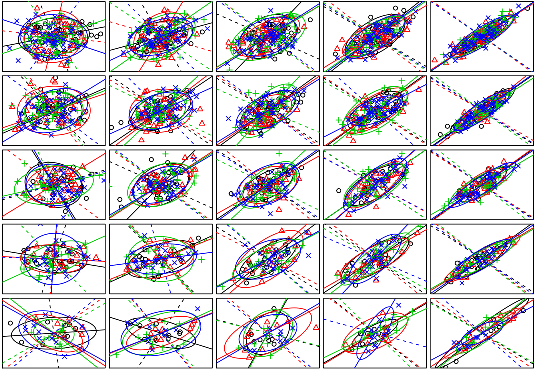
Estatística Multivariada

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 $\mathbf{x}' = (x_1, ..., x_p) \frown N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow$ 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:

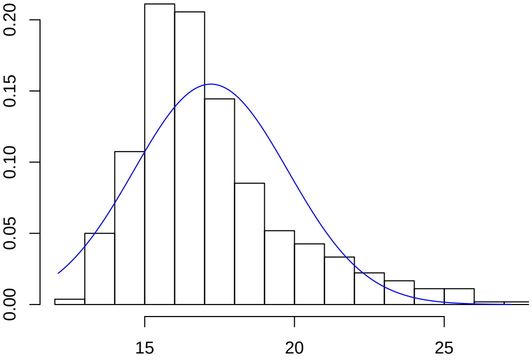
  1. Testes de ajustamento (e.g., testes de Shapiro-wilk, Kolmogorov-Smirnov (Lillefors), etc) e
  2. Métodos gráficos de análise do ajustamento

- 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)
>
```



- 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*:

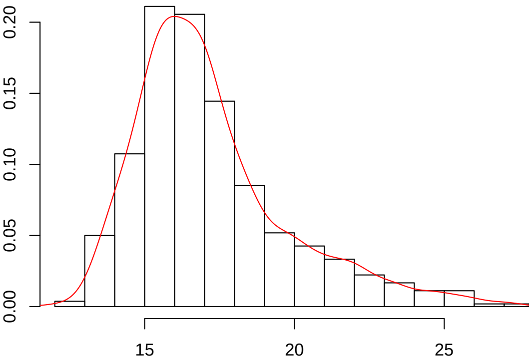
$$K(u) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{se } |u| \leq 1 \\ 0 & \text{c.c.} \end{cases}$$

e *tri-cúbica*:

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

Código R:

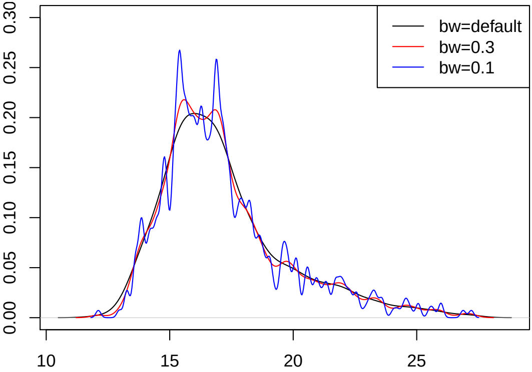
```
> hist(dados2$imc,probability = T,
      xlim = c(min(dados2[,1]),max(dados2[,1])),main=NULL,
      xlab = NULL)
> lines(density(dados2[,1]),col="red")
```



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

Código R:

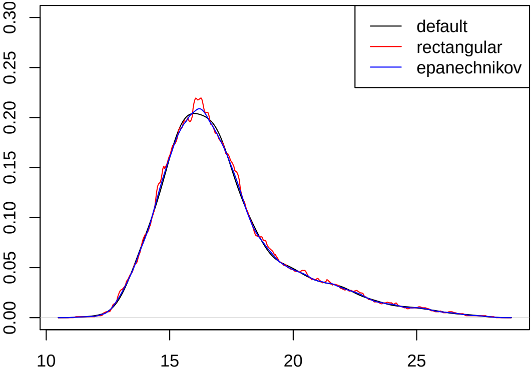
```
> plot(density(dados2[,1]),ylim=c(0,0.3),main = " ",ylab = "Densidade")
> lines(density(dados2[,1],bw = 0.3),col="red")
> lines(density(dados2[,1],bw = 0.1),col="blue")
> legend("topright",col=c(1,2,4),lty=1,legend = c("bw=default","bw=0.3","bw=0.1"))
```



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

Código R:

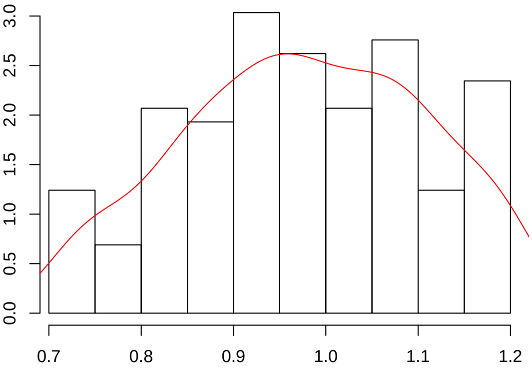
```
> plot(density(dados2[,1]),ylim=c(0,0.3),main = " ",ylab = "Densidade")
> lines(density(dados2[,1],kernel = "rectangular"),col="red")
> lines(density(dados2[,1],kernel = "epanechnikov"),col="blue")
> legend("topright",col=c(1,2,4),lty=1,legend = c("default","rectangular","epanechnikov"))
```



(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".

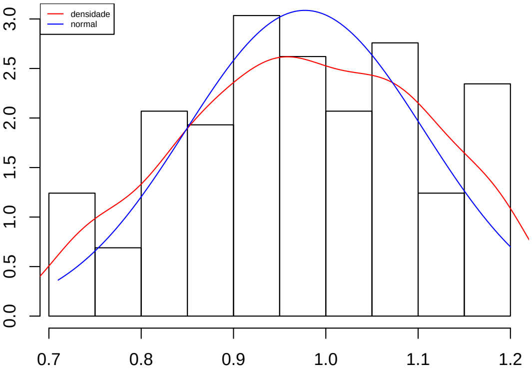
Código R:

```
> 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")
```



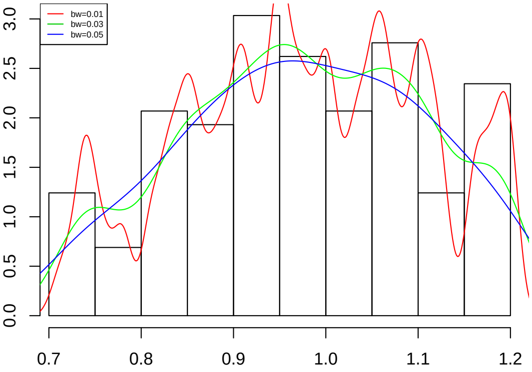
Código R:

```
> 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)
```



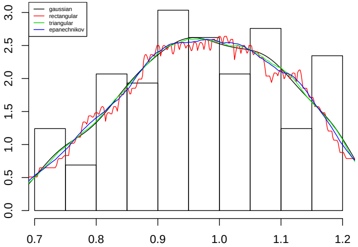
Código R:

```
> 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)
```

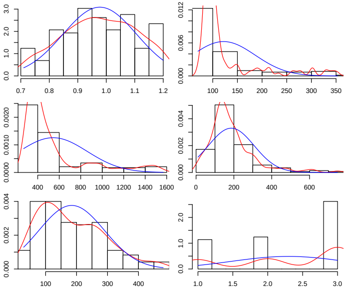


Exemplo 2

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.

Código R:

```
> par(mfrow=c(3,2),mar=c(1,2,2,1),oma=c(1,1,0,0))
> for (i in 1:6) {
 hist(dados1[,i],probability = T,
      xlim = c(min(dados1[,i]),max(dados1[,i])),main=NULL,xlab = NULL)
 lines(density(dados1[,i]),col="red")
 s<-seq(min(dados1[,i]),max(dados1[,i]),length.out = length(dados1[,i]))
 lines(s,dnorm(s,mean(dados1[,i]),sd(dados1[,i])),col=4)
 }
```



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) $x_1, ..., x_n$ (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.

$x_{(i)}$	$p_{(i)}$	$q_{(i)}$
-1.00	0.05	$\Phi^{-1}(0.05) = z_{0.05} = -1.645$
-0.10	0.15	$\Phi^{-1}(0.15) = z_{0.15} = -1.036$
0.16	0.25	$\Phi^{-1}(0.25) = z_{0.25} = -0.674$
0.41	0.35	$\Phi^{-1}(0.35) = z_{0.35} = -0.385$
0.62	0.45	$\Phi^{-1}(0.45) = z_{0.45} = -0.125$
0.80	0.55	$\Phi^{-1}(0.55) = z_{0.55} = 0.125$
1.26	0.65	$\Phi^{-1}(0.65) = z_{0.65} = 0.385$
1.54	0.75	$\Phi^{-1}(0.75) = z_{0.75} = 0.674$
1.71	0.85	$\Phi^{-1}(0.85) = z_{0.85} = 1.036$
2.30	0.95	$\Phi^{-1}(0.95) = z_{0.95} = 1.645$

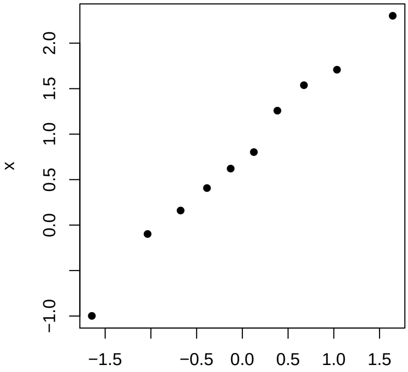
Código R:

```
> x<-c(-1,-0.1,.16,.41,.62,.8,1.26,1.54,1.71,2.3)
> pi<-((seq(1,length(x)))-0.5)/length(x); qi<-round(qnorm(pi),3)
> qi

 [1] -1.645 -1.036 -0.674 -0.385 -0.126  0.126  0.385  0.674  1.036  1.645
```

Exemplo (continuação)

Código R:

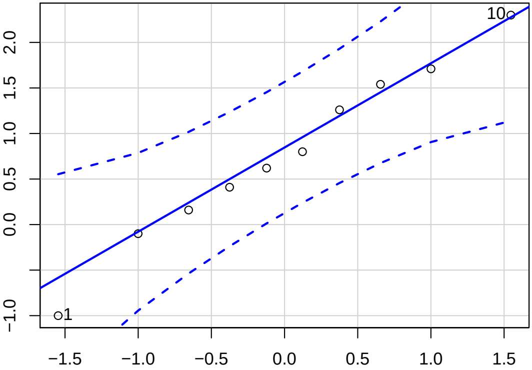
```
> par(pty="s")
> plot(qi,x,pch=16)
```



Exemplo (continuação)

Código R:

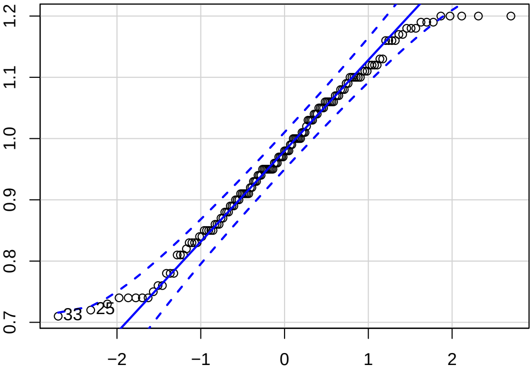
```
> library(car)
> qqp(x,distribution="norm")

[1]  1 10
```



Desenho o gráfico QQ de ajustamento à normal para a variável x_1 em "data1.xlsx".

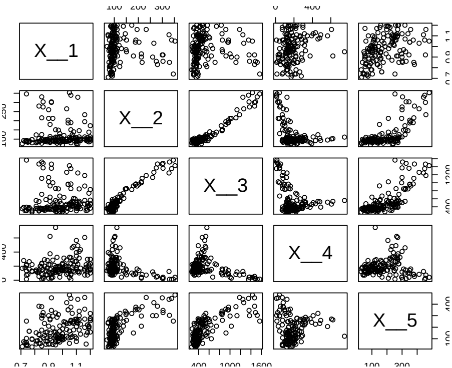
Código R:

```
> car::qqp(dados1[,1])

[1] 33 25
```



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.

Código R:

```
> x1<-c(126974,96933,86656,63438,55264,50976,39069,36156,35209,32416)
> x2<-c(4224,3835,3510,3758,3939,1809,2946,359,2480,2413)
> X<-matrix(c(x1,x2),10,2)
> c2<-numeric()
> for (i in 1:nrow(X)){
 c2[i]<-t(X[i,]-colMeans(X))%*%solve(var(X))%*%(X[i,]-colMeans(X))}
> c2<-round(c2[order(c2)],3)
```

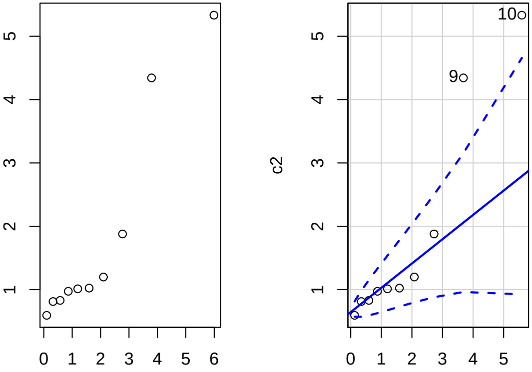
Exemplo (continuação)

Código R:

```
> #distâncias
> c2

 [1] 0.594 0.812 0.830 0.974 1.013 1.024 1.199 1.879 4.343 5.333

> #variável u
> u<-(length(c2)*c2)/(length(c2)-1)^2
> u

 [1] 0.07333333 0.10024691 0.10246914 0.12024691 0.12506173 0.12641975
 [7] 0.14802469 0.23197531 0.53617284 0.65839506

> #quantis teóricos qui-quadrado
> qi_chi<-round(qchisq(pi,ncol(X)),3)
> qi_chi

 [1] 0.103 0.325 0.575 0.862 1.196 1.597 2.100 2.773 3.794 5.991

> #quantis teóricos beta
> qi_beta<-round(qbeta(pi,shape1 = ncol(X)/2,shape2 = (nrow(X)-ncol(X)-1)/2),3)
> qi_beta

 [1] 0.015 0.045 0.079 0.116 0.157 0.204 0.259 0.327 0.418 0.575
```

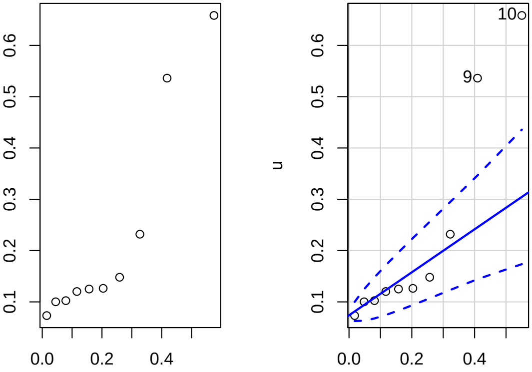
Código R:

```
> par(mfrow=c(1,2))
> plot(qi_chi,c2)
> qqp(c2,distribution="chisq",df=ncol(X))

[1] 10  9
```



Código R:

```
> par(mfrow=c(1,2))
> plot(qi_beta,u)
> qqp(u,distribution="beta",
     shape1 = ncol(X)/2,shape2 = (nrow(X)-ncol(X)-1)/2)

[1] 10  9
```



- 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

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

sendo x_i $(i=1,\ldots,n)$ variáveis iid $N(\mu,\sigma)$. Substituindo μ pelo estimador $\bar{x}$,

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

- Numa população multivariada

$$\mathbf{x}_i \frown N_p(\boldsymbol{\mu},\boldsymbol{\Sigma}) \Leftrightarrow \mathbf{w}_i = (\mathbf{x}_i-\boldsymbol{\mu}) \frown N_p(\mathbf{0},\boldsymbol{\Sigma})\ (i=1,\ldots,n)$$

- A matriz definida por $\sum_{i=1}^{n} \mathbf{w}_i\mathbf{w}_i'$ diz-se ter **distribuição de Wishart** com parâmetros n (graus de liberdade) e $\boldsymbol{\Sigma}$, i.e.,

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

- 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 $(\mathbf{x}_1 - \boldsymbol{\mu})'\boldsymbol{\Sigma}(\mathbf{x}_1 - \boldsymbol{\mu})$?

$$(\mathbf{x}_1 - \boldsymbol{\mu})'\boldsymbol{\Sigma}(\mathbf{x}_1 - \boldsymbol{\mu}) \frown \chi^2_1$$

b) Qual a distribuição de $\bar{\mathbf{x}}$ e de $\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu})$?

$$\bar{\mathbf{x}} \frown N_6\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{20}\right)$$

$$\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \frown N_6(\mathbf{0}, \boldsymbol{\Sigma})$$

c) Qual a distribuição de $(n-1)\mathbf{S}$?

$$19\mathbf{S} \frown W_6(19, \boldsymbol{\Sigma})$$