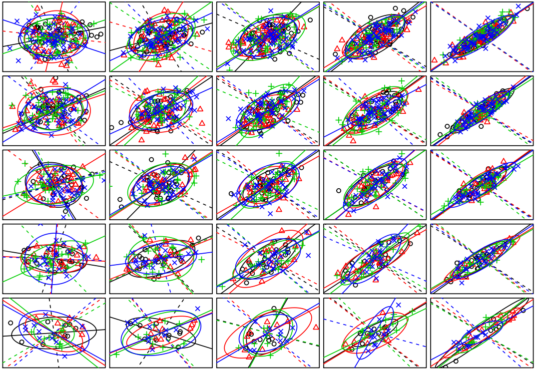
Estatística Multivariada

Slides de apoio às aulas



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

Aula 3

Estudo do ajustamento à normal multivariada

- Muitos métodos inferenciais multivariados assumem $\mathbf{x} \frown N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow$ Como verificar o pressuposto?

- Há vários procedimentos para estudar a normalidade multivariada. Abordaremos 3 processos complementares:

  1. Estudo da normalidade das margens univariadas;
  2. Análise das nuvens de pontos bivariadas;
  3. Análise das distâncias $c_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \, (i = 1, \ldots, n)$

- Note-se que do ponto de vista prático o estudo assenta em larga medida no comportamento a uma e duas dimensões, o que não garante por si só a normalidade multivariada

- Contudo, são raras as estruturas multivariadas que a uma e duas dimensões sejam normais e não o sejam para dimensões ≥ 3

Estudo do ajustamento à normal univariada

- Basicamente, existem dois grupos de métodos no estudo do ajustamento à normal univariada:

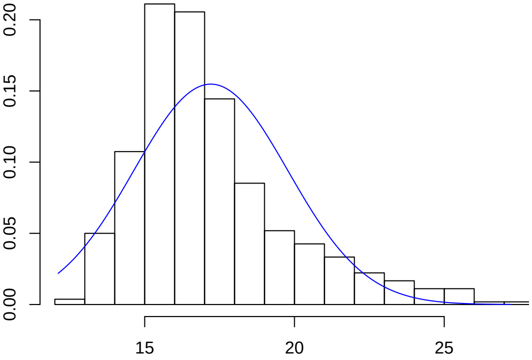
  1. Testes de ajustamento (e.g., testes de Shapiro-wilk, Kolmogorov-Smirnov (Lillefors), etc) e
  2. Métodos gráficos de análise do ajustamento

- Métodos gráficos:

  1. Comparação da distribuição empírica (observada) com a distribuição normal
  2. Estimação da densidade por métodos de *kernel* (*kernel density estimation*) e comparação gráfica entre a fdp estimada com a fdp normal
  3. Análise de gráficos QQ

Código R:

```
> dados2<-as.data.frame(readxl::read_xlsx("./Datasets/data2.xlsx",
                                         col_names = T))
> hist(dados2$imc,probability = T,
      xlim = c(min(dados2$imc),max(dados2$imc)),
      main=NULL,xlab = NULL)
> s<-seq(min(dados2$imc),max(dados2$imc),length.out = length(dados2$imc))
> lines(s,dnorm(s,mean(dados2$imc),sd(dados2$imc)),col=4)
>
```



Estimação da densidade por métodos de *kernel*

- O objetivo geral da estimação da densidade por métodos de kernel estimar a *densidade parente dos dados*
- Mais concretamente, pretende-se estimar a densidade no ponto $x = x_0$ tomando em conta a densidade dos pontos a uma distância h de x_0 (sem necessariamente atribuir o mesmo "peso" a todos os pontos nessa vizinhança)
- Os estimadores *kernel* são função de h (especifica o tamanho da janela (vizinhança) centrada em x_0) e do peso de cada ponto vizinho especificado pela função *kernel* $K(\cdot)$
- A função K satisfaz as condições de (i) simetria em torno x_0; (ii) $\int K(u)du = 1$ e (iii) $\int u^2 K(u)du > 0$.
- Uma função *kernel* frequentemente usada é a densidade Gaussiana, com suporte na reta real. Outras opões de suporte mais limitado incluem, e.g., as funções retangular, *Epanechnikov*:

$$K(u) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{se } |u| \leq 1 \\ 0 & \text{c.c.} \end{cases}$$

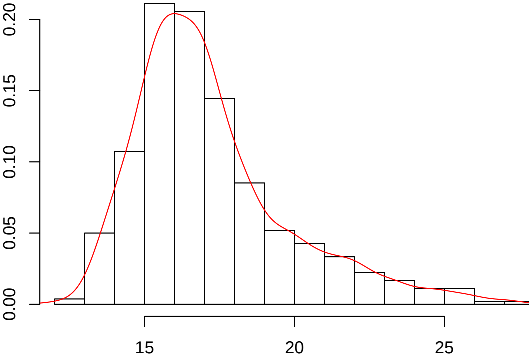
e *tri-cúbica*:

$$K(u) = \begin{cases} (1 - |u|^3)^3 & \text{se } |u| \leq 1 \\ 0 & \text{c.c.} \end{cases}$$

Estimação da densidade por métodos de *kernel*

Código R:

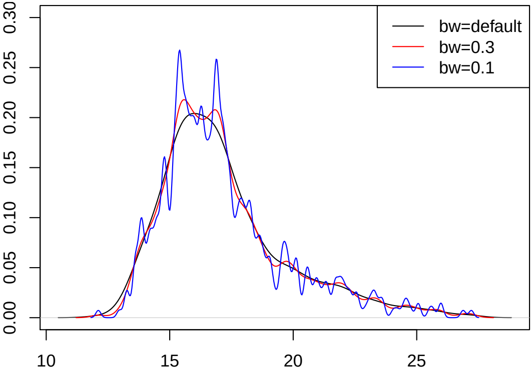
```
> hist(dados2$imc,probability = T,
      xlim = c(min(dados2[,1]),max(dados2[,1])),main=NULL,
      xlab = NULL)
> lines(density(dados2[,1]),col="red")
```



Para analisar o efeito de h (argumento "bw" no R):

Código R:

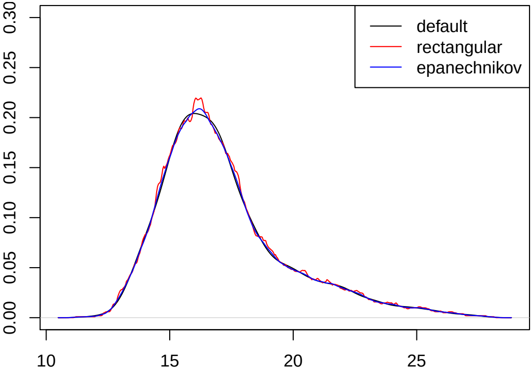
```
> plot(density(dados2[,1]),ylim=c(0,0.3),main = " ",ylab = "Densidade")
> lines(density(dados2[,1],bw = 0.3),col="red")
> lines(density(dados2[,1],bw = 0.1),col="blue")
> legend("topright",col=c(1,2,4),lty=1,legend = c("bw=default","bw=0.3","bw=0.1"))
```



Para analisar o efeito de $K(\cdot)$ (argumento "kernel" no R):

Código R:

```
> plot(density(dados2[,1]),ylim=c(0,0.3),main = " ",ylab = "Densidade")
> lines(density(dados2[,1],kernel = "rectangular"),col="red")
> lines(density(dados2[,1],kernel = "epanechnikov"),col="blue")
> legend("topright",col=c(1,2,4),lty=1,legend = c("default","rectangular","epanechnikov"))
```

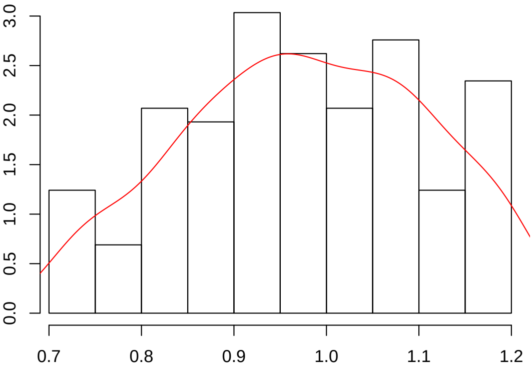


Exemplo 1

(1) Compare a distribuição empírica com a distribuição normal e (2) estude o efeito da variação de h e $K(.)$ na variável x_1 com observações em "data1.xlsx".

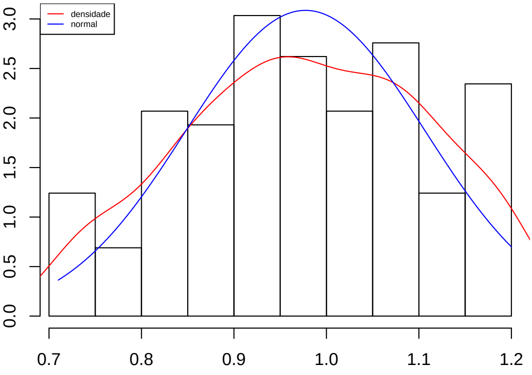
Código R:

```
> dados1<-as.data.frame(readxl::read_xlsx("./Datasets/data1.xlsx",
                                           col_names = F))
> hist(dados1[,1],probability = T,
      xlim = c(min(dados1[,1]),max(dados1[,1])),main=NULL,xlab = NULL)
> lines(density(dados1[,1]),col="red")
```



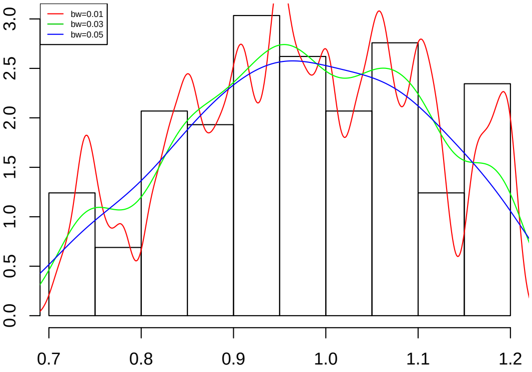
Código R:

```
> hist(dados1[,1],probability = T,
     xlim = c(min(dados1[,1]),max(dados1[,1])),main=NULL,xlab = NULL)
> lines(density(dados1[,1]),col="red")
> s<-seq(min(dados1[,1]),max(dados1[,1]),length.out = length(dados1[,1]))
> lines(s,dnorm(s,mean(dados1[,1]),sd(dados1[,1])),col=4)
> legend("topleft",col=c(2,4),lty=1,legend = c("densidade","normal"),cex=0.5)
```



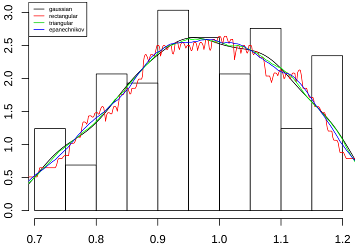
Código R:

```
> hist(dados1[,1],probability = T,
      xlim = c(min(dados1[,1]),max(dados1[,1])),main=NULL,xlab = NULL)
> lines(density(dados1[,1],bw = 0.01),col="red")
> lines(density(dados1[,1],bw = 0.03),col="green")
> lines(density(dados1[,1],bw = 0.05),col="blue")
> legend("topleft",col=c(2,3,4),lty=1,legend = c("bw=0.01","bw=0.03","bw=0.05"),cex=0.5)
```



Código R:

```
> hist(dados1[,1],probability = T,
      xlim = c(min(dados1[,1]),max(dados1[,1])),main=NULL,xlab = NULL)
> lines(density(dados1[,1],kernel = "gaussian"))
> lines(density(dados1[,1],kernel = "rectangular"),col="red")
> lines(density(dados1[,1],kernel = "triangular"),col="green")
> lines(density(dados1[,1],kernel = "epanechnikov"),col="blue")
> legend("topleft",col=c(1,2,3,4),lty=1,
         legend = c("gaussian","rectangular","triangular","epanechnikov"),cex=0.5)
```



Exemplo 2

Considere os dados em "data1.xlsx". Estude a normalidade das margens univariadas comparando as distribuições empirica e teórica. Estime a densidade por métodos de kernel.

Gráficos QQ

- Além do estudo da densidade por métodos de *kernel*, os **gráficos QQ** constituem uma boa ferramenta de análise do ajustamento à distribuição normal (ou a qualquer outra distribuição teórica)

- Os gráficos QQ são gráficos de pontos que relacionam os **quantis amostrais** (observados/empíricos) com os **quantis teóricos** que se esperaria observar se as observações fossem efetivamente provenientes de uma distribuição normal

- O ajustamento à distribuição normal será tanto melhor quanto mais linear for a disposição dos pontos

Gráficos QQ

- Suponhamos a amostra observada (univariada) $x_1, \ldots, x_n$ (sendo x uma v.a. contínua). Como desenhar o gráfico QQ?

1. Calculo das probabilidades empíricas associadas aos quantis amostrais:
   - Seja $x_{(1)}, \ldots, x_{(n)}$ a representação das respetivas estatísticas ordinais, tal que $x_{(1)} \leq \ldots \leq x_{(n)}$
   - $x_{(i)}$ $(i = 1, \ldots, n)$ (quando distintos) são os *quantis amostrais* abaixo dos quais existem exatamente i observações.
   - A proporção i/n representa a **probabilidade empírica** $p_{(j)}$ de observar um valor igual ou inferior a $x_{(i)}$. Contudo, de forma a cobrir melhor o intervalo $[0, 1]$ e para que a probabilidade empírica nunca seja 1, em regra, usa-se a "correção de continuidade" (assintoticamente equivalente)

$$p_{(j)} = \frac{i - 0.5}{n}$$

2. Calculo dos quantis teóricos: Seja $q_{(1)}, \ldots, q_{(n)}$ a representação dos quantis teóricos, tal que

$$q_{(i)} = \Phi^{-1}\left(\frac{i - 0.5}{n}\right) (i = 1, \ldots, n)$$

3. Representar graficamente os pontos $(x_{(i)}, q_{(i)})$ $(i = 1, \ldots, n)$

Exemplo 3

Considere as observações ordenadas:

$$-1.00 \quad -0.10 \quad 0.16 \quad 0.41 \quad 0.62 \quad 0.80 \quad 1.26 \quad 1.54 \quad 1.71 \quad 2.3$$

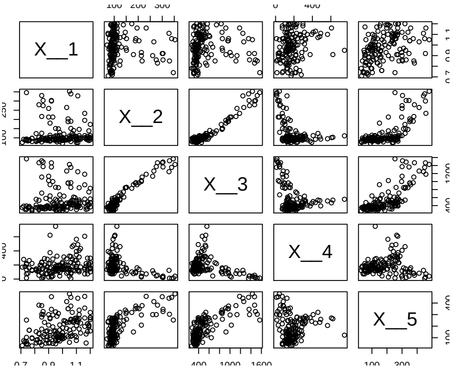
Desenhe o gráfico QQ de ajustamento à distribuição normal padrão.

Exemplo 4

Desenhe o gráfico QQ de ajustamento à normal para a variável x_1 em ''data1.xlsx''.

Código R:

```
> pairs(dados1[,-6])
```



- Vimos anteriormente que $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \frown \chi^2_p$.

- Assim, considerando a amostra $\boldsymbol{x}_1, ..., \boldsymbol{x}_x$ podem calcular-se as distâncias

$$c_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})'\mathbf{S}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\,(i = 1, ..., n)$$

  e estudar o seu ajustamento à distribuição χ^2_p

- Contudo, esta abordagem pode induzir em erro sobretudo para pequenos valores de p (Small, 1978)

- Em alternativa, Gnanadesikan and Kettenring (1972) mostraram que a transformação

$$\boldsymbol{u}_i = \frac{n\,\boldsymbol{c}_i^2}{(n-1)^2}$$

  tem distribuição beta com parâmetros de forma $\alpha = p/2$ e $\beta = (n-p-1)/2$

- Deste modo, depois de calculados os valores $\boldsymbol{u}_i$, pode estudar-se o seu ajustamento à distribuição beta e inferir sobre o ajustamento à normal multivariada.

Exemplo 5

Considere as observações:

i	x_1	x_2
1	126974	4224
2	96933	3835
3	86656	3510
4	63438	3758
5	55264	3939
6	50976	1809
7	39069	2946
8	36156	359
9	35209	2480
10	32416	2413

Calcule as distâncias c^2 e u e estude a normalidade do vetor.

- Considere-se uma amostra aleatória proveniente de uma população normal multivariada com média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$, i.e.

$$(\mathbf{x}_1, ..., \mathbf{x}_n) \text{ iid } \mathbf{x}_i \frown N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- É possível mostrar que os **estimadores** de máxima verosimilhança de $\boldsymbol{\mu}$ e de $\boldsymbol{\Sigma}$, são respetivamente

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{(n-1)}{n}\mathbf{S}$$

- Os correspondentes valores observados, $\bar{x}$ e $(n-1)S/n$, são as **estimativas** de máxima verosimilhança de $\boldsymbol{\mu}$ e de $\boldsymbol{\Sigma}$
- Note-se que o estimador $\hat{\boldsymbol{\Sigma}}$ é enviesado pelo que, em regra, usa-se $\mathbf{S}$ para estimar $\boldsymbol{\Sigma}$
- As distribuição de probabilidade das estatísticas amostrais são designadas por **distribuições de amostragem** e são a base da inferência estatística (clássica)

- Recorde-se que numa população univariada $N(\mu, \sigma^2)$ então

$$\bar{x} \frown N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Seja $\mathbf{x}_1, ...\mathbf{x}_n$ uma amostra aleatória de uma população normal multivariada com média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$, isto é, $\mathbf{x}_i \frown N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,(i, ..., n)$ independentes, então

$$\bar{\mathbf{x}} \frown N_p\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{n}\right)$$

e

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \frown \chi^2_p$$

- Pelo TLC, se $\mathbf{x}_i$ $(i = 1, ..., n)$ são réplicas iid de uma qualquer população multivariada com média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$ então

$$\bar{\mathbf{x}} \xrightarrow[n\to\infty]{d} N_p\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{n}\right) \Leftrightarrow \bar{\mathbf{x}} \overset{a}{\sim} N_p\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{n}\right)$$

isto é,

$$\sqrt{n}\,(\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{d} N_p(\mathbf{0}, \boldsymbol{\Sigma})$$

ou

$$\sqrt{n}\boldsymbol{\Sigma}^{-1/2}\,(\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{d} N_p(\mathbf{0}, \mathbf{I})$$

- Do resultado anterior (com $n >> p$), considerando $\mathbf{x}_i$ $(i = 1, ..., n)$ réplicas iid de uma qualquer população multivariada, tem-se que

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \overset{a}{\sim} \chi^2_p$$

- Numa população univariada

$$\sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(x_i - \mu)(x_i - \mu)}{\sigma^2} \frown \chi_n^2$$

sendo x_i variáveis iid $N(\mu, \sigma)$. Substituindo μ pelo estimador $\bar{x}$,

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \frown \chi_{n-1}^2$$

- Sendo $\mathbf{x}_i \frown N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Leftrightarrow (\mathbf{x}_i - \boldsymbol{\mu}) \frown N_p(\mathbf{0}, \boldsymbol{\Sigma})$ $(i = 1, ..., n)$ uma a.a. de uma população multivariada, então

$$\sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' \frown W_p(n, \boldsymbol{\Sigma})$$

onde $W_p(n, \boldsymbol{\Sigma})$ representa a **distribuição de Wishart** com parâmetros n (graus de liberdade) e $\boldsymbol{\Sigma}$

- Tal como no caso univariado, substituindo $\boldsymbol{\mu}$ por $\bar{\mathbf{x}}$

$$\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = (n-1)\mathbf{S} \frown W_p(n-1, \boldsymbol{\Sigma})$$

Exemplo 6

Seja $\mathbf{x}_1, ..., \mathbf{x}_{20}$ uma a.a. proviniente de uma população $N_6(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
a) Qual a distribuição de $(\mathbf{x}_1 - \boldsymbol{\mu})' \boldsymbol{\Sigma} (\mathbf{x}_1 - \boldsymbol{\mu})$? **(xi-u)'Cov(xi-u) segue uma qui quadrado com P graus**
b) Qual a distribuição de $\bar{\mathbf{x}}$ e de $\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu})$?
c) Qual a distribuição de $(n-1)\mathbf{S}$?

Inferência sobre um vetor de médias

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

- Note-se que rejeitar H_0 para valores elevados de $|t|$ equivale a rejeitar a hipótese para valores elevados de t^2, sendo

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

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

- Intervalo de confiança para μ a $(1 - \alpha) \times 100\%$

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

- Considerem-se agora as hipóteses estatísticas

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

- Sendo **x** um vetor aleatório, pode generalizar-se a expressão univariada de t^2, obtendo-se

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

- 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

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

- 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 $F_{(p,n-p)}$

Exemplo 7

Considere-se a amostra observada:

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

Iremos testar a hipótese

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

Exemplo 8

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

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

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

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

tal que

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

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

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

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

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

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

Exemplo 9

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.

c) Estudo o ajustamento à distribuição normal bivariada.

Código R:

```
> #a)
> dados3<-as.data.frame(readxl::read_xlsx("./Datasets/data3.xlsx",col_names = T))
> m<-colMeans(dados3); S<-var(dados3); eigen<-eigen(S)
> p<-ncol(dados3); n<-nrow(dados3)
> 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]

       v1        v2
0.1621041 0.2175663

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

        v1         v2
0.10105381 0.09506531

> #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]

        v1         v2
0.08510906 0.17947478

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

       v1        v2
0.1780488 0.1331568
```

Código R:

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

             [,1]
[1,] 251646598456

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

[1] 10.78734
```

- 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 $s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$

- Então, 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 $1-\alpha$.

- 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

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

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

$$\ldots$$

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

  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 10

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

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

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

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

$$\cdots$$

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

- 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

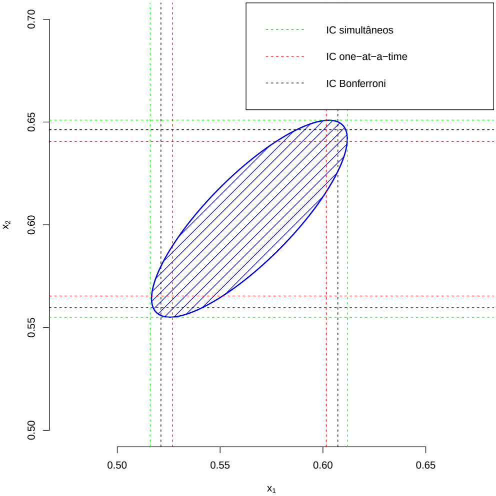
$$(1-\alpha)^p$$

- 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 μ (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

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

- 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*:

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

$$\ldots$$

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