Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Atividade 3
Estatística Multivariada 1

Douglas de Paula Nestlehner - 752728

Novembro, 2021

Capítulo 1

Problema Apresentado

Os escores obtidos por n = 87 estudantes no exame CLEP (College Level Examination Program), variável X1, e no exame CQT (College Qualification Test), variáveis X2 (verbal) e X3 (ciência), são apresentados na tabela abaixo.

X_1	X_2	X_3
468	41	26
428	39	26
514	53	21
547	67	33
...	...	...
527	49	30
474	41	16
441	47	26
607	67	32

Tabela 1.1: Base de dados

I) Determine os dois extremos outliers. Comente
II) Teste normalidade univariada, todas as bivariadas e trivariada. Use transformação, se necessário.
III) Teste, supondo normalidade e considerando o Teorema Central do Limite, que o vetor de médias μ é (500, 50, 25). Compare os dois resultados. Quais os possíveis valores para μ_0 em que você aceitaria Ho?

A seguir vou apresentar as respostas dos itens **I, II, III**, e no final os códigos utilizados para a resolução.

1.1 I) Outliers

Uma das maneiras mais comuns de detectar outliers em dados univariado é usando o gráfico de boxplot, porém os dados em que estamos analisando são multivariados e o mais interessante é analisar os outliers de forma conjunta.

Para isso utilizei a distância de mahalanobis (distancia de média) para poder encontrar os valores que mais se diferem (de forma conjunta) da média geral. Obtendo então os seguintes resultados:

X_1	X_2	X_3	Mahalanobis
468	41	26	2.1904869
428	39	26	3.6528798
514	53	21	0.9369832
547	67	33	5.0471274
...	...	...	...
527	49	30	2.1551612
474	41	16	4.9353476
441	47	26	2.4507607
607	67	32	2.3356446

Tabela 1.2: Distância de mahalanobis

Em seguida, ordenei as distâncias obtidas de forma decrescente, tendo então um "rank"das observações que mais se diferem da média, ou seja, os outliers.

X_1	X_2	X_3	Mahalanobis
501	25	26	11.7536848
733	73	33	7.4694158
507	32	27	7.2244111
468	57	14	7.1556187
647	58	23	6.8043086
348	28	18	6.5644915
408	28	17	6.4976688
620	71	36	5.7896308
...	...	...	...

Tabela 1.3: Distancia de mahalanobis

Com isso, é possível concluir que a observação com os resultados "501 25 26"é um outlier, pois tem a distância com valor bem maior do que as demais.

Não tive certeza em afirmar apenas observando para o valor da distância do segundo caso "733 73 33"com distancia igual a 7.4694 se era ou não um outlier. com isso plotei o gráfico representado na Figura 1.2, para poder observar o comportamento das distancia obtidas:

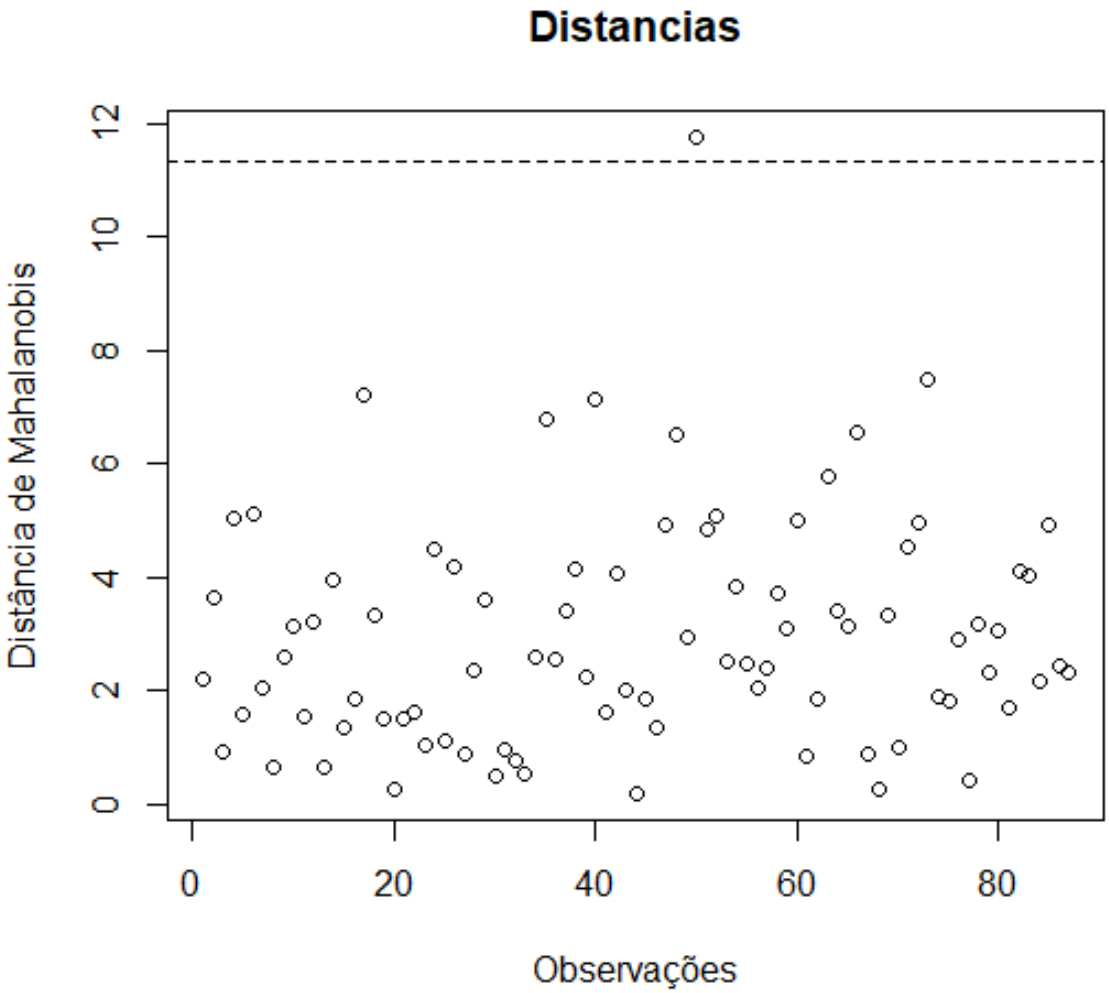

Figura 1.1: Distancia mahalanobis x Observações

A linha pontilhada representa um possível critério de diagnóstico de outlier, nesse caso optei por definir esse critério como sendo a estatística teste do Qui-quadrado(χ^2) considerando um $\alpha = 0.01$ e com 3 graus de liberdade que é o numero de covariaveis da base de dados.

Com esse critério, nota-se que apenas a observação "501 25 26"é um outlier multivariado.

1.1.1 Distancia em relação a origem

Outra maneira de detectar possiveis outlier é utilizar a distancia de mahalanobis com relação a o origem. Fazendo isso obtive o seguinte resultado:

X_1	X_2	X_3	Mahalanobis(média)	Mahalanobis(origem)
501	25	26	11.7536848	90.5432
733	73	33	7.4694158	112.1352
507	32	27	7.2244111	67.7574
468	57	14	7.1556187	56.2466
647	58	23	6.8043086	86.4551
348	28	18	6.5644915	34.9938
408	28	17	6.4976688	42.3328
620	71	36	5.7896308	90.2845
...	...	...	...	...

Tabela 1.4: Distancias

Em que a observação (733, 73, 33) se destaca das demais em relação ao valor da distancia, o que indica um possível outlier.

Também plotei o grafico do comportamento dessas distancias em relação a origem.

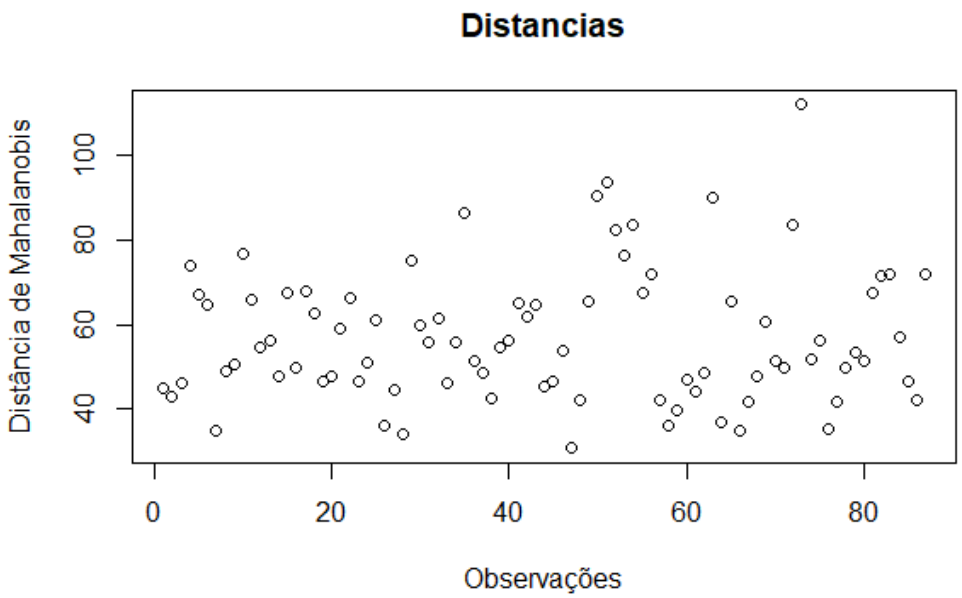

Figura 1.2: Distancia mahalanobis x Observações

Analisando as duas distâncias e os resultados obtidos concluo que os dois extremos outliers são: (501, 25, 26) e (733, 73, 33).

Depois de detectar os outiliers devemos decidir o que fazer com ele, nesse caso, decidi seguir o estudo considerando os outliers encontrados, mas se eu encontrar algum problema mais a frente (como não normalidade), poderei retira-los, substitui-los, etc.

1.2 II) - Normalidade

Primeiramente verifiquei a normalidade univariada para cada uma das covaraiveis, plotando o gráfico qqnorm, e realizando alguns testes:

1.2.1 Covariavel X1:

Gráfico qqnorm:

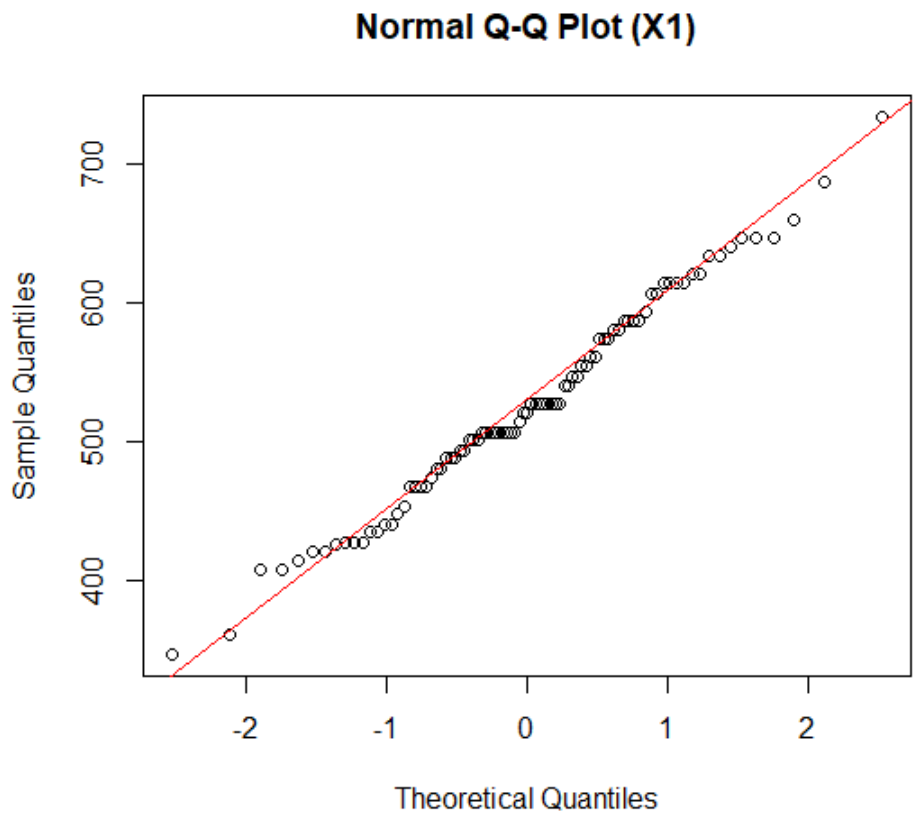

Figura 1.3: gráfico normal QQ

Aparentemente os pontos estão proximos da reta, o que indica normalidade, mas para poder confirmar a normalidade, realizei os testes de Shapiro-Wilk, Lilliefors e Anderson-Darling:

Testes:

–	Shapiro-Wilk	Lilliefors	Anderson–Darling
p-valor	0.6861	0.0481	0.3775

Tabela 1.5: Teste de normalidade para a covariavel X1

Definindo o nível de significância como sendo $\alpha = 0.01$, em todos os testes não rejeitamos H0, ou seja, a normalidade é aceita para a covariavel X1.

1.2.2 Covariavel X2:

Gráfico qqnorm:

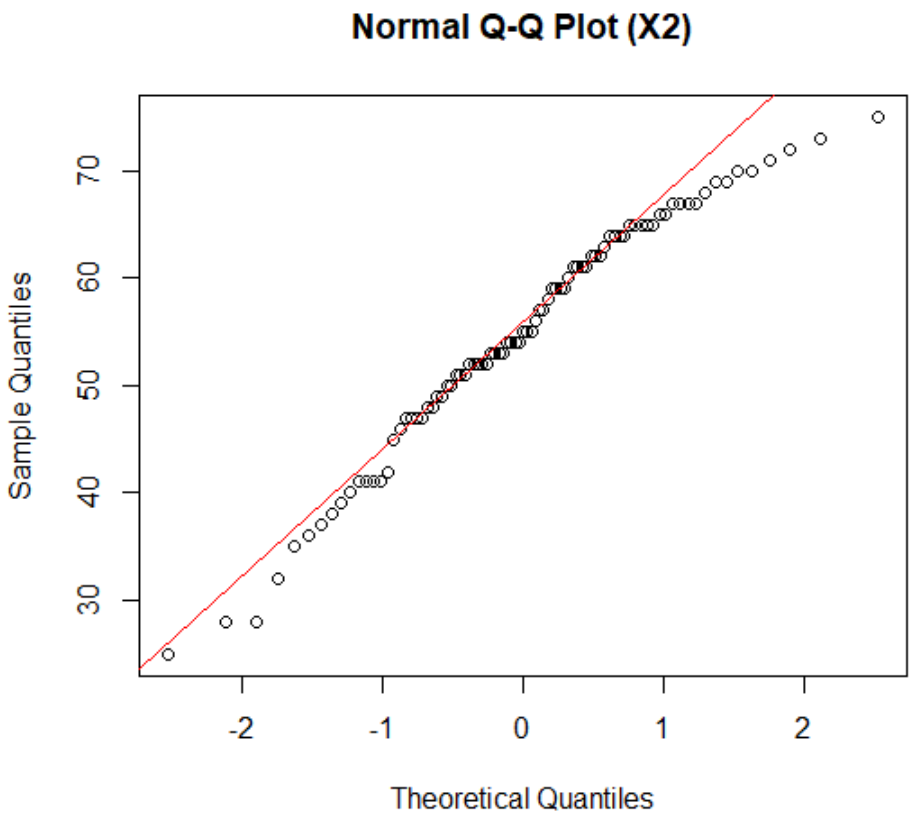

Figura 1.4: gráfico normal QQ

Alguns pontos estão distantes da reta porem outros estão bem "comportados", a conclusão apenas olhando o gráfico se torna um pouco difícil, a realização dos testes é o mais indicado:

Testes:

–	Shapiro-Wilk	Lilliefors	Anderson–Darling
p-valor	0.0387	0.1744	0.0679

Tabela 1.6: Teste de normalidade para a covariavel X2

Os p-valores estão bem próximos ao valor de decisão = 0.01, porém em todos os casos rejeitamos H0, concluindo estão que a normalidade é aceita para a covariavel X2.

1.2.3 Covariavel X3:

Gráfico qqnorm:

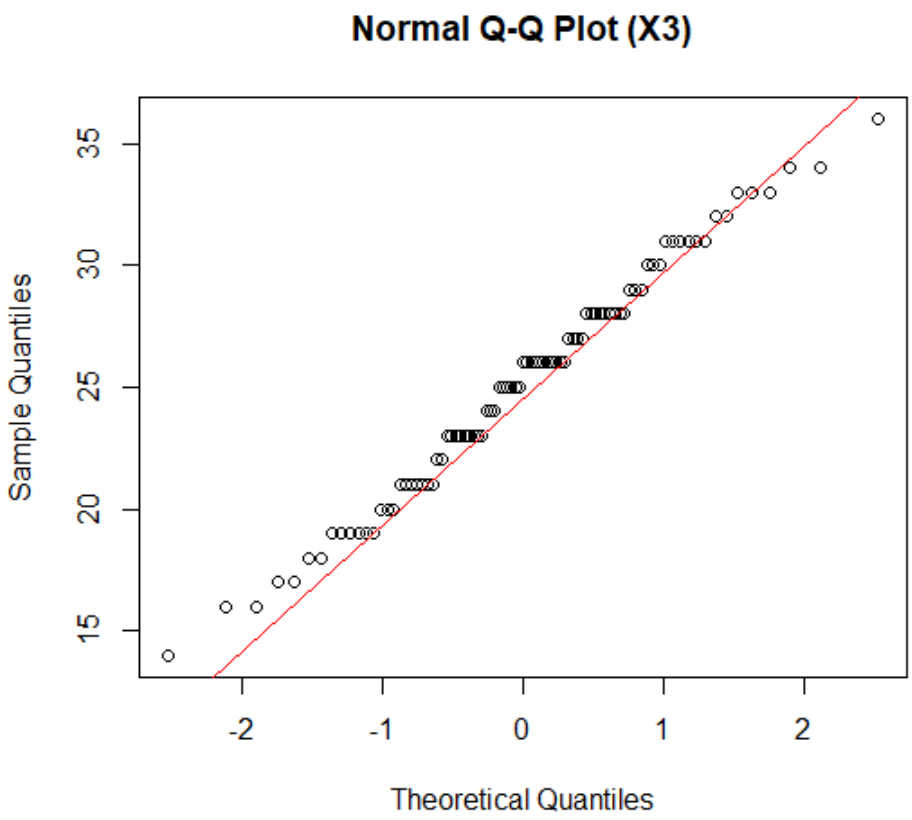

Figura 1.5: gráfico normal QQ

Os pontos estão próximos da reta e aparentemente a normalidade vai ser aceita, o que será confirmada pelos testes:

Testes:

–	Shapiro-Wilk	Lilliefors	Anderson–Darling
p-valor	0.5358	0.2168	0.3775

Tabela 1.7: Teste de normalidade para a covariavel X3

Em todos os testes rejeitamos H0, ou seja, a normalidade é aceita para a covariavel X3.

1.2.4 Normalidade Bivariada

Em seguida, assim como foi pedido no exercício, testei a normalidade de todos os possíveis casos bivariados.

X1-X2

Plotei o gráfico Gama Plot - Quantis QuiQuadrado para a normalidade Multivariada, considerando apenas as covariaveis X1 e X2:

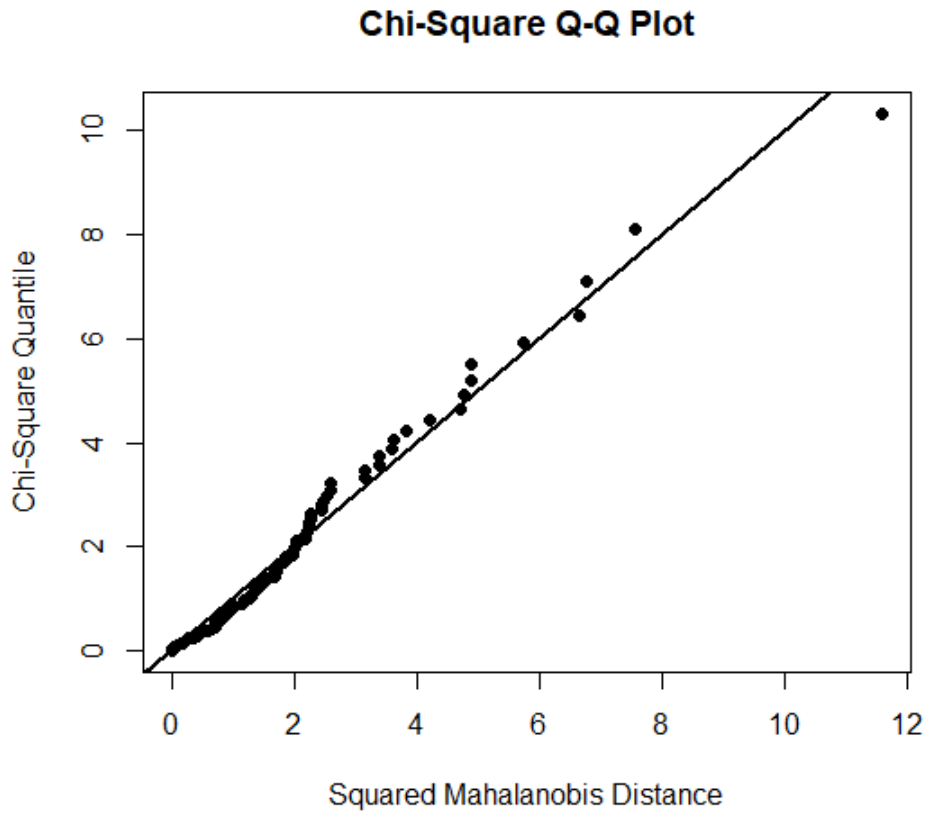

Figura 1.6: Grafico Gama Plot

Aparentemente a normalidade sera aceita, pois os pontos parecem estar proximos da reta. Para confirmar a normalidade multivariada realizei os teste de Royston, teste de Henze-Zinklers e o teste de Mardia.

–	Royston	Henze-Zinklers	Mardia
p-valor	0.1101	0.4465	0.1235 e 0.6251

Tabela 1.8: Teste de normalidade multivariada (X1 e X2)

Em todos os testes rejeita-se H0 ($\alpha = 0.01$), ou seja em todos os testes a normalidade bivariada (X1 e X2) foi aceita.

X1-X3

Grafico Gama Plot para as covariaveis X1 e X3:

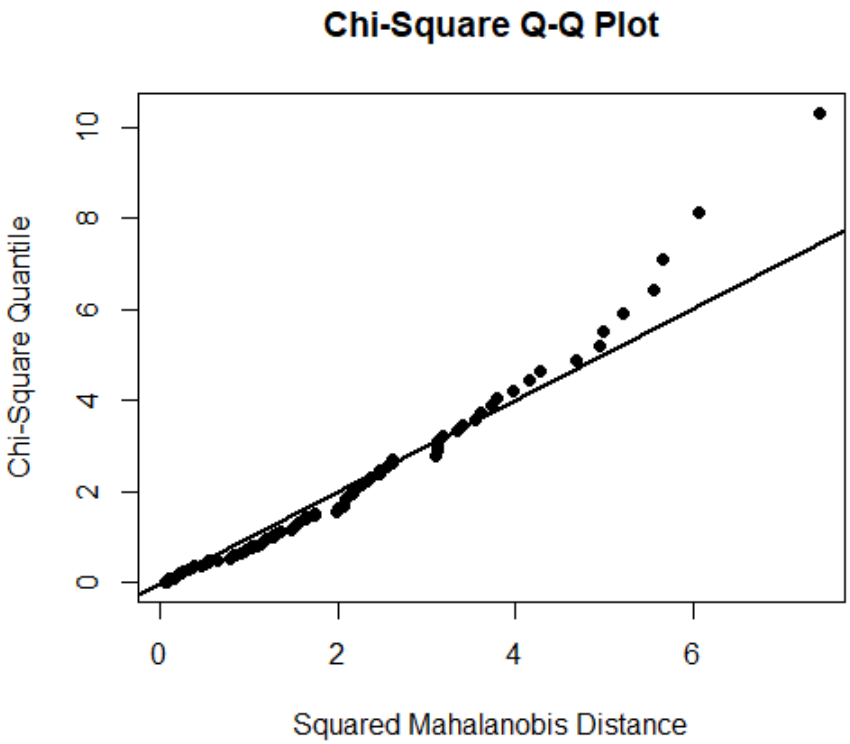

Figura 1.7: Grafico Gama Plot

Aparentemente a normalidade será aceita, pois os pontos parecem estar proximos da reta. Para confirmar a normalidade multivariada realizei os teste de Royston, teste de Henze-Zinklers e o teste de Mardia.

–	Royston	Henze-Zinklers	Mardia
p-valor	0.7614	0.6825	0.8945 e 0.0895

Tabela 1.9: Teste de normalidade multivariada (X1 e X3)

Em todos os testes rejeita-se H0 ($\alpha = 0.01$), ou seja em todos os testes a normalidade bivariada (X1 e X3) foi aceita.

X2-X3

Grafico Gama Plot para as covariaveis X2 e X3:

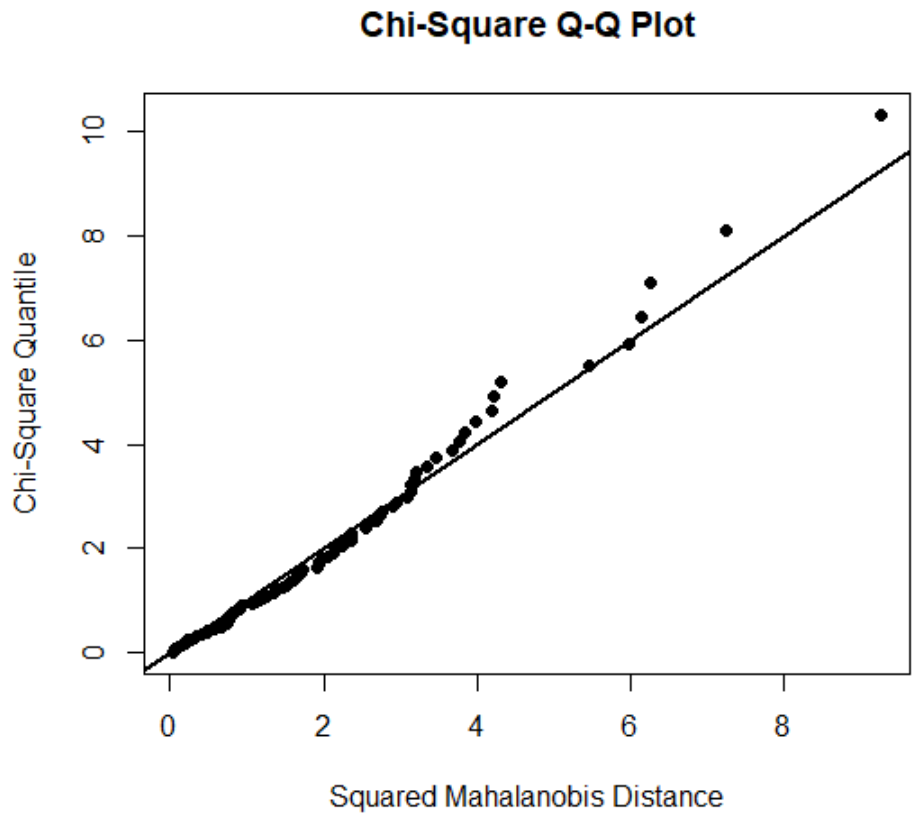

Figura 1.8: Grafico Gama Plot

Aparentemente a normalidade será aceita, pois os pontos parecem estar próximos da reta. Para confirmar a normalidade multivariada realizei os teste de Royston, teste de Henze-Zinklers e o teste de Mardia.

–	Royston	Henze-Zinklers	Mardia
p-valor	0.0971	0.3577	0.20788 e 0.2864

Tabela 1.10: Teste de normalidade multivariada (X2 e X3)

Em todos os testes rejeita-se H0 ($\alpha = 0.01$), ou seja em todos os testes a normalidade bivariada (X2 e X3) foi aceita.

1.2.5 Normalidade Trivariada

Após verificar a normalidade univariada e bivariada, só restou testar a normalidade trivariada, de forma semenlhante a bivaraiada plotei o gráfico Gama e realizei alguns testes de normalidade:

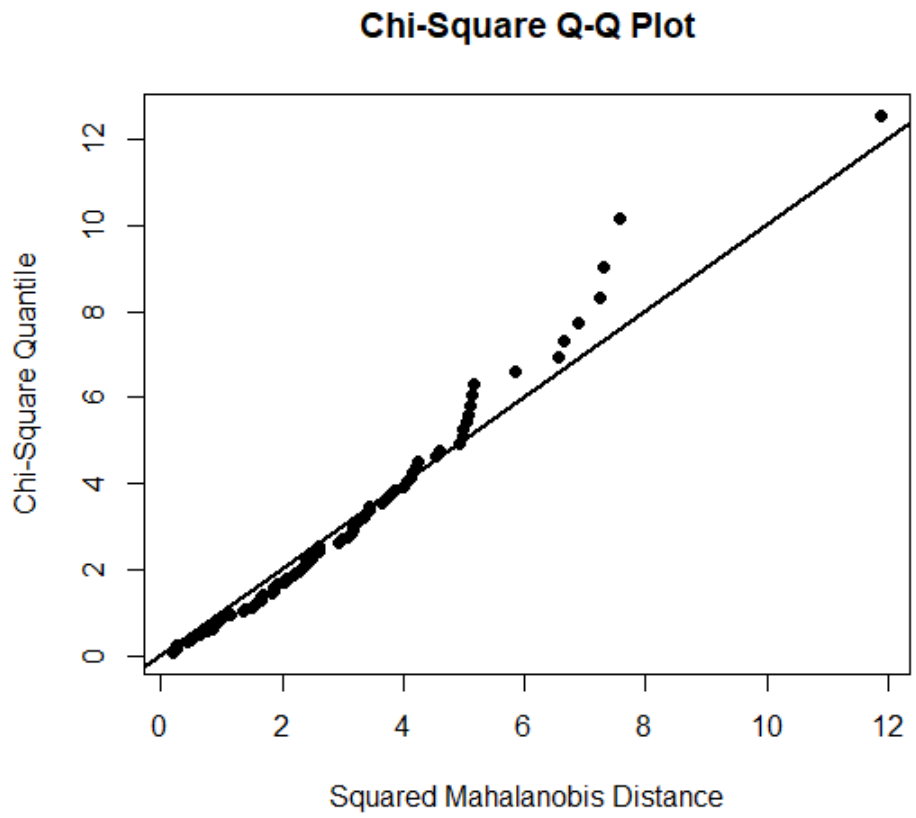

Figura 1.9: Grafico Gama Plot

Observa-se que alguns pontos estão distantes da reta, porém a um alto numero de pontos que estão "comportados", a realização dos testes é a mais indicada nesses casos.

–	Royston	Henze-Zinklers	Mardia
p-valor	0.1859	0.7235	0.4311 e 0.1095

Tabela 1.11: Teste de normalidade multivariada

A normalidade multivariada foi aceita nos três testes feitos.

Conclusão Normalidade:
Como foi apresentado até agora a normalidade foi aceita: Univariada, Bivariada e Trivariada ao nível de significância $\alpha = 0.01$ (mesmo não retirando os outliers encontrados anteriormente).

1.3 III) Teste de Hipótese

O objetivo imposto pelo problema é testar :

$$\begin{cases} H0 : \mu = (500, 50, 25) \\ H1 : \mu \neq (500, 50, 25) \end{cases}$$

Para isso, utilizei um teste apresentado em aula, rejeitamos H0, ao nivel $\alpha = 0.01$ de significância, se:

$$T^2 = n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0) > \frac{(n-1)p}{(n-p)} F_{p,(n-p),\alpha}$$

obs: Esse teste exige normalidade multivariada, então utilizei os dados que não contém o outlier encontrado (pois sem ele verifiquei normalidade multivariada).

Calculando T^2 no R, obtive:

$$T^2 = 20.6170$$

verificando a tabela F e calculando $\frac{(87-1)3}{(87-3)} F_{3,(87-3),0.01}$ também no R:

$$\frac{(87-1)^3}{(87-3)} F_{3,(87-3),0.01} = 12.3584$$

Ou seja:

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

Concluindo então que ao nível de significância $\alpha = 0.01$ rejeita-se a hipótese que $\mu = (500, 50, 25)$

Considerando o Teorema Central do Limite

Outro teste apresentado em aula e que tem a mesma finalidade de testar:

$$\begin{cases} H0 : \mu = \mu_0 \\ H1 : \mu \neq \mu_0 \end{cases}$$

É o teste considerando o Teorema Central do Limite, que tem como critério: Rejeita-se H0 se:

$$\Lambda = \left(\frac{|\sum (x_i - \bar{x})(x_i - \bar{x})'|}{|\sum (x_i - \mu_0)(x_i - \mu_0)'|} \right)^{n/2} < c_\alpha$$

O que é equivalente a:

$$\Lambda^{2/n} = \left(\frac{1}{1 + \frac{T^2}{n-1}} \right)^{-1}$$

Ou seja, rejeita-se H0 se $\Lambda^{2/n}$ menor que um valor muito baixo (c_α), sendo assim calculei:

$$\Lambda^{2/n} = 0.2369$$

Tendo então que $\Lambda^{2/n}$ é um valor bem baixo e menor que o valor critico ($\frac{(n-1)^p}{(n-p)} F_{p,(n-p),\alpha}$). Conclui-se então, pelo teste considerando o Teorema Central do Limite também rejeita-se H0.

1.3.1 Regiões de Confiança - RC

Em ambos os testes concluiu-se que não ha evidências que $\mu = (500, 50, 25)$, então quais seriam os possíveis valores para μ_0 talque $\mu = \mu_0$.

Para encontrarSS esses valores calculei as regiões de confiança da matriz de dados, a qual denotei como R(X):

$$R(X) = \left(\bar{x}_p - \sqrt{\frac{p(n-1)}{n-p} F_{p,n-p,\alpha}} \sqrt{\frac{S_{pp}}{n}} \leq \mu_p \leq \bar{x}_p + \sqrt{\frac{p(n-1)}{n-p} F_{p,n-p,\alpha}} \sqrt{\frac{S_{pp}}{n}} \right)$$

para todo p.

Sendo assim calculei a região de confiança para para todas as colunas e obtive o seguinte resultado:

$$\mu_1 = (497.8626; 555.3098)$$

$$\mu_2 = (50.45809; 58.92122)$$

$$\mu_3 = (23.31452; 26.93836)$$

Apenas para provar se os valores encontrados realmente seriam aceitos no teste, tomei $\mu_0 = (530, 55, 25)$ (valores dentro da região de confiança) e realizei o teste novamente, obtendo:

$$T^2 = 0.5686$$

O valor encontrado T^2 é menor do que a decisão já calculada anteriormente (12.3584), ou seja NÃO rejeita-se H0, tendo então que ao nível de significância $\alpha = 0.01$ existe evidencias que $\mu = (500, 55, 25)$

Apêndice A

Código

```r
# DADOS
df = read.csv2("E:/Trabalho_03/df_T3.csv", sep = ",", header = T)
df

df$X = NULL

head(df)
tail(df)

plot(df)

boxplot(df)
boxplot(df$Ciencia.Social.e.Historia..CLEP.)
boxplot((df$Verbal..CQT.))
boxplot(df$Ciencia..CQT.)

d = mahalanobis(df, colMeans(df), cov(df))
d

df$mahalanobis = d
df

d2 = mahalanobis(df, 0, cov(df))
d2

df$mahalanobis0 = d2
df

library(dplyr)
dfOrdem = df%>%
  arrange(desc(df$mahalanobis))

dfOrdem

# plot das distancia
```

```r
quant = qchisq(0.01, 3, lower.tail = F)
plot(1:length(d), d, xlab = "Observa  es",
     ylab= "Dist ncia de Mahalanobis",
     main = "Distancias")
abline(h=quant, lty=2)

plot(1:length(d2), d2, xlab = "Observa  es",
     ylab= "Dist ncia de Mahalanobis",
     main = "Distancias")
abline(h=quant, lty=2)


##########################################
############   Normalidade   #############
##########################################
# X1
qqnorm(df$Ciencia.Social.e.Historia..CLEP., main = "Normal Q-Q Plot (X1)
    ")
qqline(df$Ciencia.Social.e.Historia..CLEP., col="red")

library(nortest)
shapiro.test(df$Ciencia.Social.e.Historia..CLEP.)
lillie.test(df$Ciencia.Social.e.Historia..CLEP.)
ad.test(df$Ciencia.Social.e.Historia..CLEP.)

# X2
qqnorm(df$Verbal..CQT., main = "Normal Q-Q Plot (X2)")
qqline(df$Verbal..CQT., col="red")

hist(df$Verbal..CQT.)
shapiro.test(df$Verbal..CQT.)
lillie.test(df$Verbal..CQT.)
ad.test(df$Verbal..CQT.)

#X3
qqnorm(df$Ciencia..CQT., main = "Normal Q-Q Plot (X3)")
qqline(df$Ciencia..CQT., col="red")

shapiro.test(df$Ciencia..CQT.)
lillie.test(df$Ciencia..CQT.)
ad.test(df$Ciencia..CQT.)



#################################################################
############## Normalidade bivariada  ###########################
#################################################################
# install.packages("MVA")
# install.packages("royston")
```

```r
library(MVN)
########## A-B ###############
ab = data.frame(df$Ciencia.Social.e.Historia..CLEP., df$Verbal..CQT.)
ab

# Teste de Royston
library(royston)
mvn(ab, mvnTest = c("royston"), covariance = T, scale = F, desc = F,
    multivariatePlot = "qq")
royston.test(ab)

# Teste de Henze-Zirklers
mvn(ab, mvnTest = c("hz"), covariance = T, scale = F, desc = F,
    multivariatePlot = "qq")

# Teste de Mardia
mvn(ab, mvnTest = c("mardia"), covariance = T, scale = F, desc = F,
    multivariatePlot = "qq")



########### A-C ###############
ac = data.frame(df$Ciencia.Social.e.Historia..CLEP., df$Ciencia..CQT.)
ac

# Teste de Royston
mvn(ac, mvnTest = c("royston"), covariance = T, scale = F, desc = F,
    multivariatePlot = "qq")

# Teste de Henze-Zirklers
mvn(ac, mvnTest = c("hz"), covariance = T, scale = F, desc = F,
    multivariatePlot = "qq")

# Teste de Mardia
mvn(ac, mvnTest = c("mardia"), covariance = T, scale = F, desc = F,
    multivariatePlot = "qq")



######### B-C ################
bc = data.frame(df$Verbal..CQT., df$Ciencia..CQT.)
bc

# Teste de Royston
mvn(bc, mvnTest = c("royston"), covariance = T, scale = F, desc = F,
    multivariatePlot = "qq")

# Teste de Henze-Zirklers
mvn(bc, mvnTest = c("hz"), covariance = T, scale = F, desc = F,
```

```
    multivariatePlot = "qq")

# Teste de Mardia
mvn(bc, mvnTest = c("mardia"), covariance = T, scale = F, desc = F,
    multivariatePlot = "qq")



#######################################################
#####         TRIVARIADA    ###########################
#######################################################
tri = data.frame(df$Ciencia.Social.e.Historia..CLEP., df$Verbal..CQT.,
    df$Ciencia..CQT.)
tri

# Teste de Royston
mvn(tri, mvnTest = c("royston"), covariance = T, scale = F, desc = F,
    multivariatePlot = "qq")

# Teste de Henze-Zirklers
mvn(tri, mvnTest = c("hz"), covariance = T, scale = F, desc = F,
    multivariatePlot = "qq")

# Teste de Mardia
mvn(tri, mvnTest = c("mardia"), covariance = T, scale = F, desc = F,
    multivariatePlot = "qq")

#######################################################
#################   Teste de hipotese \mu ########
#######################################################
df3 = data.frame(df$Ciencia.Social.e.Historia..CLEP., df$Verbal..CQT.,
    df$Ciencia..CQT.)
df3

n = count(df3)
n = as.numeric(n)
p = length(df3)

mu = matrix(colMeans(df3))
mu
mu0 = matrix(c(500, 50, 25))
mu0

S = cov(df3)
T2 = n*t(mu - mu0) %*% solve(S) %*% (mu-mu0)
T2

# Tabela F
f = qf(0.01, df1 = 3, df2 = 84, lower.tail = FALSE) # tail = F, calda da
```

```
    direita
f

final = f*(n-1)*p/(n-p)
final


# Outro
lambda2 = (1/(1+(T2/n-1)))^{-1}
lambda2




#########################################################
#############   Regiao de confian a  ################
#########################################################

x1 = c(mean(df3$df.Ciencia.Social.e.Historia..CLEP.) - sqrt(p*(n-1)/(n-p
    )*f)*sqrt(S[1,1]/n), mean(df3$df.Ciencia.Social.e.Historia..CLEP.) +
    sqrt(p*(n-1)/(n-p)*f)*sqrt(S[1,1]/n))
x1

x2 = c(mean(df3$df.Verbal..CQT.) - sqrt(p*(n-1)/(n-p)*f)*sqrt(S[2,2]/n),
    mean(df3$df.Verbal..CQT.) + sqrt(p*(n-1)/(n-p)*f)*sqrt(S[2,2]/n))
x2

x3 = c(mean(df3$df.Ciencia..CQT.) - sqrt(p*(n-1)/(n-p)*f)*sqrt(S[3,3]/n)
    , mean(df3$df.Ciencia..CQT.) + sqrt(p*(n-1)/(n-p)*f)*sqrt(S[3,3]/n))
x3


# Testando novamente
mu0 = matrix(c(530, 55,25))
mu0
S = cov(df3)
T2 = n*t(mu - mu0) %*% solve(S) %*% (mu-mu0)
T2

# Tabela F
f = qf(0.01, df1 = p, df2 = (n-p), lower.tail = FALSE) # tail = F, calda
    da direita
f

final = f*(n-1)*p/(n-p)
final
```