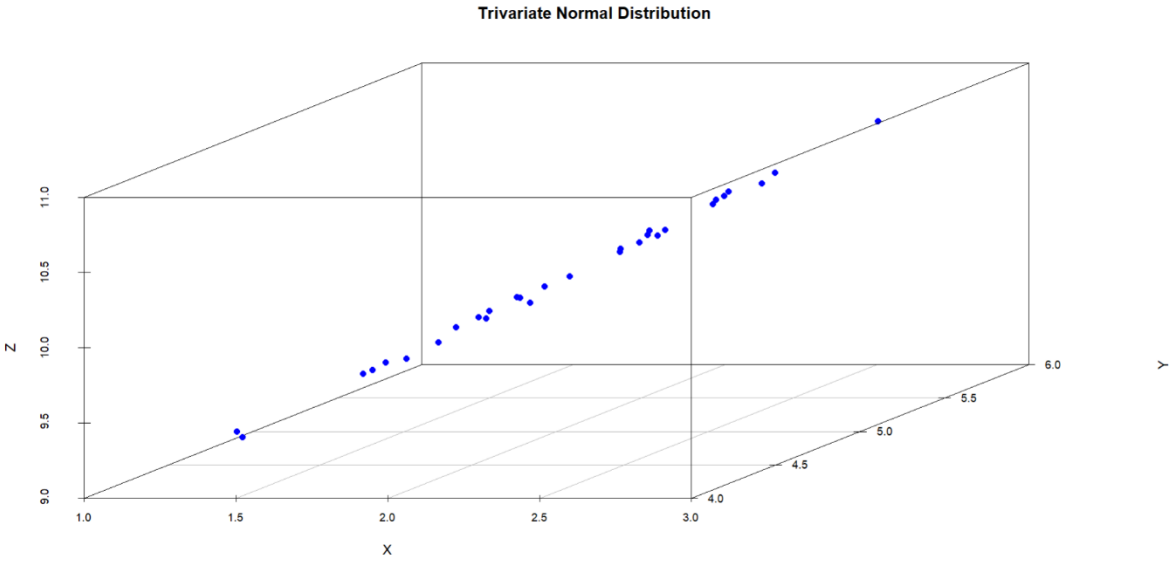

Figure (1)

- The scatter plot for the three variables shows a positive correlation.

2- Using the parameter given, Plot the surface of the joint pdf of each bivariate normal distribution (for each pair of variables) using (plot3D) package in R software. Repeat this step using I2 as variance-covariance matrix. Interpret each graph and compare between the three graphs and the graph corresponding to the bivariate standard normal distribution.

Now we will plot the surface of the joint pdf the standard bivariate normal distribution with $\mu=[0,0]T$, Σ which is identity matrix I2

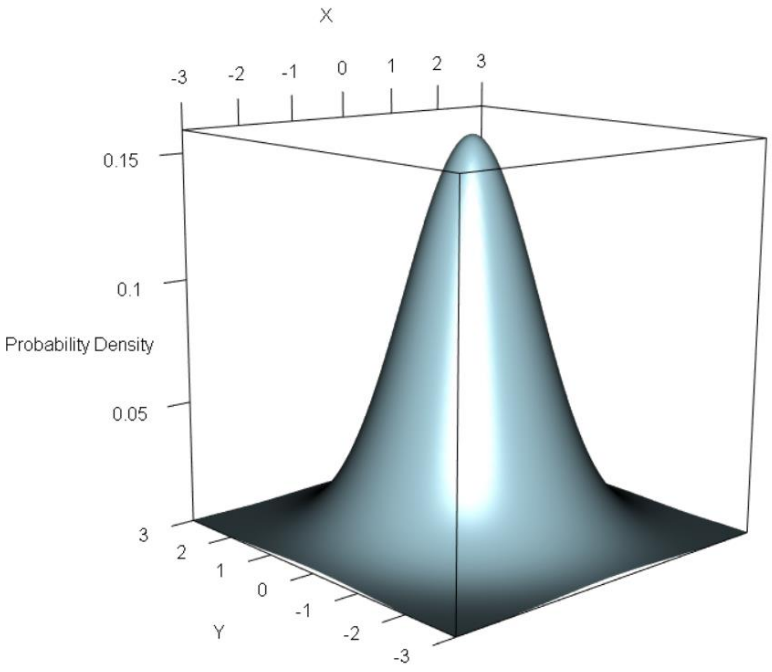

Figure (2)

We can assume both variables are uncorrelated; zero correlation, with same variance= 1 which makes the cone shape also we can see the graph centered (0,0) because of the mean vector defined. Hence we can see a perfect bell shape graph

Now we will plot the surface of the joint pdf of each pair of variables (bivariate normal distribution) one plot with our defined mean and defined matrix covariance and one with our defined mean but with the identity matrix (I2)

X1 and X2

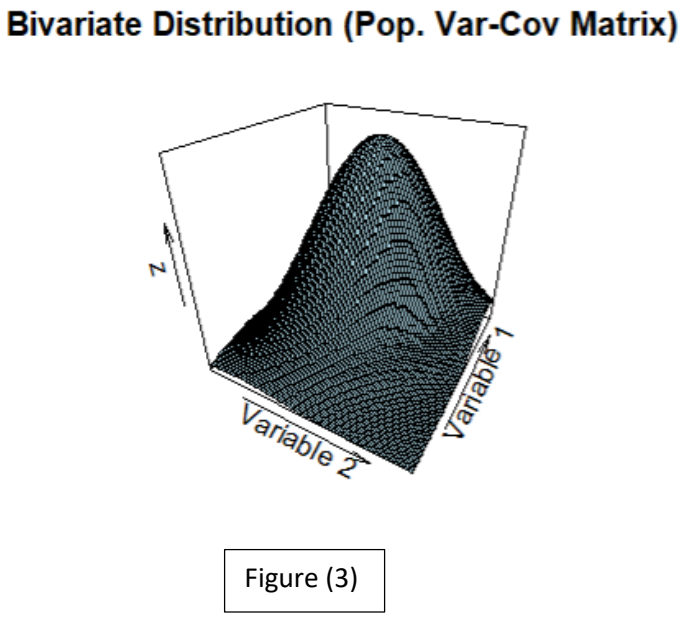

Figure (3)

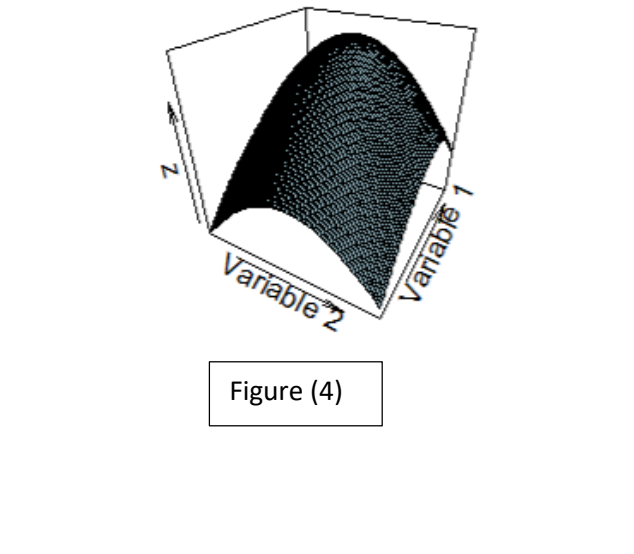

Figure (4)

From figure 3 we can see a bell shaped normal distribution between X1 and X2 where X1 ranges between 1 and 3 and X2 ranges between 4 and 6 and we can see the graph close to standard bivariate normal distribution.

From figure 4 we can see a bell shaped normal distribution between X1 and X2 where X1 ranges between 1 and 3 and X2 ranges between 4 and 6; but here we can see that the graph is more bell shaped than figure 3 due to the use of identity matrix I2 which makes it look like more as a cone due to zero correlation and variance equal 1.

X1 and X3

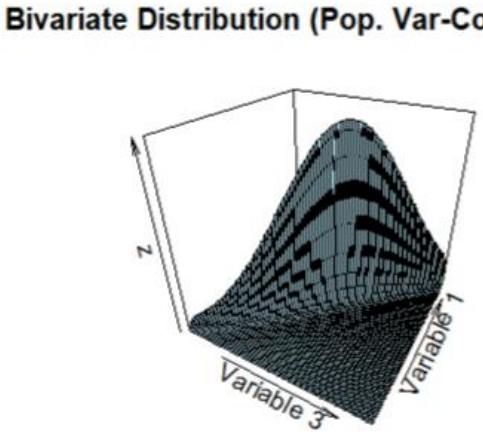

Figure (5)

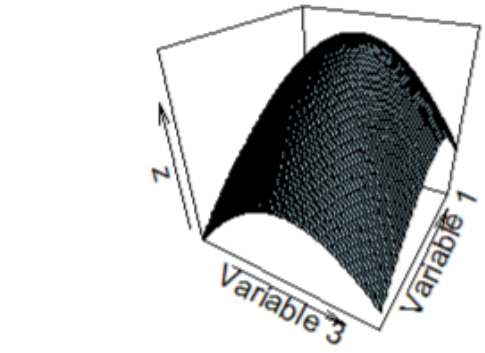

Figure (6)

From figure 5 we can see a bell shaped normal distribution between X1 and X3 where X1 ranges between 1 and 3 and X3 ranges between 9 and 11 and we can see the graph close to standard bivariate normal distribution.

From figure 6 we can see a bell shaped normal distribution between X1 and X3 where X1 ranges between 1 and 3 and X3 ranges between 9 and 11; but here we can see that the graph is more bell shaped than figure 5 due to the use of identity matrix I2 which makes it look like more as a cone due to zero correlation and variance equal 1.

X2 and X3

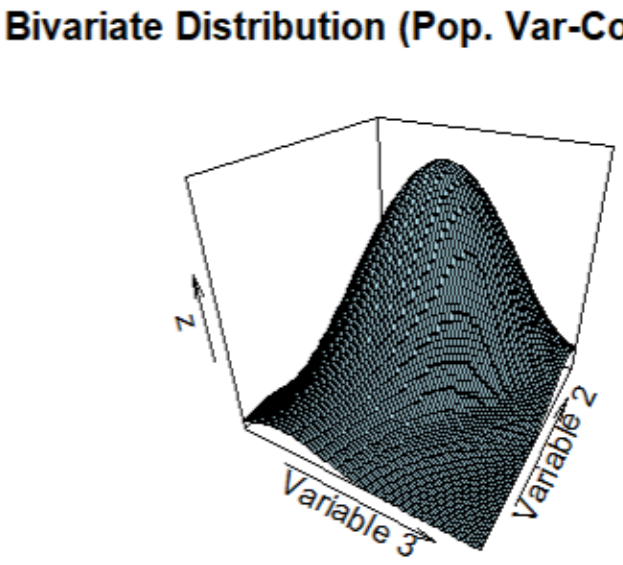

Figure (7)

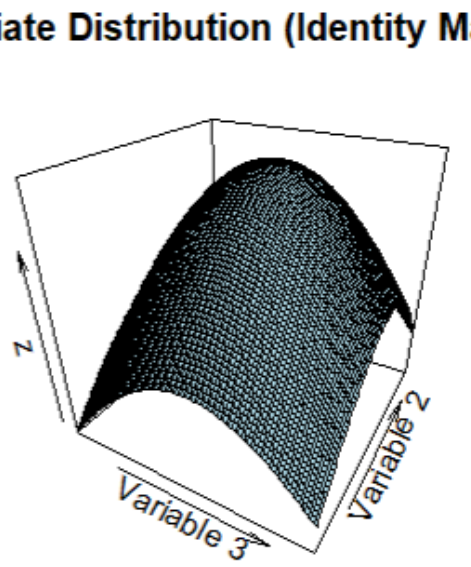

Figure (8)

From figure 7 we can see a bell shaped normal distribution between X2 and X3 where X2 ranges between 4 and 6 and X3 ranges between 9 and 11 and we can see the graph close to standard bivariate normal distribution.

From figure 8 we can see a bell shaped normal distribution between X2 and X3 where X2 ranges between 4 and 6 and X3 ranges between 9 and 11; but here we can see that the graph is more bell shaped than figure 7 due to the use of identity matrix I2 which makes it look like more as a cone due to zero correlation and variance equal 1.

3- Plot the histogram of each variable of the simulated data n=30.

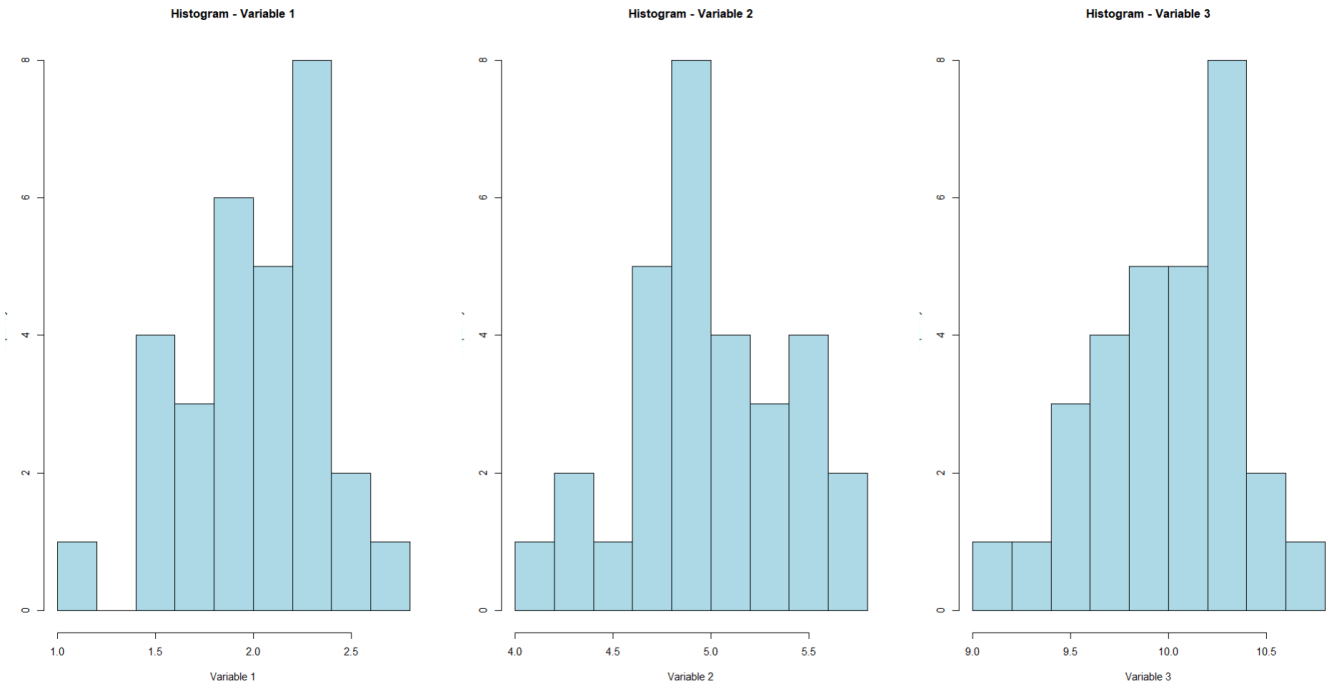

Figure (9)

- For the histogram of $\boldsymbol{X_1}$: The graph is not bell shaped. The graph is one peaked with mean equals to 2.018 and, it doesn't show signs of symmetry.
- For the histogram of $\boldsymbol{X_2}$: The graph is not bell shaped. The graph is one peaked with mean equals to 4.997 and, it doesn't show signs of symmetry.
- For the histogram of $\boldsymbol{X_3}$: The graph is nearly symmetrical with one peak and mean equals to 10.016.

4- Compute the sample mean vector and the sample variancecovariance matrix

```
> means
[1]  2  5 10
> sample_mean
[1]  2.017806  4.997106 10.016032
```

Output (1)

- We can observe that there is a very slight difference between the sample & population means.

```
> sample_var_cov_matrix
            [,1]       [,2]      [,3]
[1,] 0.12880909 0.08240849 0.1333745
[2,] 0.08240849 0.16005527 0.1032477
[3,] 0.13337448 0.10324774 0.1410930
> cov_matrix
3 x 3 Matrix of class "dpoMatrix"
            [,1]       [,2]       [,3]
[1,] 0.19788241 0.05652169 0.19320531
[2,] 0.05652169 0.15380521 0.07817222
[3,] 0.19320531 0.07817222 0.19247703
```

Output (2)

- We can observe that there is a slight difference between the sample and population variances, as well as the samples covariances between each two variables are quite close to the population covariances.

5- Sketch the scatter matrix plot to describe the relations between all pairs of the three random variables.

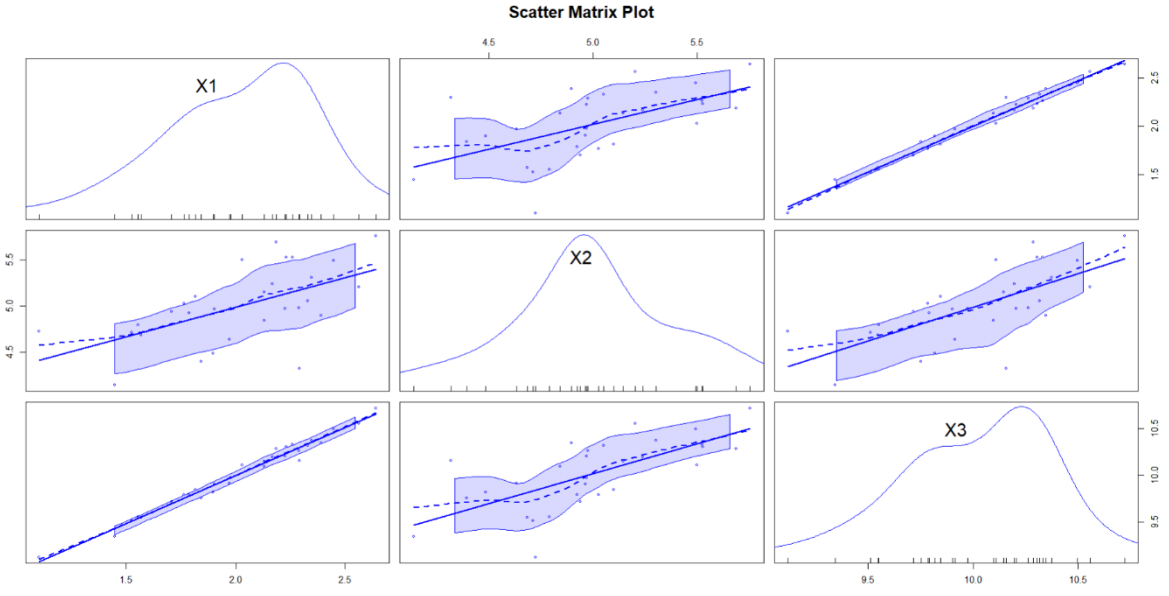

Figure (10)

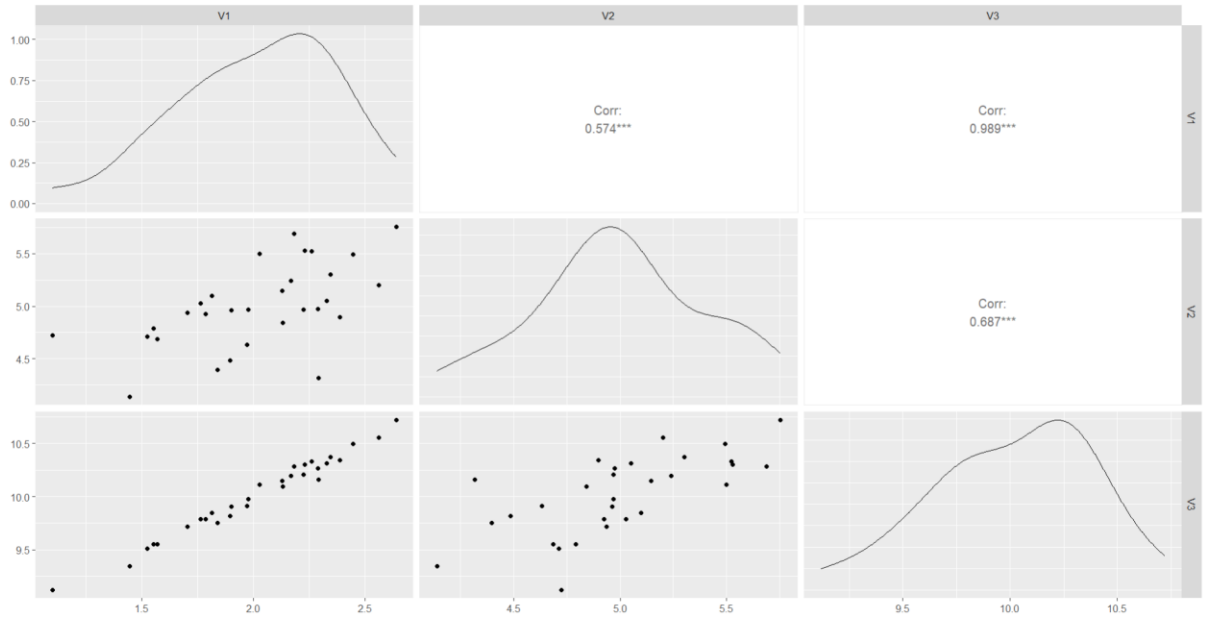

Figure (11)

- There exists a strong positive correlation between X_1 & X_3 that equals to 0.989, there also exists a strong positive correlation between X_2 & X_3 that equals to 0.687 and, there exists a moderate positive correlation between X_1 & X_2 that equals to 0.574.

6- Compute the population as well as the sample correlation matrix.

- Sample & Population Correlation Matrix (*n=30*):

```
> cov2cor(cov_matrix)
3 x 3 Matrix of class "corMatrix"
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.3239858 0.9899791
[2,] 0.3239858 1.0000000 0.4543363
[3,] 0.9899791 0.4543363 1.0000000
> cov2cor(sample_var_cov_matrix)
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.5739363 0.9893425
[2,] 0.5739363 1.0000000 0.6870570
[3,] 0.9893425 0.6870570 1.0000000
```

Output (3)

- We can observe that the correlation between X_1 & X_3 is almost the same for the sample and the population. However, for the correlation between X_1 & X_2 it's different, it's a

weak to moderate correlation in the population with correlation equals to 0.32 while it's a moderate to strong correlation in the sample with correlation equals to 0.57. Also, for the correlation between $\boldsymbol{X_2}$ & $\boldsymbol{X_3}$, it's different as well, the population correlation is 0.45 which is considered a moderate correlation, as for the sample correlation it's 0.68 which is considered a strong correlation.

7- Test the significance of each of the population mean of each of the three random variables.

```
> t.test(sample_data[,1])

        One Sample t-test

data:  sample_data[, 1]
t = 30.794, df = 29, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.883791 2.151822
sample estimates:
mean of x
 2.017806
```

Output (4)

- Since the p-value is less than alpha (0.05), therefore we reject the null hypothesis, there is enough evidence to prove that the population mean does not equal zero for $\boldsymbol{X_1}$.

```
> t.test(sample_data[,2])

        One Sample t-test

data:  sample_data[, 2]
t = 68.414, df = 29, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 4.847717 5.146494
sample estimates:
mean of x
 4.997106
```

Output (5)

- Since the p-value is less than alpha (0.05), therefore we reject the null hypothesis, there is enough evidence to prove that the population mean does not equal zero for $\boldsymbol{X_2}$.

```
> t.test(sample_data[,3])

        One Sample t-test

data:  sample_data[, 3]
t = 146.05, df = 29, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  9.875772 10.156292
sample estimates:
mean of x
 10.01603
```

Output (6)

- Since the p-value is less than alpha (0.05), therefore we reject the null hypothesis, there is enough evidence to prove that the population mean does not equal zero for $\boldsymbol{X_3}$.

Now we will repeat the same steps but for a greater sample n=1000

- **Use R software to simulate a sample of size 1000 observations from tri-variate normal distribution**
- Now we will be generating a random sample of size 1000 observations from a trivariate normal distribution with $\mu^T = [\,2, 5, 10\,]^T$,

$\boldsymbol{\Sigma}$ =

	[,1]	[,2]	[,3]
[1,]	0.19788241	0.05652169	0.19320531
[2,]	0.05652169	0.15380521	0.07817222
[3,]	0.19320531	0.07817222	0.19247703

- Some descriptive measurements for our sample:

	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$
Minimum	0.5481	3.817	8.666
1st Quartile	1.6984	4.736	9.709

Median	2.0077	4.997	10.011
Mean	2.0099	4.998	10.009
3rd Quartile	2.3095	5.241	10.298
Maximum	3.3856	6.042	11.406

Table (2)

From table 2 we can conclude that:

- For $\boldsymbol{X_1}$ **:** The minimum value equals 0.5, while the 1st quartile equals 1.6948 which means that 25% of the sample data is less than 1.6948, with median and mean equal to 2.007 & 2.0099 respectively. The 3rd quartile equals 2.3095 which means that 75% of the sample data is less than 2.3095 and, the maximum value equals 3.3856.
- For $\boldsymbol{X_2}$ **:** The minimum value equals 3.817, while the 1st quartile equals 4.736 which means that 25% of the sample data is less than 4.736, with median and mean equal to 4.997 & 4.998 respectively. The 3rd quartile equals 5.241 which means that 75% of the sample data is less than 5.241 and, the maximum value equals 6.042.
- For $\boldsymbol{X_3}$ **:** The minimum value equals 8.666, while the 1st quartile equals 9.709 which means that 25% of the sample data is less than 9.709 , with median and mean equal to 10.011& 10.009 respectively. The 3rd quartile equals 10.298 which means that 75% of the sample data is less than 10.298 and, the maximum value equals 11.406.
- The value of the mean of each variable (2.0099, 4.998, 10.009) are very close to the population mean vector (2, 5, 10) ; the larger the sample the better indication about the population and the estimates are closer to the actual value.

8- Plot the histogram of each variable of the simulated data n=30.

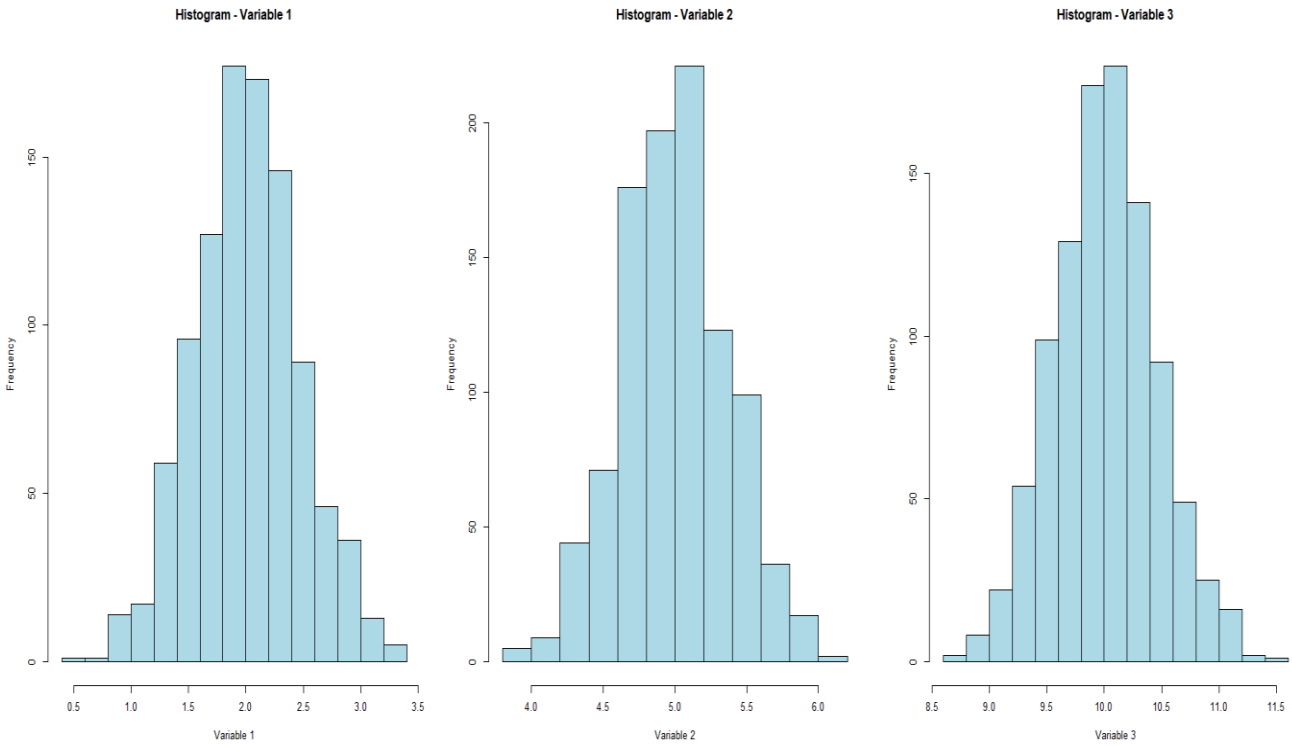

Figure (12)

- For the histogram of $\boldsymbol{X_1}$: after comparing the histograms of X1 in n=30 and n=1000, we find that when we increased the sample size this improved the bell shape obviously, it became closer to the normal shape (showing a slight of symmetric distribution)
- For the histogram of $\boldsymbol{X_2}$: after comparing the histograms of X2 in n=30 and n=1000, we find that when we increased the sample size this improved the bell shape obviously, it became closer to the normal shape
- For the histogram of $\boldsymbol{X_3}$: after comparing the histograms of X3 in n=30 and n=1000, we find that when we increased the sample size this improved the bell shape obviously, it became closer to the normal shape (showing a slight of symmetric distribution)

9- Compute the sample mean vector and the sample variancecovariance matrix

```
> sample_mean
[1]  2.009931  4.998055 10.008899
> sample_var_cov_matrix
           [,1]       [,2]       [,3]
[1,] 0.20854388 0.04728188 0.20156368
[2,] 0.04728188 0.14041076 0.06735564
[3,] 0.20156368 0.06735564 0.19843340
```

Output (7)

- The values of the sample means are extremely close to the mean vector.
- We can observe that there is a very slight difference between the sample and population variances, as well as the samples covariances between each two variables are quite close to the population covariances and more close than the small sample.
- When conducting the analysis with a larger sample size while keeping the population mean vector and variance-covariance matrix constant, we noticed an enhancement in the results, with values converging more closely to those of the population. Therefore, the increased sample size, along with the fixation of the population mean vector and variance-covariance matrix, positively influenced the improvement of values, bringing them closer to the population values.

10- Sketch the scatter matrix plot to describe the relations between all pairs of the three random variables.

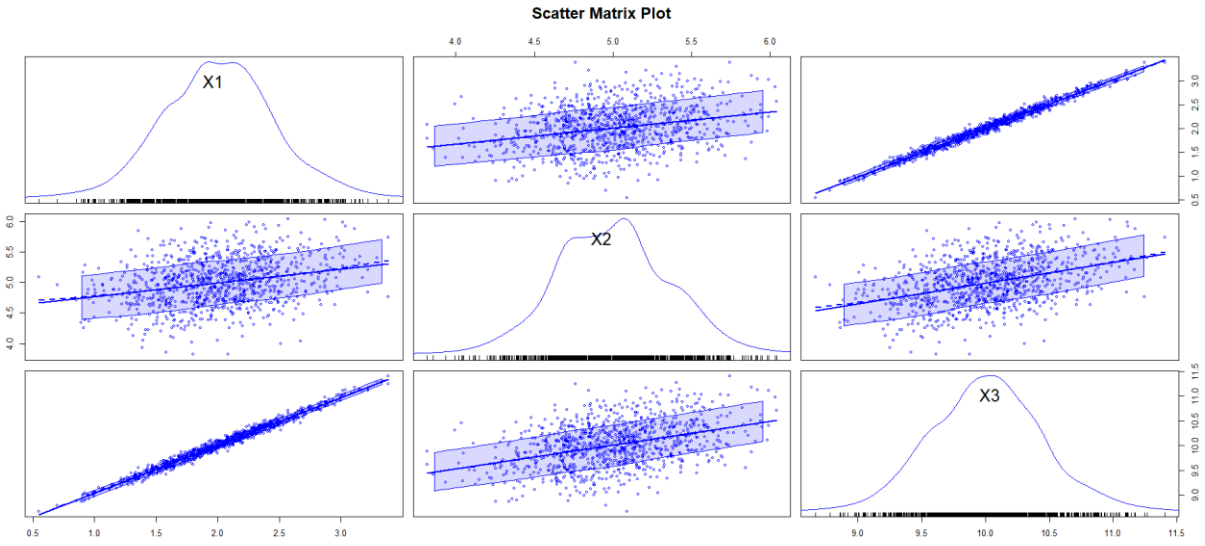

Figure (13)

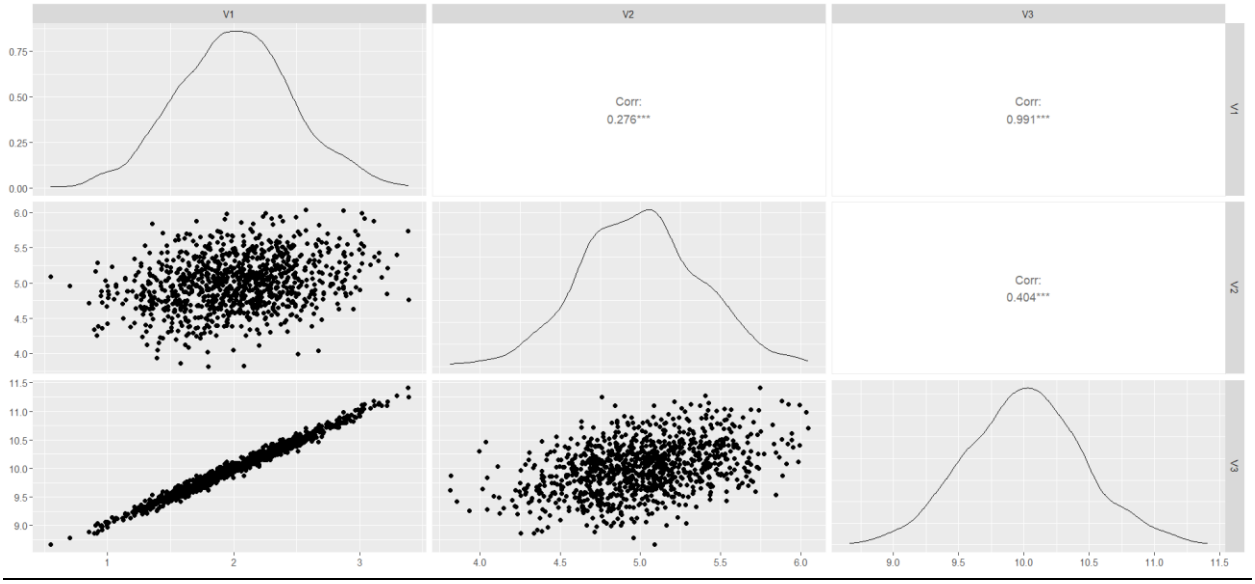

Figure (14)

- There exists a strong positive correlation between $\boldsymbol{X_1}$ & $\boldsymbol{X_3}$ that equals to 0.991, there also exists a moderate positive correlation between $\boldsymbol{X_2}$ & $\boldsymbol{X_3}$ that equals to 0.404 and, there exists a weak positive correlation between $\boldsymbol{X_1}$ & $\boldsymbol{X_2}$ that equals to 0.276. The results are close to the smaller sample but patterns are better visualized when the sample size got bigger.

11- Compute the population as well as the sample correlation matrix.

```
> population_corr_matrix
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.2763095 0.9908460
[2,] 0.2763095 1.0000000 0.4035212
[3,] 0.9908460 0.4035212 1.0000000
> sample_corr_matrix
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.2763095 0.9908460
[2,] 0.2763095 1.0000000 0.4035212
[3,] 0.9908460 0.4035212 1.0000000
```

Output (8)

- We can observe that the correlations between the variables are exactly the same in the sample and the populations indicating that when having a larger sample size the sample values become closer to the population values

12- Test the significance of each of the population mean of each of the three random variables.

```
> t.test(larger_sample_data[,1])

        One Sample t-test

data:  larger_sample_data[, 1]
t = 139.18, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.981592 2.038269
sample estimates:
mean of x
 2.009931
```

Output (9)

- Since the p-value is less than alpha (0.05), therefore we reject the null hypothesis, there is enough evidence to prove that the population mean does not equal zero for $\boldsymbol{X_1}$.

```
> t.test(larger_sample_data[,2])

        One Sample t-test

data:  larger_sample_data[, 2]
t = 421.79, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 4.974802 5.021308
sample estimates:
mean of x
 4.998055
```

Output (9)

- Since the p-value is less than alpha (0.05), therefore we reject the null hypothesis, there is enough evidence to prove that the population mean does not equal zero for $\boldsymbol{X_2}$.

```
> t.test(larger_sample_data[,3])

        One Sample t-test

data:  larger_sample_data[, 3]
t = 710.52, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  9.981256 10.036542
sample estimates:
mean of x
  10.0089
```

Output (10)

- Since the p-value is less than alpha (0.05), therefore we reject the null hypothesis, there is enough evidence to prove that the population mean does not equal zero for $\boldsymbol{X_3}$.
- The results concerning the mean significance are more reliable when the sample size got bigger.

Conclusion

In summary, our analysis in R showed that increasing the sample size from 30 to 1000 consistently improved the accuracy of our results. Here's what we found:

1. **Histograms:**
    - Histograms of each variable became more accurate representations of the population distribution as the sample size increased.
2. **Sample Statistics:**
    - Sample mean and variance-covariance matrix estimates approached population values with a larger sample size.
3. **Scatter Plots:**
    - Relationships between variables became clearer in scatter plots as the sample size grew, resembling the population patterns more closely.

4. **Correlation Matrices:**
    - Both population and sample correlation matrices showed improved accuracy with a larger sample size.
5. **Mean Significance Tests:**
    - Testing the significance of population means became more reliable with a larger sample size.

Appendix

```
# Set the seed for reproducibility
set.seed(123)
library(mvtnorm)

# Specify the population means for the three variables
means <- c(2, 5, 10)

# Generate a positive-definite matrix for the variance-covariance matrix
library(Matrix)
cov_matrix <- nearPD(matrix(rnorm(9), ncol = 3))$mat
cov_matrix

# Simulate a sample of size 30 from the tri-variate normal distribution
sample_data <- MASS::mvrnorm(n = 30, mu = means, Sigma = cov_matrix)

# Display the simulated data
print(sample_data)

#Scatter plot for the trivariate distribution
scatterplot3d(sample_data[,1], sample_data[,2], sample_data[,3],
              xlab = "X", ylab = "Y", zlab = "Z",
              color = "blue", pch = 16, main = "Trivariate Normal Distribution")

# Function to compute bivariate normal density
bivariate_normal_density <- function(x, y, means, cov_matrix) {
  vec <- cbind(x, y)
  dmvnorm(vec, mean = means, sigma = cov_matrix)
}
```

```
# Step 2: Plot the surface of the joint pdf for each bivariate normal distribution
# Using var_cov_matrix
library(plot3D)
means <- colMeans(sample_data)
cov_matrix <- cov(sample_data)

par(mfrow=c(2,2))
for (i in 1:3) {
  for (j in 1:3) {
    if (i != j) {
      x_vals <- seq(min(sample_data[, i]), max(sample_data[, i]), length.out = 100)
      y_vals <- seq(min(sample_data[, j]), max(sample_data[, j]), length.out = 100)
      z_vals <- outer(x_vals, y_vals, FUN = Vectorize(function(x, y) bivariate_normal_density(x,
y, means[c(i, j)], cov_matrix[c(i, j), c(i, j)])))

      plot_bivariate <- plot3D::persp3D(x = x_vals, y = y_vals, z = z_vals,
                                        phi = 30, theta = 30,
                                        col = "lightblue", border = "black",
                                        xlab = paste("Variable", i), ylab = paste("Variable", j),
                                        main = paste("Bivariate Distribution (Pop. Var-Cov Matrix)"))
    }
  }
}
##I2
cov_matrix_identity <- diag(3)
par(mfrow=c(2,2))
for (i in 1:3) {
  for (j in 1:3) {
    if (i != j) {
      x_vals <- seq(min(sample_data[, i]), max(sample_data[, i]), length.out = 100)
      y_vals <- seq(min(sample_data[, j]), max(sample_data[, j]), length.out = 100)
```

```
    z_vals <- outer(x_vals, y_vals, FUN = Vectorize(function(x, y) bivariate_normal_density(x,
y, means[c(i, j)], cov_matrix_identity[c(i, j), c(i, j)])))

    plot_bivariate <- plot3D::persp3D(x = x_vals, y = y_vals, z = z_vals,
                                      phi = 30, theta = 30,
                                      col = "lightblue", border = "black",
                                      xlab = paste("Variable", i), ylab = paste("Variable", j),
                                      main = paste("Bivariate Distribution (Identity Matrix)"))
  }
 }
}


# Step 3: Plot the histogram of each variable
par(mfrow=c(1,3))
for (i in 1:3) {
 hist(sample_data[, i], main = paste("Histogram - Variable", i), xlab = paste("Variable", i), col =
"lightblue", border = "black")
}

# Step 4: Compute the sample mean vector and sample variance-covariance matrix
sample_mean <- colMeans(sample_data)
sample_var_cov_matrix <- cov(sample_data)

# Step 5: Scatter matrix plot
library(car)
scatterplotMatrix(sample_data, main = "Scatter Matrix Plot")
#another way for step 5:
library(ggplot2)
library(GGally)
ggpairs(as.data.frame(sample_data))
```

```
# Step 6: Compute population and sample correlation matrix
population_corr_matrix <- cov2cor(cov_matrix)
sample_corr_matrix <- cov2cor(sample_var_cov_matrix)

# Step 7: Test significance of population mean for each variable
for (i in 1:3) {
  t_test_result <- t.test(sample_data[, i], mu = means[i])
  print(paste("Variable", i, "p-value:", t_test_result$p.value))
}
#another way for step 7
t.test(sample_data[,1])
t.test(sample_data[,2])
t.test(sample_data[,3])

# Step 8: Repeat steps 3-7 with a sample size of 1000
larger_sample_size <- 1000


larger_sample_data <- MASS::mvrnorm(n = 1000, mu = means, Sigma = cov_matrix)
summary(larger_sample_data)
# Repeat steps 3-7 with larger sample size
# Step 3: Plot the histogram of each variable
par(mfrow=c(1,3))
for (i in 1:3) {
  hist(larger_sample_data[, i], main = paste("Histogram - Variable", i), xlab = paste("Variable", i),
col = "lightblue", border = "black")
}

# Step 4: Compute the sample mean vector and sample variance-covariance matrix
sample_mean <- colMeans(larger_sample_data)
```

```
sample_var_cov_matrix <- cov(larger_sample_data)
sample_mean
sample_var_cov_matrix
# Step 5: Scatter matrix plot
library(car)
scatterplotMatrix(larger_sample_data, main = "Scatter Matrix Plot")
#another way for step 5:
library(ggplot2)
library(GGally)
ggpairs(as.data.frame(larger_sample_data))

# Step 6: Compute population and sample correlation matrix
population_corr_matrix <- cor(larger_sample_data)
sample_corr_matrix <- cor(larger_sample_data)
population_corr_matrix
sample_corr_matrix



# Step 7: Test significance of population mean for each variable
for (i in 1:3) {
  t_test_result <- t.test(larger_sample_data[, i], mu = means[i])
  print(paste("Variable", i, "p-value:", t_test_result$p.value))
}
#another way for step 7
t.test(larger_sample_data[,1])
t.test(larger_sample_data[,2])
t.test(larger_sample_data[,3])
```