Teste T-quadrado de Hotteling e Região de Confiança

Beatriz Lima e Vitória Sesana

Vitória, UFES

Julho de 2023

Teste de T-Quadrado de Hotelling

Teste de T-Quadrado de Hotelling

- Foi desenvolvido por Harold Hotteling, um estatístico e influente teórico econômico;
- É uma alternativa multivariada para o teste T de Student;
- Esse teste é utilizado em duas situações:

- **Uma amostra (One Sample)** : Verifica se um determinado vetor de médias tem valores plausiveis para a média da população de uma amostra multivariada. Possui a seguintes hipóteses:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

- **Duas amostras (Two Samples)**: Verifica se um determinado vetor tema valores plausiveis para a diferença de médias entre observações multivariadas de duas amostras. Possui a seguintes hipóteses:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

T-student e T-Quadrado de Hotelling

No contexto de análise univariada, para descobrir se um determinado valor μ_0 é um valor plausivel para a média de uma populaçao μ, cria-se o seguinte teste de hipótese:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Sendo $X_1, X_2, \ldots, X_n$ uma amostra aleatória de uma variável normal, o teste estatistico é:

$$t = \frac{(\overline{X} - \mu_0)}{s/\sqrt{n}}$$

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$$

T-student e T-Quadrado de Hotelling

Considerando em parametros multivariados, é preciso determinar se um dado vetor p x 1 μ_0 é um valor palusivel para o vetor de médias de uma distribuição normal multivariada. A generalização da tstudent para seu analógo multivariado é:

$$T^2 = (\overline{X} - \mu_0)´(\frac{1}{n}S)^{-1}(\overline{X} - \mu_0) = n(\overline{X} - \mu_0)´S^{-1}(\overline{X} - \mu_0)$$

$$\overline{X}_{p\times 1} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

$$S_{p\times p} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})(X_j - \overline{X})´$$

$$\mu_{0_{1\times p}} = \begin{bmatrix} \mu_{10} \\ \vdots \\ \mu_{p0} \end{bmatrix}$$

Distribuição de Fisher e do T-Quadrado de Hotelling

Se o valor T^2 observado é muito grande, então $\overline{x}$ é está muito "afastado" de μ_0 e a hipótese nula é rejeitada. Visto que T^2 é distribuida como $\frac{(n-1)p}{(n-p)}F_{p,n-p}$. Sendo $F_{p,n-p}$ uma variável aleatória com distribuição F com p e n-p graus de liberdade. Rejeita-se H_0 para nivel de significancia de α se

$$T^2 = n(\overline{X} - \mu_0)´S^{-1}(\overline{X} - \mu_0) > \frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)$$

Banco de Dados

```
knitr::kable(summary(iris), format = 'html')
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
	Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100
	1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300
	Median :5.800	Median :3.000	Median :4.350	Median :1.300
	Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199
	3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800
	Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500

Exemplo no R: Teste T-student

```
t.test(iris$Sepal.Length, mu= 6)
```

```
##
##     One Sample t-test
##
## data:  iris$Sepal.Length
## t = -2.3172, df = 149, p-value = 0.02186
## alternative hypothesis: true mean is not equal to 6
## 95 percent confidence interval:
##  5.709732 5.976934
## sample estimates:
## mean of x
##  5.843333
```

Exemplo no R: Teste T-student

```r
t.test(iris$Sepal.Width, mu= 3)
```

```
##
##     One Sample t-test
##
## data:  iris$Sepal.Width
## t = 1.611, df = 149, p-value = 0.1093
## alternative hypothesis: true mean is not equal to 3
## 95 percent confidence interval:
##  2.987010 3.127656
## sample estimates:
## mean of x
##  3.057333
```

Exemplo no R: Teste T-student

```
t.test(iris$Petal.Length, mu = 4)
```

```
##
##     One Sample t-test
##
## data:  iris$Petal.Length
## t = -1.679, df = 149, p-value = 0.09525
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  3.473185 4.042815
## sample estimates:
## mean of x
##     3.758
```

Exemplo no R: Teste T-student

```
t.test(iris$Petal.Width, mu = 1)
```

```
##
##     One Sample t-test
##
## data:  iris$Petal.Width
## t = 3.2028, df = 149, p-value = 0.001664
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
##  1.076353 1.322313
## sample estimates:
## mean of x
##  1.199333
```

Exemplo no R: Teste T-quadrado de Hotteling

```
n <- nrow(iris)
p <- ncol(iris)

vetor_medias <- colMeans(iris)
matriz_cov <- cov(iris)
matriz_inv_cov <- solve(matriz_cov)

medias_0 <- c(5,3,4,1)

T_2 <- n*t(vetor_medias - medias_0)%*%
        matriz_inv_cov%*%
        (vetor_medias - medias_0)

alpha <- 0.01
valorF <- qf(p = alpha,df1 = p,df2 = n-p, lower.tail = FALSE)

VC <- (((n-1)*p)/(n-p))*valorF
```

Resultados

Vetor de médias:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Médias	5.843333	3.057333	3.758	1.199333

Matriz de Covariância:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.6856935	-0.0424340	1.2743154	0.5162707
Sepal.Width	-0.0424340	0.1899794	-0.3296564	-0.1216394
Petal.Length	1.2743154	-0.3296564	3.1162779	1.2956094
Petal.Width	0.5162707	-0.1216394	1.2956094	0.5810063

Resultados

Matriz Inversa da Covariância:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	10.314699	-6.713189	-7.314483	5.739951
Sepal.Width	-6.713189	11.058417	6.480589	-6.170932
Petal.Length	-7.314483	6.480589	10.031679	-14.513766
Petal.Width	5.739951	-6.170932	-14.513766	27.693635

Testando o vetor de médias $\mu_0 = [5, 3, 4, 1]$. Dado n = 150 e p = 4, o valor critico de F para o teste é 3.45 e o T^2 é igual a 2160.86. Assim, rejeita-se H_0 com nível de significancia de 1%.

Exemplo no R: Uma Amostra

```
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.2.3
```

```
HotellingsT2Test(iris, mu = medias_0, test = "chi")
```

```
##
##     Hotelling's one sample T2-test
##
## data:  iris
## T.2 = 2160.9, df = 4, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to c(5,3,4,1)
```

Exemplo no R: Duas amostras

Selecionando as observações da especies Setosa e Versicolor

```
setosa <- iris[iris$Species == "setosa",]
setosa <- setosa[,-5]
knitr::kable(summary(setosa), format = 'html')
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
	Min. :4.300	Min. :2.300	Min. :1.000	Min. :0.100
	1st Qu.:4.800	1st Qu.:3.200	1st Qu.:1.400	1st Qu.:0.200
	Median :5.000	Median :3.400	Median :1.500	Median :0.200
	Mean :5.006	Mean :3.428	Mean :1.462	Mean :0.246
	3rd Qu.:5.200	3rd Qu.:3.675	3rd Qu.:1.575	3rd Qu.:0.300
	Max. :5.800	Max. :4.400	Max. :1.900	Max. :0.600

Exemplo no R: Duas amostras

Selecionando as observações da especies Setosa e Versicolor

```
versicolor <- iris[iris$Species == "versicolor",]
versicolor <- versicolor[,-5]
knitr::kable(summary(versicolor), format = 'html')
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
	Min. :4.900	Min. :2.000	Min. :3.00	Min. :1.000
	1st Qu.:5.600	1st Qu.:2.525	1st Qu.:4.00	1st Qu.:1.200
	Median :5.900	Median :2.800	Median :4.35	Median :1.300
	Mean :5.936	Mean :2.770	Mean :4.26	Mean :1.326
	3rd Qu.:6.300	3rd Qu.:3.000	3rd Qu.:4.60	3rd Qu.:1.500
	Max. :7.000	Max. :3.400	Max. :5.10	Max. :1.800

Exemplo no R: Duas amostras

Testando o vetor $\mu = [0.9, -0.65, 2.8, 1]$ para diferença de médias.

```
mu <- c(0.9, -0.65, 2.8, 1)

HotellingsT2Test(x=versicolor,y=setosa, mu=mu, test = "chi")
```

```
##
##     Hotelling's two sample T2-test
##
## data:  versicolor and setosa
## T.2 = 15.445, df = 4, p-value = 0.003861
## alternative hypothesis: true location difference is not equal to c(0.9,-0.
```

Região de Confiança para a Média

Intervalo de Confiança

- Caso univariado de tamanho n;
- Para μ e σ desconhecidos: teste t-student.

$$P\left[\left|\frac{\bar{X}-\mu}{S/\sqrt{n}}\right| \leq t_{n-1}(\alpha)\right] = 1-\alpha$$

$$P\left[\bar{X} - t_{n-1}(\alpha)\frac{S}{\sqrt{n}} < \mu < \bar{X} - t_{n-1}(\alpha)\frac{S}{\sqrt{n}}\right] = 1-\alpha$$

Portanto, o intervalo para μ com 100(1- α)% de confiança é:

$$IC = \left[\bar{X} - t_{n-1}(\alpha)\frac{S}{\sqrt{n}}; \bar{X} + t_{n-1}(\alpha)\frac{S}{\sqrt{n}}\right]$$

- Com $\bar{X}$ sendo a média amostral e S^2 a variância amostral.

Região de Confiança

- Caso Multivariado de tamanho n e com p variáveis;
- Teste de Hotteling.

$$P\left[R(X) \text{ cobrir o real valor de } \underset{\sim}{\mu}\right] = 1 - \alpha$$

$$P\left[n(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu) \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)\right] = 1 - \alpha$$

Como construir a região de confiança $R(X)$ com 100(1- α)% de confiança?

Calculando os eixos e seus tamanhos relativos de uma elipsoide (centrada em $\bar{X}$) a partir dos autovalores e autovetores da matriz de covariância S!

$$Se_i = \lambda_i e_i$$

- Com λ_i sendo os autovalores e e_i os autovetores.

Construindo a Região de Confiança

- Direção e metade dos tamanhos dos eixos do elipsoide de confiança:

$$\sqrt{\lambda_i}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}e_i$$

- Vértices da elipsoide de confiança:

$$\bar{X} \pm \sqrt{\lambda i}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}e_i$$

- Indicador de alogamento da elipsoide de confiança:

$$\frac{n\sqrt{\lambda_{max}}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}}{n\sqrt{\lambda_{min}}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}} = \frac{\sqrt{\lambda_{max}}}{\sqrt{\lambda_{min}}}$$

Região de Confiança Simultânea

- Outro método para obter uma região de confiança;
- Para p variáveis indepedentes, considera-se um intervalo de confiança para cada variável.
- Intervalos simultâneos ou intervalos T^2.

$$IC_i = \left[\bar{x}_i - \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{ii}}{n}}; \bar{x}_i + \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{ii}}{n}}; \right]$$

A Região de Confiança será a região que atende aos intervalos de cada variável simultâneamente!

O nível de confiança para cada intervalo será de $100(1-\alpha)\%$, mas simultaneamente será $100(1-\alpha)^p$.

Isso ocorre já que há p intervalos de confiança individuais.

$$P\left[\text{Todos os IC's cobrirem o real valor de } \underset{\sim}{\mu} \right] = (1-\alpha) \ldots (1-\alpha) = (1-\alpha)^p$$

Região de Confiança Simultânea (Método Bonferroni)

- Intervalos simultâneos mais precisos (menores) que os intervalos T^2.
- Desigualdade Bonferroni

$$P[\text{ Todos os IC's cobrirem o real valor de } \underset{\sim}{\mu}] = 1 - P[\text{Pelo menos um IC é falso}]$$

$$P[\text{ Todos os IC's serem verdadeiros}] = 1 - \sum_{i=n}^{m} P[\text{IC}_i \text{ ser falso}]$$

$$P[\text{ Todos os IC's serem verdadeiros}] = 1 - \sum_{i=n}^{m} (1 - P[\text{IC}_i \text{ ser verdadeiro}])$$

$$P[\text{Todos os IC's serem verdadeiros}] = 1 - (\alpha_1 + \cdots + \alpha_m)$$

Região de Confiança Simultânea (Método Bonferroni)

Desse modo, $\alpha_i = \alpha/m$

$$P\left[\bar{x}_i \pm t_{n-1}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{s_{ii}}{n}}\right] = 1 - \alpha/m$$

Os intervalos de confiança individuais utilizando o método bonferroni são:

$$IC_i = \left[\bar{x}_i - t_{n-1}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{s_{ii}}{n}}; \bar{x}_i + t_{n-1}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{s_{ii}}{n}}\right]$$

Exemplo no R

- Base 'iris', colunas: 'Sepal.Length' e 'Sepal.Width';

```
base_bivariada <- iris %>%
  select(Sepal.Length, Sepal.Width)

n <- nrow(base_bivariada)
p <- ncol(base_bivariada)

vetor_medias <- colMeans(base_bivariada)
matriz_cov <- cov(base_bivariada)
matriz_inv_cov <- solve(matriz_cov)

alpha <- 0.01
valorF <- qf(1-alpha, p, n - p)

autovalores <- eigen(matriz_cov)$values
autovetores <- eigen(matriz_cov)$vectors
```

Resultados

Vetor Médias:

	Sepal.Length	Sepal.Width
Médias	5.843333	3.057333

Matriz de Variâncias e Covariâncias:

	Sepal.Length	Sepal.Width
Sepal.Length	0.6856935	-0.0424340
Sepal.Width	-0.0424340	0.1899794

Resultados

Autovalores:

Autovalor 1	Autovalor 2
0.6892997	0.1863732

Autovetores:

Autovetor 1	Autovetor 2
-0.9964083	-0.0846783
0.0846783	-0.9964083

Dado n = 150, p = 2 o valor crítico de F para o teste de hotteling é 4.75 com 1% de nível de significância.

Calculando os eixos da elipse e seus tamanhos

```
eixos_tamanhos <- c()
eixos <- c()

for (i in 1:p) {
  tamanho <- sqrt(autovalores[i]) *
    sqrt( ( p * (n - 1) / (n * (n - p)) ) *
            valorF)

  eixos_valores <- tamanho * autovetores[,i]
  eixos_tamanhos <- cbind(eixos_tamanhos, tamanho)
  eixos <- cbind(eixos, eixos_valores)
}
```

Metade dos tamanhos dos eixos:

Tamanho 1	Tamanho 2
0.2096768	0.109028

Eixos:

Eixo 1	Eixo 2
-0.2089237	-0.0092323
0.0177551	-0.1086364

Calculando os vértices da elipse

```
vertices <- c()

for (i in 1:p) {
  vertice_inf <- vetor_medias - eixos[,i]
  vertice_sup <- vetor_medias + eixos[,i]
  vertices <- rbind(vertices, vertice_inf, vertice_sup)
}

colnames(vertices) <- colnames(base_bivariada)
```

Vértices da elipse:

Sepal.Length	Sepal.Width
6.052257	3.039578
5.634410	3.075088
5.852566	3.165970
5.834101	2.948697

Indicador de Achatamento da Região de Confiança

Razão entre os tamanhos dos eixos

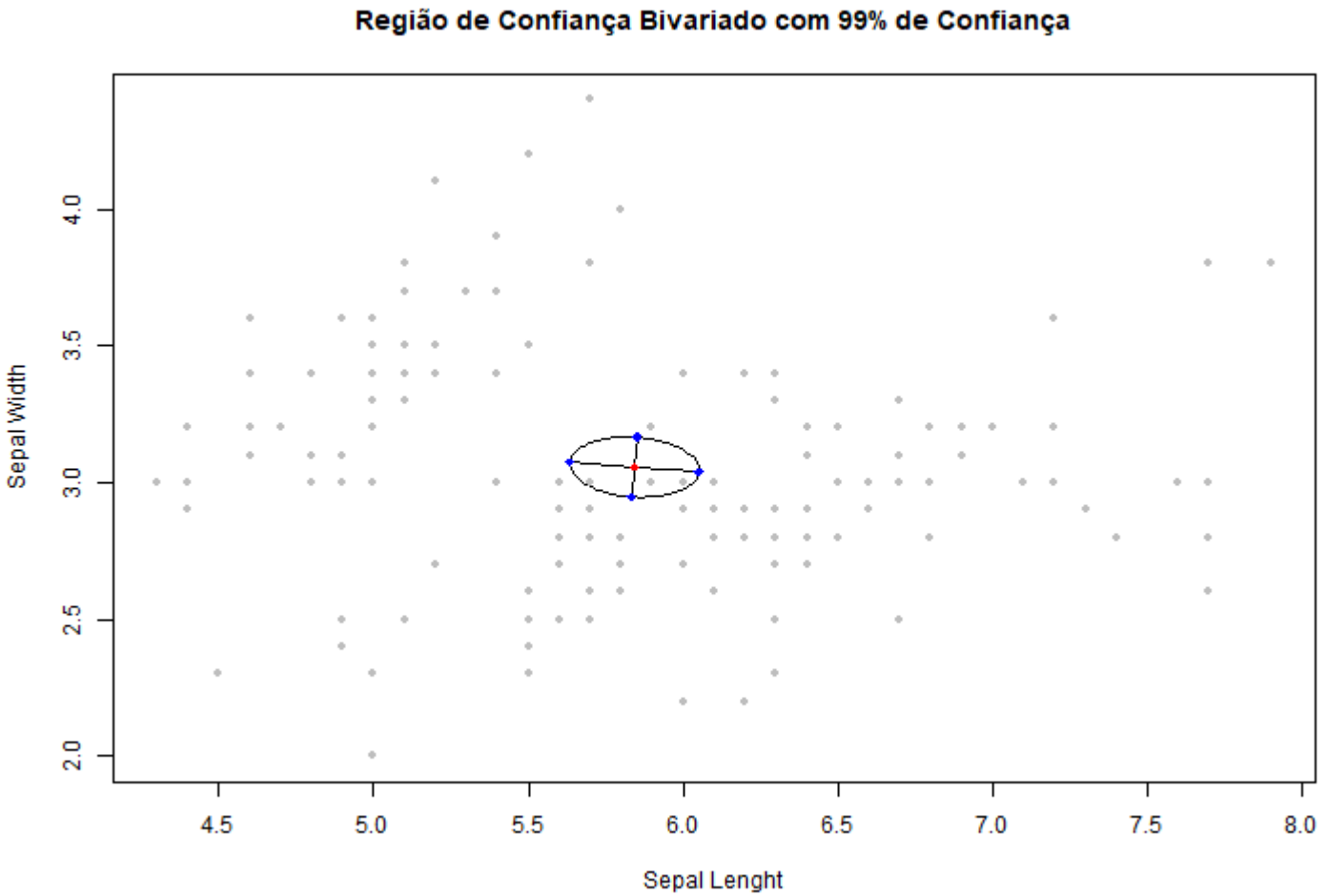
```
razao_tamanho_eixos <-
  sqrt(max(autovalores)) / sqrt(min(autovalores))
```

O tamanho do maior eixo é 1.9 vezes o tamanho do menor eixo.

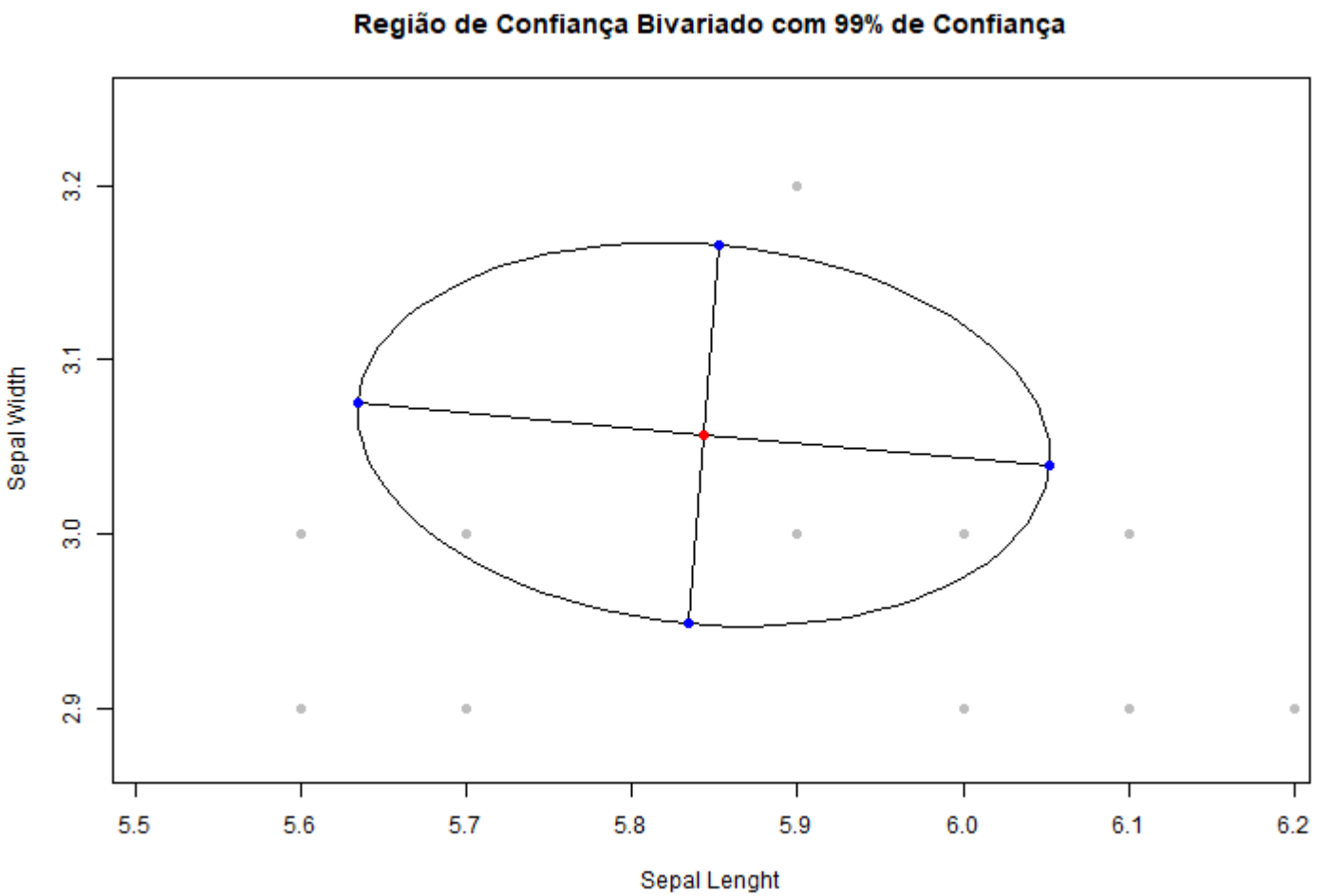
Plotando a Região de Confiança (Código)

```
library(MVQuickGraphs)

confidenceEllipse(X.mean = vetor_medias,
                  eig = eigen(matriz_cov),
                  n = n,
                  p = p,
                  alpha = alpha,
                  xl = c(min(base_bivariada$Sepal.Length),
                         max(base_bivariada$Sepal.Length)),
                  yl = c(min(base_bivariada$Sepal.Width),
                         max(base_bivariada$Sepal.Width)))

points(base_bivariada,
       pch = 20, col = "gray")

points(vetor_medias[1],
       vetor_medias[2],
       pch = 18, col = "red")

points(vertices[,1],
       vertices[,2],
       pch = 18, col = "blue")
```

Plotando a Região de Confiança (GRÁFICO)



Plotando a Região de Confiança (GRÁFICO AMPLIADO)



Calculando Intervalos Simultâneos (Teste de Hotteling)

```
intervalos <- c()
for (i in 1:p) {
  estatistica <-
    sqrt( ((p * (n-1)) / (n-p)) * valorF) *
    sqrt(matriz_cov[i,i]/n)
  lim_inf <- vetor_medias[i]-estatistica
  lim_sup <- vetor_medias[i]+estatistica
  limites <- c(lim_inf,lim_sup )
  intervalos <- rbind(intervalos,limites)
}

rownames(intervalos) <- rownames(matriz_cov)
colnames(intervalos) <- c("Limite Inferior",
                          "Limite Superior")
```

	Limite Inferior	Limite Superior
Sepal.Length	5.634206	6.052461
Sepal.Width	2.947256	3.167411

Calculando Intervalos Simultâneos (Bonferroni)

```
m <- p
valorT <- qt(1 - (alpha/(2*m)), df = n-1)

intervalos_bonferroni <- c()
for (i in 1:p) {
  estatistica <- valorT * sqrt(matriz_cov[i,i]/n)
  lim_inf <- vetor_medias[i] - estatistica
  lim_sup <- vetor_medias[i] + estatistica
  limites <- c(lim_inf, lim_sup)
  intervalos_bonferroni <- rbind(intervalos_bonferroni,limites)
}

rownames(intervalos_bonferroni) <- rownames(matriz_cov)
colnames(intervalos_bonferroni) <- c("Limite Inferior",
                                     "Limite Superior")
```

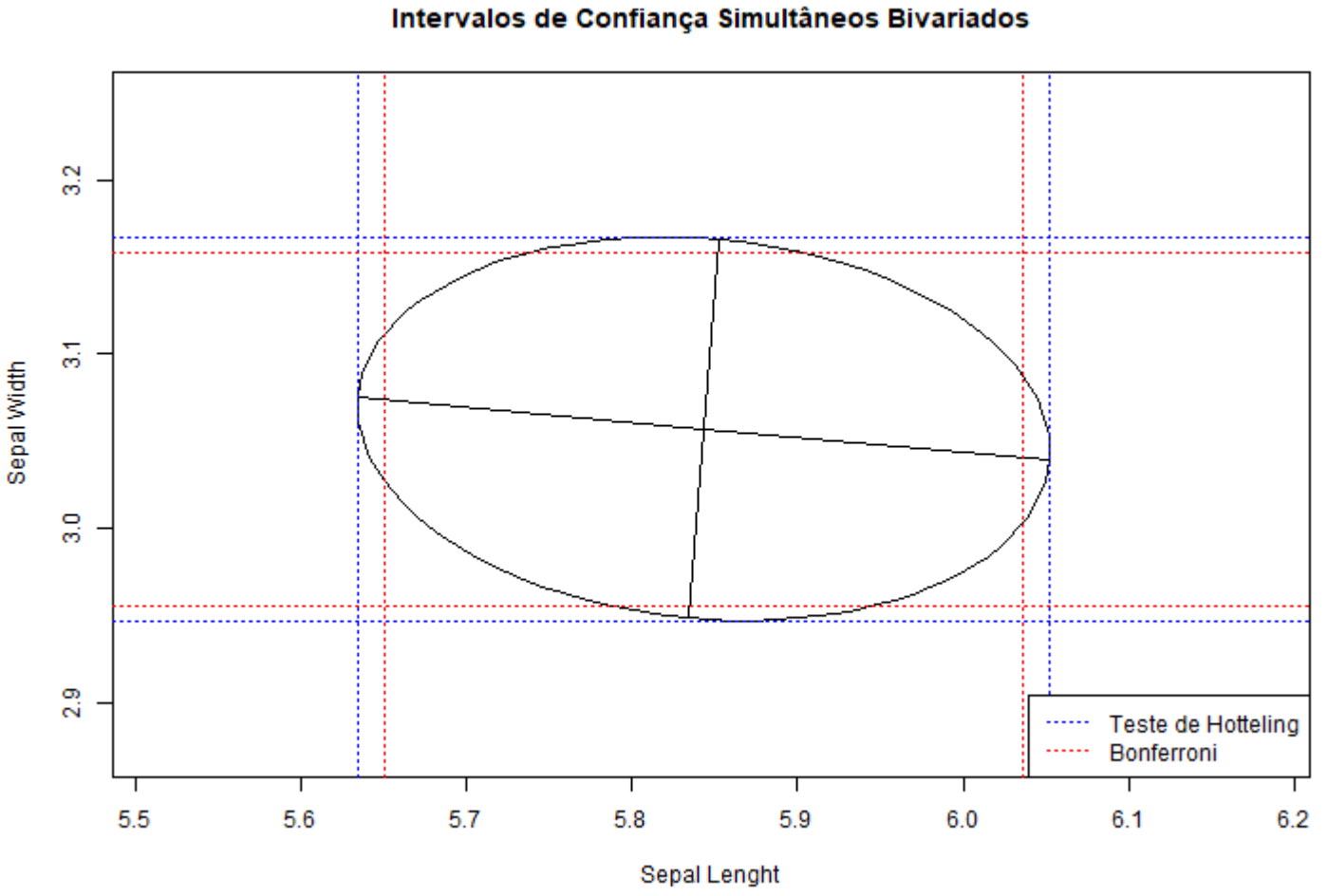
	Limite Inferior	Limite Superior
Sepal.Length	5.650679	6.035988
Sepal.Width	2.955926	3.158740

Plotando os Intervalos Simultâneos (Códigos)

```
confidenceEllipse(X.mean = vetor_medias,
                  eig = eigen(matriz_cov),
                  n = n,
                  p = p,
                  alpha = alpha)
title("Intervalos de Confiança Simultâneos com 99% de Confiança",
      xlab = "Sepal Lenght",
      ylab = "Sepal Width")

abline(v = intervalos[1,],
       h = intervalos[2,],
       lwd=1, lty=3, col = "blue")

abline(v = intervalos_bonferroni[1,],
       h = intervalos_bonferroni[2,],
       lwd=1, lty=3, col = "red")

legend("bottomright",
       legend = c("Teste de Hotteling", "Bonferroni"),
       lwd = c(1,1), lty=3, col = c("blue", "red"))
```

Plotando os Intervalos Simultâneos (Graficamente)

- Com 98.01% de confiança para $\alpha = 1\%$.



Agradecemos pela atenção!