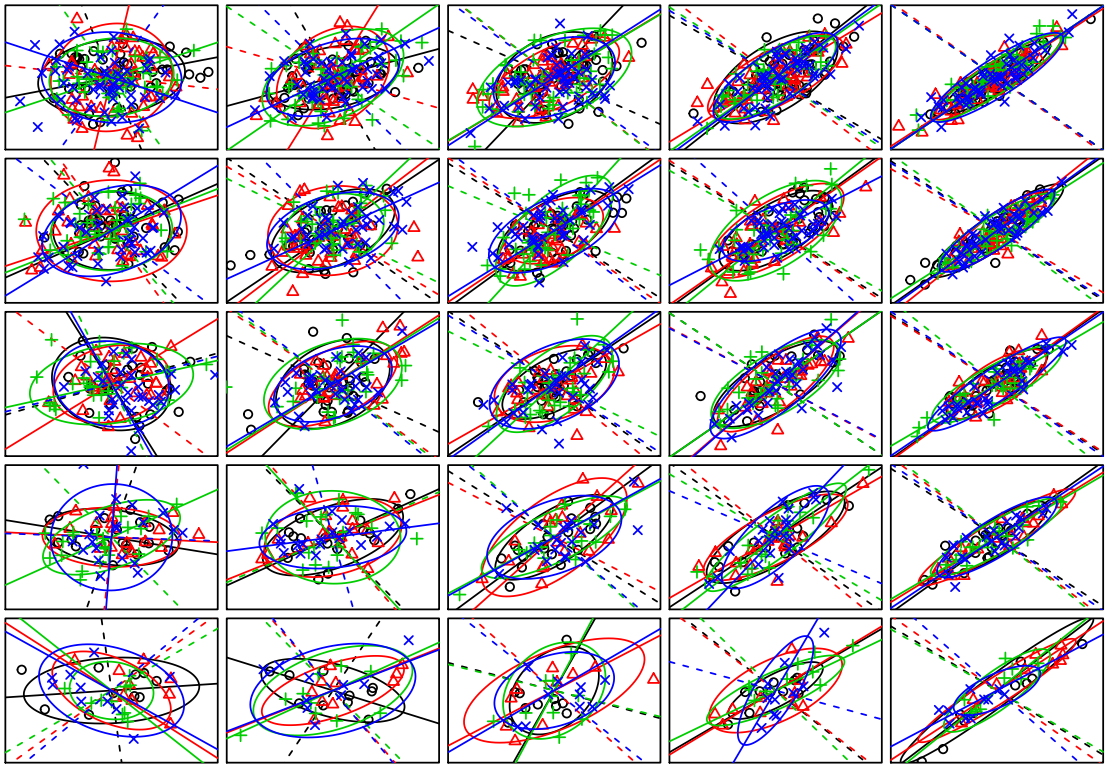
Estatística Multivariada

Slides de apoio às aulas



Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
2018/19

Aula 4

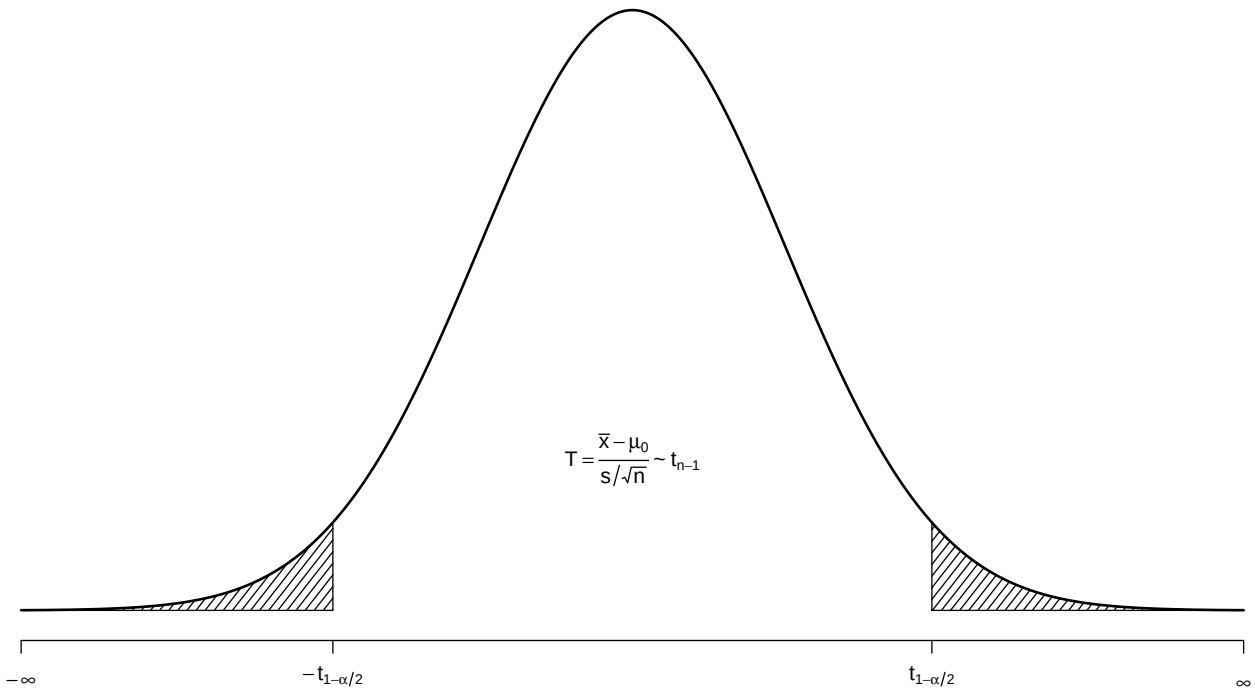
Inferência sobre um vetor médio

Inferência sobre μ (populacão normal univariada)

- Considerem-se as hipóteses estatísticas $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
- Sendo $x_1, ..., x_n$ uma amostra aleatória de uma população normal univariada de valor médio μ e variância σ^2 (sendo σ desconhecido), sabe-se que, sob H_0

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \frown t_{n-1}$$

sendo H_0 rejeitada se $|t| > t_{(n-1);1-\alpha/2}$



Exemplo 1

Nma amostra de 16 indivíduos com glaucoma de ângulo aberto registaram-se as idades dos pacientes:

62 62 68 48 51 60 51 57 57 41 62 50 53 34 62 61

Teste se existem evidências estatísticas de que a idade média da população a partir da qual a amostra foi retirada não difere significativamente de 60 anos ($\alpha = 0.01$).

Código R:

```
> x<-c(62,62,68,48,51,60,51,57,57,41,62,50,53,34,62,61)
> m0<-60
> m<-mean(x)
> m

[1] 54.9375

> s<-sd(x)
> s

[1] 8.872946

> t<-(m-m0)/(s/sqrt(length(x)))
> t

[1] -2.282218
```

- Com o objetivo de encontrar um teste com boas propriedades é usual restringir a escolha a testes que controlem a probabilidade de erro tipo I, α, a um nível especificado, usualmente, $\alpha = 0.01, 0.05, 0.10$

- Uma forma usual de apresentação do resultado de um teste de hipóteses baseia-se no **valor**-p (p-*value*), p

- p é uma **estatística** (ou seja, função da a.a.) tal que $p = p(\mathbf{x}) \in [0, 1]$, $\forall \mathbf{x} = (x_1, ..., x_n)$.

- Seja $T(\mathbf{x})$ é uma estatística apropriada ao teste , então, sob H_0 (teste bilateral), para o ponto amostral fixo $\mathbf{x} = \boldsymbol{x}$

$$p = 2P\big(T(\mathbf{x}) \geq |t(\boldsymbol{x})| \,\big|\, H_0\big)$$

  sendo $t(\boldsymbol{x})$ o valor observado para $T(\mathbf{x})$

- Ao nível de significância α deve rejeitar-se H_0 sse $p \leq \alpha$

Exemplo (continuação)

Código R:

```
> n<-length(x)
> pcum<-pt(abs(t),df=n-1)
> pvalue<-2*(1-pcum)
> pvalue

[1] 0.03748951
```

Inferência sobre $\boldsymbol{\mu}$ (populacão normal multivariada)

- Numa população multivariada, consideram-se agora as hipóteses estatísticas

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{bmatrix} \text{ vs. } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

- Note-se, numa população univariada normal (σ desconhecido), a rejeição de $H_0 : \mu = \mu_0$ para valores elevados de $|t|$ equivale à rejeição para valores elevados de t^2

$$t^2 = \frac{(\bar{x} - \mu)^2}{s^2/n} = n\,(\bar{x} - \mu)(s^2)^{-1}(\bar{x} - \mu)$$

sendo H_0 rejeitada se $t^2 > t^2_{(n-1);1-\alpha/2}$

- Sendo $\mathbf{x}$ um vetor aleatório, pode generalizar-se a expressão univariada de t^2, obtendo-se (com $\mathbf{x} \frown N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ e $\boldsymbol{\Sigma}$) desconhecida)

$$T^2 = (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \left(\frac{1}{n}\mathbf{S}\right)^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0) = n\,(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'(\mathbf{S})^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$$

(Nota: Para que $\mathbf{S}$ seja invertível, $n > p + 1$. Caso contrário, $\mathbf{S}$ terá determinante nulo.)

Inferência sobre μ (populacão normal multivariada)

- A estatística T^2 tem distribuição T^2 **de Hotelling**, com $n-1$ graus de liberdade, i.e., sob H_0

$$n\,(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'(\mathbf{S})^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \frown T_p^2(n-1)$$

sendo

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

- Assim, H_0 será rejeitada quando

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

onde $F_{(p,n-p);1-\alpha}$ representa o quantil de probabilidade $1-\alpha$ da distribuição $F_{(p,n-p)}$

Exemplo 2

Considere-se a amostra observada:

$$\boldsymbol{X} = \begin{bmatrix} 6 & 9 \\ 10 & 6 \\ 8 & 3 \end{bmatrix}$$

Iremos testar a hipótese

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 = \begin{bmatrix} 11 \\ 3 \end{bmatrix} \text{ vs. } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

1) Estatísticas amostrais (vetor de médias e matriz de covariâncias):

$$\bar{\boldsymbol{x}} = \begin{bmatrix} 8 \\ 6 \end{bmatrix} \qquad \boldsymbol{S} = \begin{bmatrix} 4 & -3 \\ -3 & 9 \end{bmatrix} \qquad \boldsymbol{S}^{-1} = \begin{bmatrix} 1/3 & 1/9 \\ 1/9 & 4/27 \end{bmatrix}$$

2) Cálculo de $T^2 = n\,(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'(\mathbf{S})^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$: $T^2_{\text{obs}} = 7$

3) Distribuição de T^2, sob H_0:

$$T^2 \overset{d}{=} \frac{2(3-1)}{3-2} F_{(2,1)} = 4F_{(2,1)}$$

4) Quantil de probabilidade 0.95 de $4F_{(2,1)}$: Para $\alpha = 0.05$, $F_{(2,1);0.95} = 199.5$, logo $T^2_{\text{obs}} << 4F_{(2,1);0.95}$.

5) Conclusão: Não se rejeita H_0, concluindo-se que não existem evidências estatísticas contra H_0

Exemplo 3

Considere-se a seguinte amostra bivariada de dimensão $n = 42$ (radiações emitidas por microondas, ficheiro "data3.xlsx"):

v_1	0.15	0.09	0.18	0.10	0.05	0.12	0.08	0.05	0.08	0.1	0.07	0.02	0.01	0.1
	0.1	0.1	0.02	0.1	0.01	0.4	0.1	0.05	0.03	0.05	0.15	0.1	0.15	0.09
	0.08	0.18	0.10	0.2	0.11	0.3	0.02	0.2	0.2	0.3	0.3	0.4	0.3	0.05
v_2	0.3	0.09	0.3	0.1	0.1	0.12	0.09	0.1	0.09	0.1	0.07	0.05	0.01	0.45
	0.12	0.2	0.04	0.1	0.01	0.6	0.12	0.1	0.05	0.05	0.15	0.3	0.15	0.09
	0.09	0.28	0.1	0.1	0.1	0.3	0.12	0.25	0.2	0.4	0.33	0.32	0.12	0.12

e as transformações $x_1 = v_1^{1/4}$ e $x_2 = v_2^{1/4}$, por forma a garantir o ajuste à distribuição normal bivariada. Teste as hipóteses

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 = \begin{bmatrix} 0.562 \\ 0.589 \end{bmatrix} \text{ vs. } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

Código R:

```
> dados3<-as.data.frame(readxl::read_xlsx("./Datasets/data3.xlsx",col_names = TRUE))
> x1<-dados3$v1^(1/4)
> x2<-dados3$v2^(1/4)
> X<-matrix(c(x1,x2),42,2)
> m<-colMeans(X)
> S<-var(X)
```

Exemplo (continuação)

$T^2_{\text{obs}} = 1.257$

Código R:

```
> n<-nrow(X)
> mu0<-c(0.562,0.589)
> t2<-n*t(c(m[1]-mu0[1],m[2]-mu0[2]))%*%solve(S)%*%(c(m[1]-0.562,m[2]-0.589))
> t2

       [,1]
[1,] 1.2573
```

Sob H_0, $T^2 \overset{d}{=} \frac{2(41)}{40} F_{(2,40)}$. Para $\alpha = 0.05$, $F_{(2,40);0.95} = 3.23$ e $2.05 F_{(2,40)} = 6.63$.

Código R:

```
> p<-ncol(X)
> F<-qf(p = 0.95,df1 = p,df2 = n-p)
> vc<-((p*(n-1))/((n-p)))*F
> vc

[1] 6.62504
```

Logo $T^2_{\text{obs}} < 6.63$. Não se rejeita H_0, concluindo-se que não existem evidências estatísticas contra H_0

Região de confiança para $\boldsymbol{\mu}$ (populacão normal multivariada)

- A região de confiança a $(1-\alpha)\times 100$ para o vetor $\boldsymbol{\mu}$ normal $p-$variado é dada por

$$n\,(\bar{\mathbf{x}}-\boldsymbol{\mu})'(\mathbf{S})^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \leq \frac{p(n-1)}{n-p}F_{(p,n-p);1-\alpha} \Leftrightarrow$$

$$\Leftrightarrow (\bar{\mathbf{x}}-\boldsymbol{\mu})'(\mathbf{S})^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \leq \frac{p(n-1)}{n(n-p)}F_{(p,n-p);1-\alpha}$$

  tal que

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

- Face a um conjunto de n observações multivariadas $\boldsymbol{x}_1,...,\boldsymbol{x}_n$ a região de confiança para $\boldsymbol{\mu}$ é dada por

$$(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'(S)^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) \leq \frac{p(n-1)}{n(n-p)}F_{(p,n-p);1-\alpha}$$

- A região de confiança elipsóide tem centro $\bar{\boldsymbol{x}}$ e eixos

$$\pm\sqrt{\ell_j}\sqrt{\frac{p(n-1)}{n(n-p)}F_{(p,n-p);1-\alpha}}\mathbf{e}_j$$

  sendo $\ell_j\,(j=1,...,p)$ os valores próprios e $\mathbf{e}_j\,(j=1,...,p)$ os vetores próprios de $\boldsymbol{S}$.

Exemplo (continuação)

Considerando o exemplo anterior, defina a região de confiança a 95% para o vetor (μ_1, μ_2) (esboce graficamente a região de confiança).

A região de confiança a 95% corresponde a todos os pontos da superfície definida por

$$[0.564-\mu_1, 0.603-\mu_2]\begin{bmatrix}203.498 & -163.907\\ -163.907 & 200.769\end{bmatrix}\begin{bmatrix}0.564-\mu_1\\ 0.603-\mu_2\end{bmatrix} \leq \frac{2(41)}{42(40)} \times 3.23 \Leftrightarrow$$

$$\Leftrightarrow [0.564-\mu_1, 0.603-\mu_2]\begin{bmatrix}203.498 & -163.907\\ -163.907 & 200.769\end{bmatrix}\begin{bmatrix}0.564-\mu_1\\ 0.603-\mu_2\end{bmatrix} \leq 0.158$$

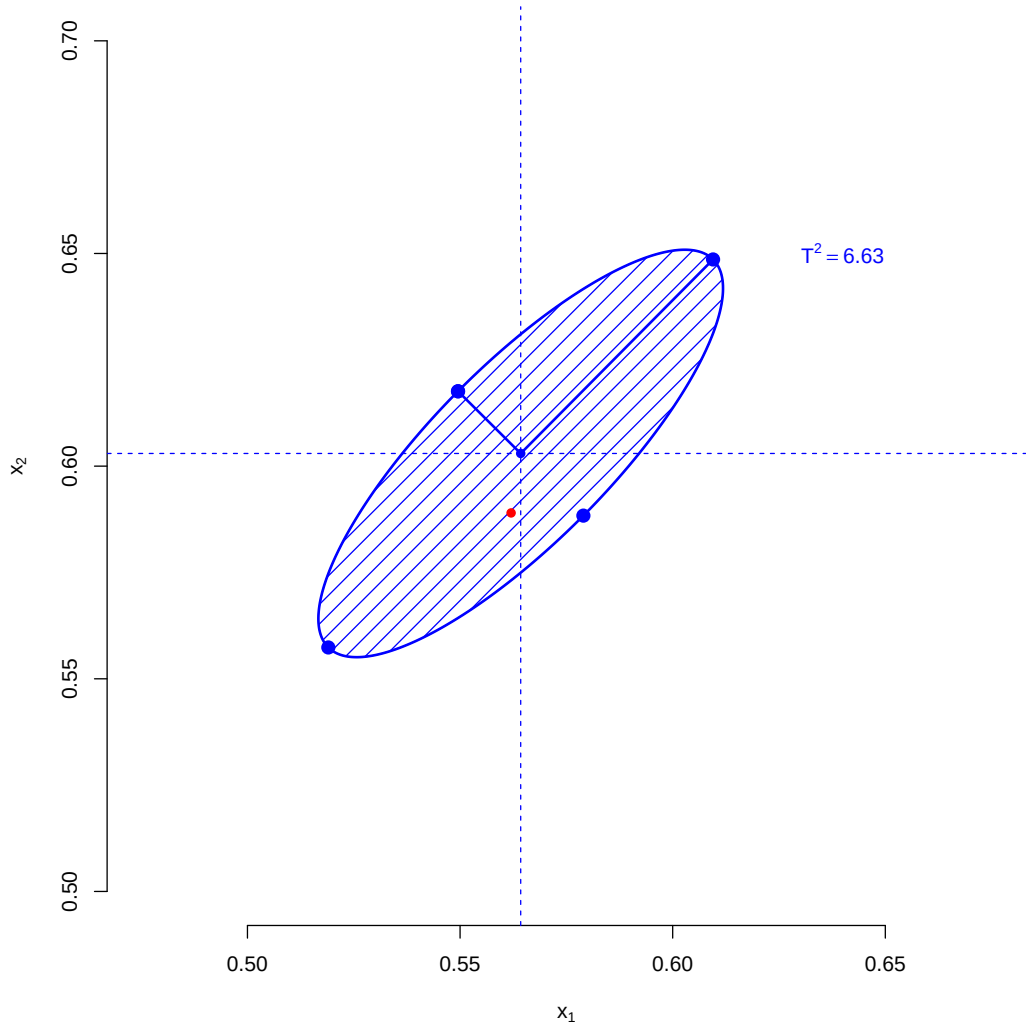
A região de confiança elíptica tem eixos definidos por

$$\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\end{bmatrix} \pm \sqrt{\ell_1}\sqrt{\frac{p(n-1)}{n(n-p)}F_{(p,n-p);1-\alpha}}\mathbf{e}_1 = \sqrt{0.026}\sqrt{0.158}\begin{bmatrix}0.704\\ 0.710\end{bmatrix} = \begin{bmatrix}0.609\\ 0.649\end{bmatrix}, \begin{bmatrix}0.519\\ 0.557\end{bmatrix}$$

$$\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\end{bmatrix} \pm \sqrt{\ell_2}\sqrt{\frac{p(n-1)}{n(n-p)}F_{(p,n-p);1-\alpha}}\mathbf{e}_2 = \sqrt{0.00273}\sqrt{0.158}\begin{bmatrix}-0.710\\ 0.704\end{bmatrix} = \begin{bmatrix}0.549\\ 0.618\end{bmatrix}, \begin{bmatrix}0.578\\ 0.588\end{bmatrix}$$

Note-se que o ponto $\boldsymbol{\mu}_0' = (0.562, 0.589)$ está dentro da região definida pela elipse de confiança

Exemplo (continuação)



Exemplo 4

O ficheiro "data4.xlsx" contem medições do grau de rigidez e resistência à flexão de 30 toros de madeira.

a) Construa e desenhe a elipse de confiança a 99% para $\boldsymbol{\mu}$

b) Suponha que os valores 2000 (rigidez) e 10000 (resistência) são aceites como os valores médios característicos das variáveis em estudo. Teste se, com base na amostra obtida, é possível corroborar a afirmação (calcule o *p-value*).

c) Estudo o ajustamento à distribuição normal bivariada.

Exemplo (continuação)

Código R:

```
> dados4<-as.data.frame(readxl::read_xlsx("./Datasets/data4.xlsx",col_names = TRUE))
> m<-colMeans(dados4); S<-var(dados4); eigen<-eigen(S)
> p<-ncol(dados4); n<-nrow(dados4)
> alpha<-0.01
> #Maior eixo
> m+sqrt(eigen$values[1])*
   sqrt(((p*(n-1))/(n*(n-p)))*qf(1-alpha,df1 = p,df2 = n-p))*eigen$vectors[,1]

      x1       x2
1982.322 9499.446

> m-sqrt(eigen$values[1])*
   sqrt(((p*(n-1))/(n*(n-p)))*qf(1-alpha,df1 = p,df2 = n-p))*eigen$vectors[,1]

      x1       x2
1738.678 7208.220

> #Menor eixo
> m+sqrt(eigen$values[2])*
   sqrt(((p*(n-1))/(n*(n-p)))*qf(1-alpha,df1 = p,df2 = n-p))*eigen$vectors[,2]

      x1       x2
1681.985 8372.816

> m-sqrt(eigen$values[2])*
   sqrt(((p*(n-1))/(n*(n-p)))*qf(1-alpha,df1 = p,df2 = n-p))*eigen$vectors[,2]

      x1       x2
2039.015 8334.850
```

Exemplo (continuação)

Código R:

```
> m0<-c(2000,10000)
> T2<-n*t(m-m0)%*%solve(S)%*%(m-m0)
> T2

         [,1]
[1,] 23.65452

> F<-qf(1-alpha,df1 = p,df2 = n-p)
> vc<-((p*(n-1))/(n-p))*F
> vc

[1] 11.29537

> ifelse(T2>vc,"Rejeita-se H0","Não se rejeita H0")

     [,1]
[1,] "Rejeita-se H0"

> p<-1-pf(((n-p)/(p*(n-1)))*T2,df1=p,df2=n-p)
> p

              [,1]
[1,] 0.0002363051
```

Intervalos de confiança simultâneos

- Do ponto de vista prático, em regra, em estudos multivariados é necessário e útil a construção de IC para cada um dos valores médios do vetor $\boldsymbol{\mu}$.

- A determinação destes IC implica o entendimento de que estes se verificam *simultaneamente* para uma dada probabilidade de confiança ("simultânea")

- Seja $\mathbf{x} \frown N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ e considere-se a combinação linear

$$z = a_1 x_1 + ... + a_p x_p = \mathbf{a}'\mathbf{x}$$

com $\mu_z = \mathbf{a}'\boldsymbol{\mu}$ e $\sigma_z^2 = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$, i.e., $z \frown N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$

- Considerando a a.a. $\mathbf{x}_1, ..., \mathbf{x}_n$, $\bar{z} = \mathbf{a}'\bar{\mathbf{x}}$ e $s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$, e, simultaneamente para todos os valores de $\mathbf{a}$, o intervalo

$$\left(\mathbf{a}'\bar{\mathbf{x}} - \sqrt{\frac{p(n-1)}{n(n-p)} F_{(p,n-p);1-\alpha} \mathbf{a}'\mathbf{S}\mathbf{a}}; \mathbf{a}'\bar{\mathbf{x}} + \sqrt{\frac{p(n-1)}{n(n-p)} F_{(p,n-p);1-\alpha} \mathbf{a}'\mathbf{S}\mathbf{a}} \right)$$

contem $\mathbf{a}'\boldsymbol{\mu}$ com probabilidade $1 - \alpha$.

Intervalos de confiança simultâneos

- As sucessivas escolhas $\mathbf{a}' = (1, 0, ..., 0)$, $\mathbf{a}' = (0, 1, ..., 0)$,..., $\mathbf{a}' = (0, 0, ..., 1)$ permitem obter, respetivamente, os sucessivos intervalos para os p valores médios μ_1, μ_2,..., μ_p

$$\left(\bar{x}_1 - \sqrt{\frac{p(n-1)}{(n-p)} F_{(p,n-p);1-\alpha} \times \frac{s_{11}}{n}}; \bar{x}_1 + \sqrt{\frac{p(n-1)}{(n-p)} F_{(p,n-p);1-\alpha} \times \frac{s_{11}}{n}}\right)$$

$$\left(\bar{x}_2 - \sqrt{\frac{p(n-1)}{(n-p)} F_{(p,n-p);1-\alpha} \times \frac{s_{22}}{n}}; \bar{x}_2 + \sqrt{\frac{p(n-1)}{(n-p)} F_{(p,n-p);1-\alpha} \times \frac{s_{22}}{n}}\right)$$

...

$$\left(\bar{x}_p - \sqrt{\frac{p(n-1)}{(n-p)} F_{(p,n-p);1-\alpha} \times \frac{s_{pp}}{n}}; \bar{x}_p + \sqrt{\frac{p(n-1)}{(n-p)} F_{(p,n-p);1-\alpha} \times \frac{s_{pp}}{n}}\right)$$

Estes intervalos são designados por *intervalos de confiança simultâneos* a $(1-\alpha) \times 100\%$ ou T^2-*intervalos*

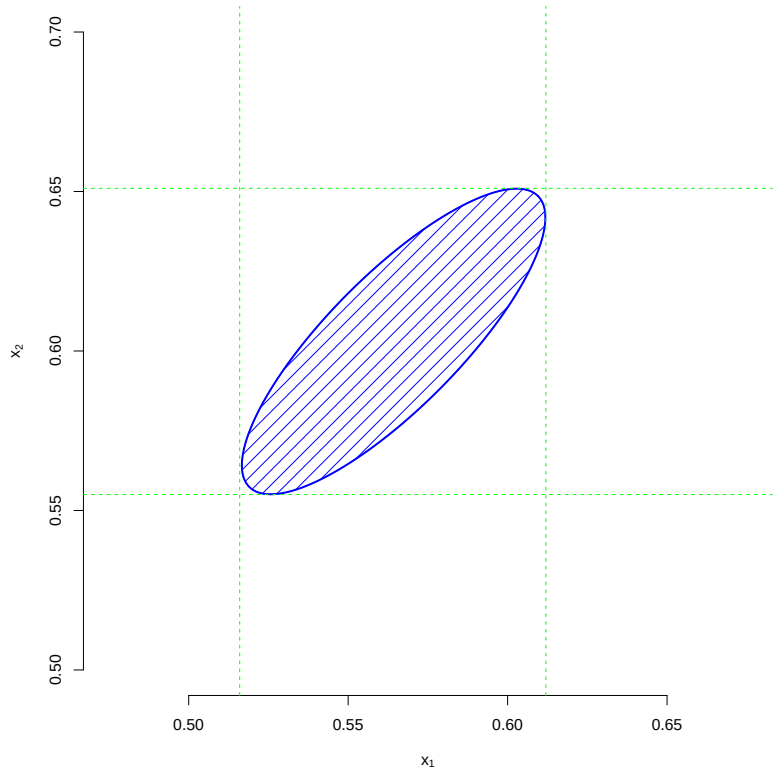
- Estes intervalos correspondem às projeções ("sombras") da região de confiança elipsóide sobre os eixos.

Exemplo 5

Considerando os dados do exemplo anterior, defina os IC simultâneos a 95% para o vetor $\boldsymbol{\mu}' = (\mu_1, \mu_2)$.

$$\left(\bar{x}_1 - \sqrt{\frac{2(41)}{40} F_{(2,40);0.95} \times \frac{s_{11}}{n}}; \bar{x}_1 + \sqrt{\frac{2(41)}{40} F_{(2,40);0.95} \times \frac{s_{11}}{n}}\right) = (0.516, 0.612)$$

$$\left(\bar{x}_2 - \sqrt{\frac{2(41)}{40} F_{(2,40);0.95} \times \frac{s_{22}}{n}}; \bar{x}_2 + \sqrt{\frac{2(41)}{40} F_{(2,40);0.95} \times \frac{s_{22}}{n}}\right) = (0.555, 0.651)$$



One-at-a-time intervalos de confiança

- Uma abordagem alternativa consiste em considerar os p intervalos univariados para os valores médios μ_1, μ_2,..., μ_p

$$\left(\bar{x}_1 - t_{(n-1);1-\alpha/2}\sqrt{\times\frac{s_{11}}{n}}; \bar{x}_1 + t_{(n-1);1-\alpha/2}\sqrt{\frac{s_{11}}{n}}\right)$$

$$\left(\bar{x}_2 - t_{(n-1);1-\alpha/2}\sqrt{\frac{s_{22}}{n}}; \bar{x}_2 + t_{(n-1);1-\alpha/2}\sqrt{\frac{s_{22}}{n}}\right)$$

$$\ldots$$

$$\left(\bar{x}_p - t_{(n-1);1-\alpha/2}\sqrt{\frac{s_{pp}}{n}}; \bar{x}_p + t_{(n-1);1-\alpha/2}\sqrt{\frac{s_{pp}}{n}}\right)$$

- Estes intervalos ignoram a estrutura de covariância das p covariáveis;
- Por outro lado, a confiança (probabilidade conjunta) associada a todas as estimativas intervalares não é $1-\alpha$, mas sim

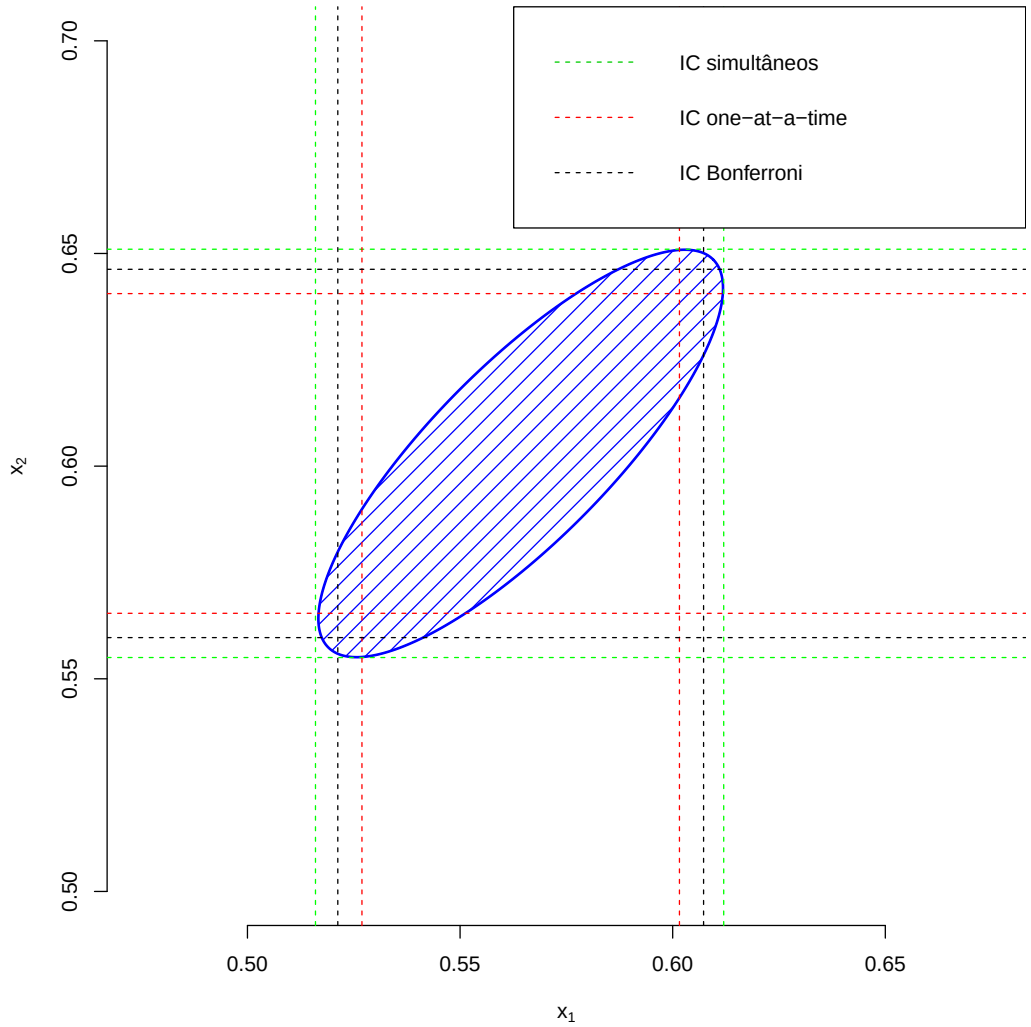
$$(1-\alpha)^p$$

- Para corrigir esta situação pode fazer-se a *correção de Bonferroni* que consiste em considerar uma probabilidade de erro α/p (confiança $= 1-\alpha/p$) em cada intervalo:

$$\bar{x}_i \pm t_{(n-1);(1-(\alpha/p)/2)}\sqrt{\frac{s_i^2}{n}}\ (i = 1, ..., p)$$

assegurando uma confiança global não inferior a $1-\alpha$

Exemplo (continuação)



Inferência sobre $\boldsymbol{\mu}$ (grandes amostras)

- Vimos anteriormente que para n suficientemente grande ($n >> p$)

$$n\,(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{d} \chi^2_p$$

- Com base neste resultado, ao nível de significância α, rejeita-se $H_0 = \boldsymbol{\mu} = \boldsymbol{\mu}_0$ contra $H_0 = \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ quando o valor observado para

$$n\,(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) > \chi^2_{p;(1-\alpha)}$$

- Os IC simultâneos assintóticos são definidos por

$$\left(\mathbf{a}'\bar{\mathbf{x}} - \sqrt{\chi^2_{p;(1-\alpha)} \times \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}\,;\, \mathbf{a}'\bar{\mathbf{x}} + \sqrt{\chi^2_{p;(1-\alpha)} \times \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}\right)$$

sendo os sucessivos intervalos simultâneos assintóticos para μ_1, μ_2,..., μ_p a $(1-\alpha)\times 100\%$, respetivamente, dados por

$$\left(\bar{x}_1 - \sqrt{\chi^2_{p;(1-\alpha)} \times \frac{s_{11}}{n}}\,;\, \bar{x}_1 + \sqrt{\chi^2_{p;(1-\alpha)} \times \frac{s_{11}}{n}}\right)$$

...

$$\left(\bar{x}_p - \sqrt{\chi^2_{p;(1-\alpha)} \times \frac{s_{pp}}{n}}\,;\, \bar{x}_p + \sqrt{\chi^2_{p;(1-\alpha)} \times \frac{s_{pp}}{n}}\right)$$

Inferência sobre μ (grandes amostras)

- Também neste caso se podem obter os intervalos univariados por aproximação à distribuição normal

$$\left(\bar{x}_1 - z_{1-\alpha/2}\sqrt{\frac{s_{11}}{n}}; \bar{x}_1 + z_{1-\alpha/2}\sqrt{\frac{s_{11}}{n}}\right)$$

$$\ldots$$

$$\left(\bar{x}_p - z_{1-\alpha/2}\sqrt{\frac{s_{pp}}{n}}; \bar{x}_p + z_{1-\alpha/2}\sqrt{\frac{s_{pp}}{n}}\right)$$

- Com *correcção de Bonferroni*:

$$\left(\bar{x}_1 - z_{1-\alpha/(2p)}\sqrt{\frac{s_{11}}{n}}; \bar{x}_1 + z_{1-\alpha/(2p)}\sqrt{\frac{s_{11}}{n}}\right)$$

$$\ldots$$

$$\left(\bar{x}_p - z_{1-\alpha/(2p)}\sqrt{\frac{s_{pp}}{n}}; \bar{x}_p + z_{1-\alpha/(2p)}\sqrt{\frac{s_{pp}}{n}}\right)$$

Inferência sobre dois vetores médios

Comparação de 2 valores médios (populacão normal univariada, amostras emparelhadas)

- Em muitos estudos é usual proceder-se à medição (repetida) das variáveis em análise nas mesmas unidades estatísticas, em diferentes tempos $\rightarrow$ *amostras emparelhadas*
- Considere-se o caso univariado para a situação de 2 amostras emparelhadas. Se x representa a variável em estudo (população normal univariada), então x_{i1} e x_{i2} $(i = 1, ..., n)$ representam as amostras emparelhadas a partir das quais é possível calcular

$$d_i = x_{i1} - x_{i2} \, (i = 1, ..., n)$$

sendo d_i variáveis iid $N(\mu_d, \sigma_d^2)$.

- Neste caso, sob $H_0 : \mu_d = 0$

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \frown t_{n-1}$$

- Assim, H_0 será rejeitada quando $|t| > t_{1-\alpha/2;(n-1)}$
- Intervalo de confiança para μ a $(1 - \alpha) \times 100\%$

$$\left(\bar{d} - t_{(n-1);1-\alpha/2} \times \frac{s_d}{\sqrt{n}}; \bar{d} + t_{(n-1);1-\alpha/2} \times \frac{s_d}{\sqrt{n}}\right)$$

Comparação de 2 vetores médios (populacão normal multivariada, amostras emparelhadas)

- Considerem-se agora as hipóteses estatísticas

$$H_0 : \boldsymbol{\mu}_d = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ vs. } H_1 : \boldsymbol{\mu}_d \neq \mathbf{0}$$

- Sendo $\mathbf{d}$ um vetor aleatório tal que $\mathbf{d} \frown N_p(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$, pode generalizar-se a expressão univariada de t^2, obtendo-se, sob H_0

$$T^2 = n\,(\bar{\mathbf{d}} - \boldsymbol{\mu}_{d0})'(\mathbf{S}_d)^{-1}(\bar{\mathbf{d}} - \boldsymbol{\mu}_{d0}) = n\,\bar{\mathbf{d}}'(\mathbf{S}_d)^{-1}\bar{\mathbf{d}} \frown T_p^2(n-1)$$

onde

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

- Assim, H_0 será rejeitada quando

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

onde $F_{(p,n-p);1-\alpha}$ representa o quantil de probabilidade $1-\alpha$ da distribuição $F_{(p,n-p)}$

Região de confiança para a diferença de 2 vetores médios (populacão normal multivariada, amostras emparelhadas)

- A região de confiança a $(1-\alpha)\times 100$ para o vetor $\boldsymbol{\mu}_d$ de uma população normal $p-$variada é dada por

$$(\bar{\mathbf{d}}-\boldsymbol{\mu}_d)'\mathbf{S}^{-1}(\bar{\mathbf{d}}-\boldsymbol{\mu}_d) \leq \frac{p(n-1)}{n(n-p)}F_{(p,n-p);1-\alpha}$$

- Os p IC simultâneos para μ_{d_j} $(j=1,...,p)$ são dados por a $(1-\alpha)\times 100\%$

$$\bar{d}_j \pm \sqrt{\frac{p(n-1)}{n(n-p)}F_{(p,n-p);1-\alpha}}\sqrt{\frac{s_{d_j}^2}{n}}$$

sendo $\bar{d}_j$ o j-ésimo elemento do vetor $\bar{\mathbf{d}}$ e $s_{d_j}^2$ o i-ésimo elemento da diagonal da matriz $\mathbf{S}_d$

- Os p IC simultâneos de Bonferroni para μ_{d_j} $(j=1,...,p)$ são dados por a $(1-\alpha)\times 100\%$

$$\bar{d}_j \pm t_{(n-1);1-\alpha/(2p)}\sqrt{\frac{s_{d_j}^2}{n}}$$

Exemplo 6

A análise de 11 amostras de águas residuais em dois laboratórios distintos revelou as seguintes medições de duas componentes orgânicas x_1 e x_2:

	Lab. 1		Lab. 2	
	x_1	x_2	x_1	x_2
1	6	27	25	15
2	6	23	28	13
3	18	64	36	22
4	8	44	35	29
5	11	30	15	31
6	34	75	44	64
7	28	26	42	30
8	71	124	54	64
9	43	54	34	56
10	33	30	29	20
11	20	14	39	21

Serão os resultados médios laboratoriais significativamente diferentes? Calcule os IC simultâneos a 95% de confiança.

Código R:

```
> dx1
 [1] -19 -22 -18 -27  -4 -10 -14  17   9   4 -19
> dx2
 [1] 12 10 42 15 -1 11 -4 60 -2 10 -7
```

Exemplo 7

Estatísticas amostrais:

Código R:

```
> d<-matrix(c(dx1,dx2),11,2)
> md<-colMeans(d)
> md

[1] -9.363636 13.272727

> Sd<-var(d)
> Sd

          [,1]      [,2]
[1,] 199.25455  88.30909
[2,]  88.30909 418.61818
```

Estatística do teste:

Código R:

```
> n<-nrow(d)
> T2<-n*t(md)%*%solve(Sd)%*%md
> T2

         [,1]
[1,] 13.63931
```

Exemplo (continuação)

Valor crítico:

Código R:

```
> p<-ncol(d)
> vc<-(p*(n-1))/(n-p)*qf(0.95,p,n-p)
> vc
[1] 9.458877
```

Logo, rejeita-se H_0 existindo evidências de diferenças significativas entre os resultados médios dos dois laboratórios.

IC simultâneos:

Código R:

```
> md[1]-sqrt((p*(n-1))/(n*(n-p))*qf(0.95,p,n-p))*sqrt(Sd[1,1]/n)
[1] -13.31031
> md[1]+sqrt((p*(n-1))/(n*(n-p))*qf(0.95,p,n-p))*sqrt(Sd[1,1]/n)
[1] -5.416963
> md[2]-sqrt((p*(n-1))/(n*(n-p))*qf(0.95,p,n-p))*sqrt(Sd[2,2]/n)
[1] 7.552199
> md[2]+sqrt((p*(n-1))/(n*(n-p))*qf(0.95,p,n-p))*sqrt(Sd[2,2]/n)
[1] 18.99326
```

Comparação de 2 vetores médios (populacão normal multivariada, amostras independentes)

- Sejam
  - $(\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1})$ a a.a. 1 de dimensão n_1 extraída de uma população $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
  - $(\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2})$ a a.a. 2 de dimensão n_2 extraída de uma população $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$
  - $(\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1})$ e $(\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2})$ amostras independentes e $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$
- Considerem-se agora as hipóteses estatísticas

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ vs. } H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \Leftrightarrow H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0} \text{ vs. } H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$$

- Sob H_0

$$T^2 = [(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{pooled} \right]^{-1} [(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \frown T_p^2(n_1 + n_2 - 2)$$

onde

$$\mathbf{S}_{pooled} = \frac{n_1 - 1}{n_1 + n_2 - 2} \mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} \mathbf{S}_2$$

e

$$T^2 \overset{d}{=} \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{(p, n_1 + n_2 - p - 1)}$$

- Assim, H_0 será rejeitada quando $T^2 > \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{(p,n_1+n_2-p-1);1-\alpha}$ onde $F_{(p,n_1+n_2-p-1);1-\alpha}$ representa o quantil de probabilidade $1 - \alpha$ da distribuição $F_{(p,n_1+n_2-p-1)}$

Região de confiança para a diferença de 2 vetores médios (populacão normal multivariada, amostras independentes)

- A região de confiança elíptica a $(1-\alpha)\times 100$ para a diferença $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ de uma população normal $p-$variada tem eixos

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \pm c\sqrt{\ell_j\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\,\mathbf{e}_j\,(j=1,...,p)$$

com

$$c^2 = \frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{(p,n_1+n_2-p-1);1-\alpha}$$

sendo ℓ_j e $\mathbf{e}_j\,(j=1,...,p)$ os valores e vetores próprios de $\mathbf{S}_{pooled}$, respetivamente

Região de confiança para a diferença de 2 vetores médios (populacão normal multivariada, amostras independentes)

- Os IC simultâneos para $\boldsymbol{\mu}_{1j} - \boldsymbol{\mu}_{2j}\,(j = 1, ..., p)$ a $(1-\alpha) \times 100\%$ são dados por

$$(\bar{\mathbf{x}}_{1j} - \bar{\mathbf{x}}_{2j}) \pm c\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s^2_{j,pooled}}$$

sendo $s^2_{j,pooled} = \frac{(n_1-1)s^2_{1j}+(n_2-1)s^2_{2j}}{n_1+n_2-2}$ o estimador ponderado da variância da j-ésima variável, com

$$c^2 = \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{(p, n_1+n_2-p-1);1-\alpha}$$

- Os IC simultâneos de Bonferroni para $\boldsymbol{\mu}_{1j} - \boldsymbol{\mu}_{2j}\,(j = 1, ..., p)$ a $(1-\alpha) \times 100\%$ são dados por

$$(\bar{\mathbf{x}}_{1j} - \bar{\mathbf{x}}_{2j}) \pm t_{(n_1+n_2-2);1-\alpha/(2p)}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s^2_{j,pooled}}$$

Exemplo 8

Considere as seguintes estatísticas amostrais de duas amostrais independentes extraídas de populações normais bivariadas:

Código R:

```
> p<-2
> n1<-50
> n2<-50
> m1<-c(8.3,4.1)
> m2<-c(10.2,3.9)
> S1<-matrix(c(2,1,1,6),2,2)
> S2<-matrix(c(2,1,1,4),2,2)
```

Defina a região de confiança elíptica a 95% para a diferença entre os vetores médios populacionais.

Código R:

```
> Spool<-(n1-1)/(n1+n2-2)*S1+(n2-1)/(n1+n2-2)*S2
> eval<-eigen(Spool)$values
> evec<-eigen(Spool)$vectors
> c<-sqrt((n1+n2-2)*p/(n1+n2-p-1)*qf(0.95,2,n1+n2-p-1))
> (m1-m2)-c*sqrt(eval[1]*(1/n1+1/n2))*evec[,1]

[1] -2.2334961 -0.9014628

> (m1-m2)+c*sqrt(eval[1]*(1/n1+1/n2))*evec[,1]

[1] -1.566504  1.301463
```

Exemplo (continuação)

Código R:

```
> (m1-m2)-c*sqrt(eval[2]*(1/n1+1/n2))*evec[,2]
[1] -1.27685682  0.01132743
> (m1-m2)+c*sqrt(eval[2]*(1/n1+1/n2))*evec[,2]
[1] -2.5231432  0.3886726
```