Normal multivariada

Patrícia de Siqueira Ramos

UNIFAL-MG, *campus* Varginha

7 de Abril de 2016

- Assim como no caso univariado, todo vetor aleatório p-variado tem seus valores gerados por algum mecanismo probabilístico
- Há várias distribuições de probabilidade multivariadas:
  - T^2 de Hotelling (correspondente da t univariada)
  - Wishart (correspondente da χ^2 univariada)
  - Normal (a mais conhecida e importante para alguns procedimentos multivariados)

Normal multivariada

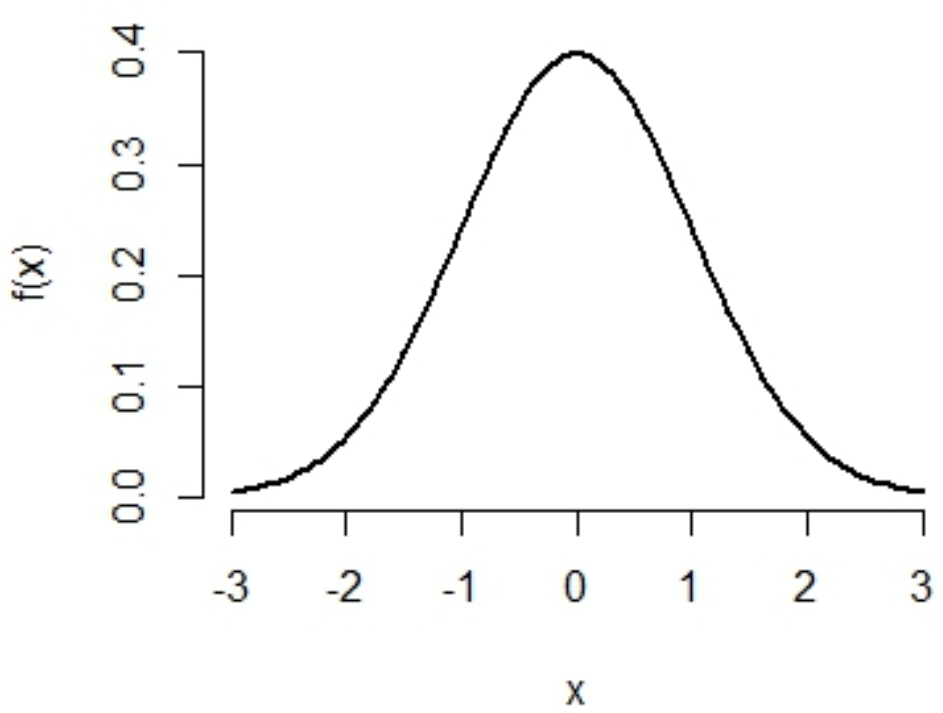
Normal univariada

- A normal univariada com média μ e variância σ^2 tem f.d.p.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\},$$

  $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$.
- Notação: $X \sim N(\mu, \sigma)$.

Forma genérica da normal univariada



$$X \sim N(\mu = 0, \sigma^2 = 1)$$

Normal multivariada

- Generalização da normal univariada quando há duas ou mais v.a.s simultaneamente
- Para um vetor aleatório com p variáveis, $\mathbf{X^T} = [X_1 \dots X_p]$, a densidade de $\mathbf{X}$ é dada por:

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right\},$$

$-\infty < X_i < \infty,\ i = 1, \dots, p$
$\boldsymbol{\mu} \in \mathbb{R}^p$ (vetor de médias das variáveis)
$\mathbf{\Sigma}$ positiva definida

Normal multivariada

- Generalização da normal univariada quando há duas ou mais v.a.s simultaneamente
- Para um vetor aleatório com p variáveis, $\mathbf{X^T} = [X_1 \dots X_p]$, a densidade de $\mathbf{X}$ é dada por:

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right\},$$

$-\infty < X_i < \infty,\ i = 1, \dots, p$
$\boldsymbol{\mu} \in \mathbb{R}^p$ (vetor de médias das variáveis)
$\mathbf{\Sigma}$ positiva definida

- Notação: $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$.
- A quantidade $(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$ é a distância de Mahalanobis do vetor $\mathbf{X}$ ao vetor de médias $\boldsymbol{\mu}$

Normal bivariada

O exemplo mais simples da normal multivariada é a bivariada ($p = 2$):

$$f(x_1, x_2) = \frac{1}{(2\pi)|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})\right\},$$

Normal bivariada

O exemplo mais simples da normal multivariada é a bivariada ($p = 2$):

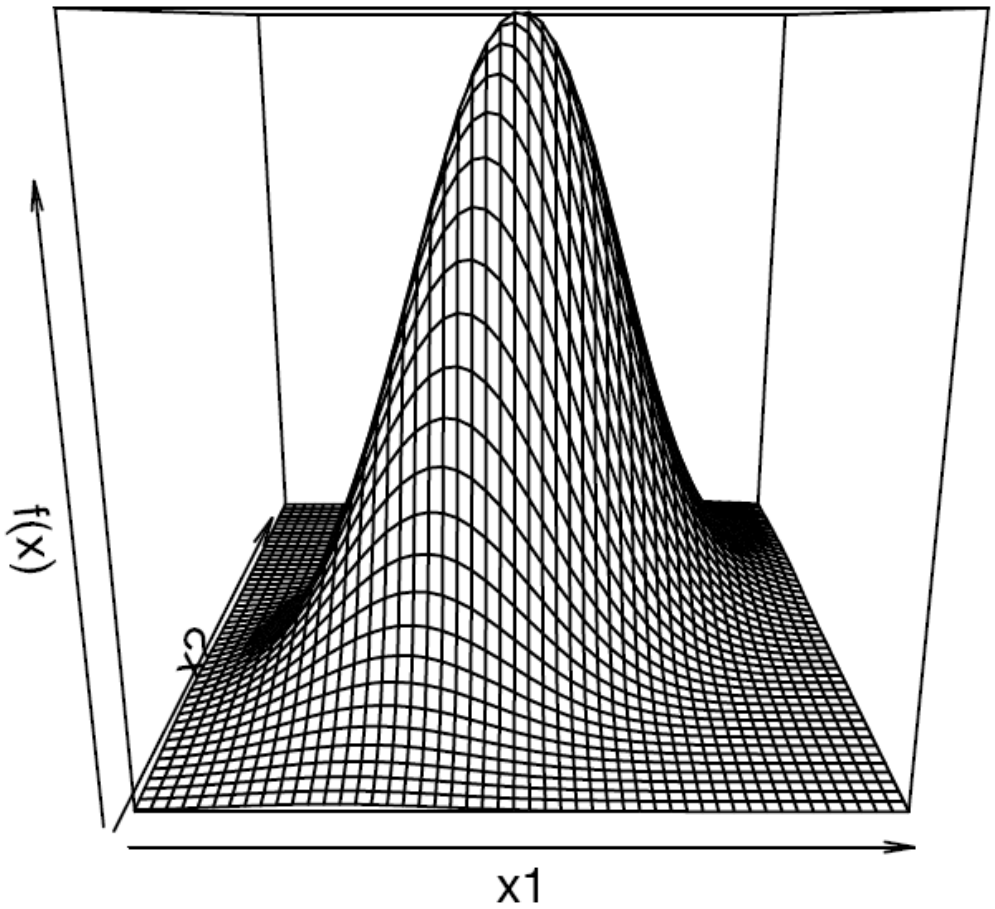
$$f(x_1, x_2) = \frac{1}{(2\pi)|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})\right\},$$

que pode ser escrita explicitamente por

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}}$$
$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

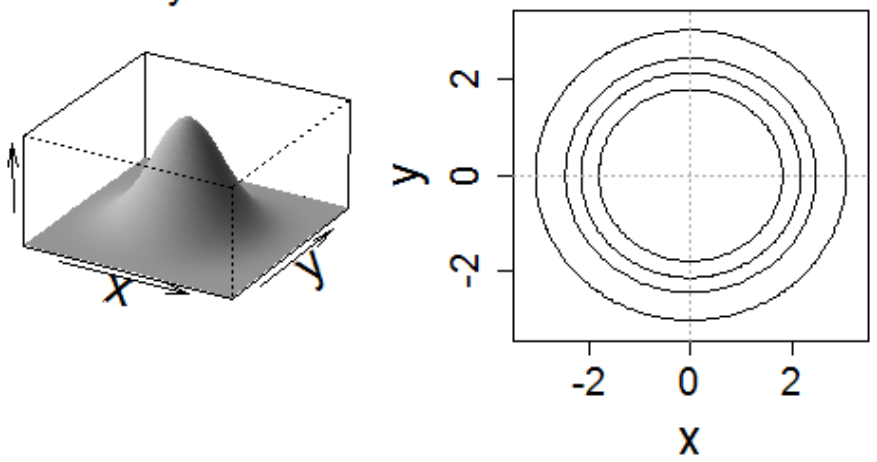
em que μ_1 e μ_2 são as médias populacionais de X_1 e X_2, σ_1^2 e σ_2^2 são suas variâncias e ρ é a correlação entre elas.

Normal bivariada com $\rho = 0,5$

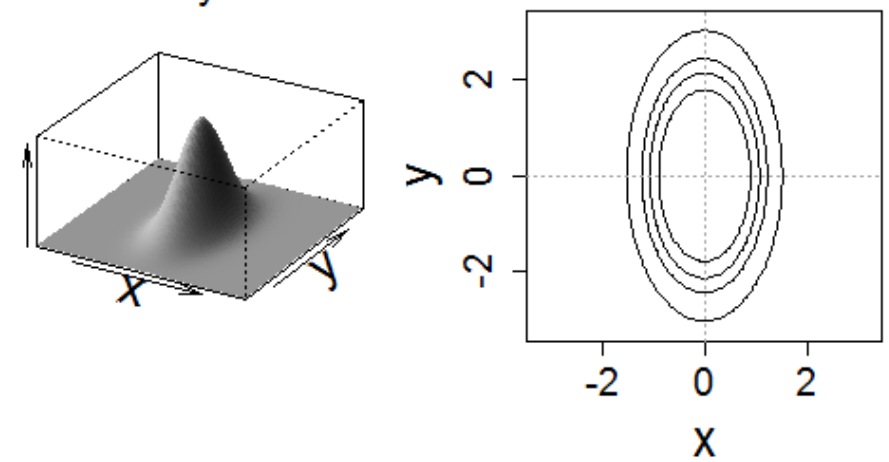


Normais bivariadas

$\sigma_x = \sigma_y, \ \rho = 0$

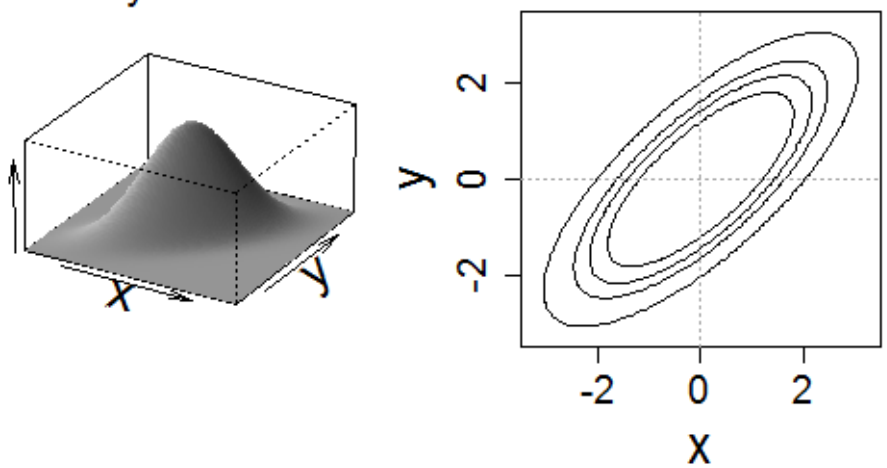


$2\sigma_x = \sigma_y, \ \rho = 0$

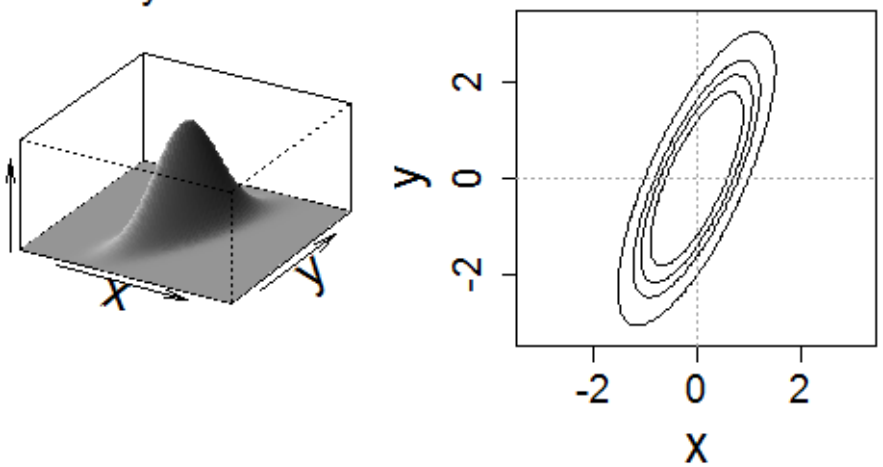


Normais bivariadas

$\sigma_x = \sigma_y, \ \rho = 0{,}75$

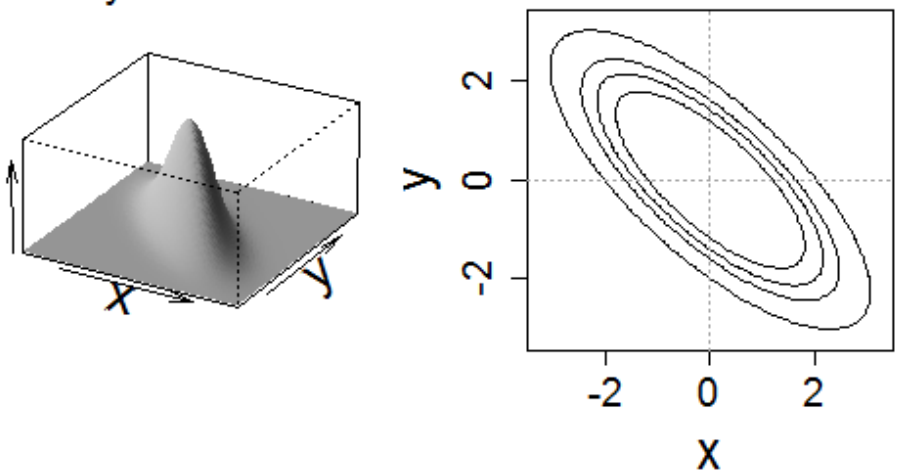


$2\sigma_x = \sigma_y, \ \rho = 0{,}75$

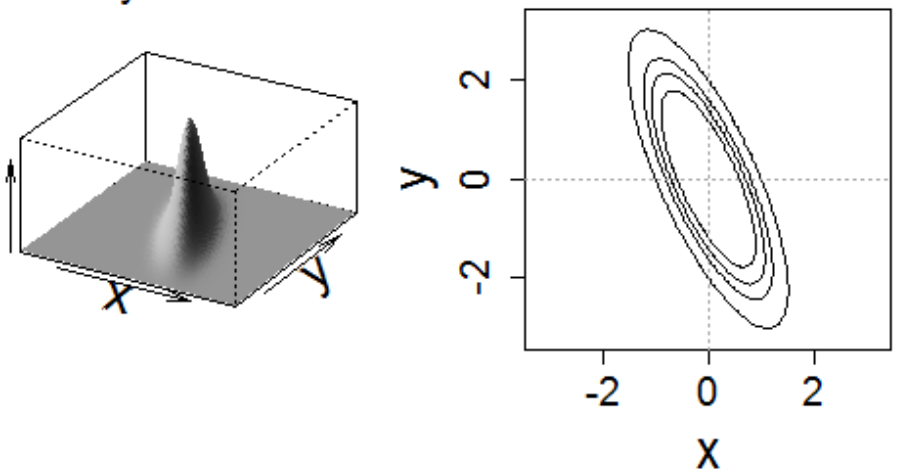


Normais bivariadas

$\sigma_x = \sigma_y, \ \rho = -0{,}75$



$2\sigma_x = \sigma_y, \ \rho = -0{,}75$



Propriedades da normal multivariada

- Combinações lineares das variáveis $\mathbf{X}$ (se $\mathbf{X} \sim N_p$),

$$Z = a_1 X_1 + \cdots + a_p X_p,$$

são normalmente distribuídas com média $\mathbf{a}^T \boldsymbol{\mu}$ e variância $\mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a}$, sendo $\mathbf{a}^T = [a_1 \quad \ldots \quad a_p]$

Propriedades da normal multivariada

- Combinações lineares das variáveis $\mathbf{X}$ (se $\mathbf{X} \sim N_p$),

$$Z = a_1 X_1 + \cdots + a_p X_p,$$

são normalmente distribuídas com média $\mathbf{a}^T \boldsymbol{\mu}$ e variância $\mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a}$, sendo $\mathbf{a}^T = [a_1 \quad \ldots \quad a_p]$

- Quando $\mathbf{X} \sim N_p$, qualquer subconjunto de k variáveis de $\mathbf{X}$, $k < p$, também terá distribuição normal k-variada
- Se $\mathbf{X} \sim N_p$ e $\boldsymbol{\Sigma}$ é diagonal, então as variáveis de $\mathbf{X}$ são normais independentemente distribuídas (i.i.d.)

Verificação da normalidade multivariada

a) Análise das distribuições uni e bivariadas

- Avaliar as distribuições marginais univariadas das p variáveis em uma a.a. de tamanho n (procedimento questionado):

$$\text{Se } \mathbf{X} \sim N_p, \text{ espera-se } X_1 \sim N, \ldots, X_p \sim N, \text{ mas}$$
$$\text{Se } X_1 \sim N, \ldots, X_p \sim N \text{ não se garante } \mathbf{X} \sim N_p$$

- Porém, realizar testes univariados de normalidade para $X_1, \ldots, X_p$ dão algum indicativo da normalidade multivariada e podem ser úteis

a) Análise das distribuições uni e bivariadas

Para cada variável do conjunto de dados, a suposição de normalidade **univariada** pode ser verificada por meio de

- Métodos gráficos (PP e QQ *plots*)
- Testes de hipóteses formais (Shapiro-Wilk, Anderson-Darling, D'Agostino etc.)

a) Análise das distribuições uni e bivariadas

Para cada variável do conjunto de dados, a suposição de normalidade **univariada** pode ser verificada por meio de

- Métodos gráficos (PP e QQ *plots*)
- Testes de hipóteses formais (Shapiro-Wilk, Anderson-Darling, D'Agostino etc.)

Já a suposição de normalidade **bivariada** pode ser verificada pela construção de gráficos de dispersão $X_i \times X_j$, $i \neq j$:

- Todos os pares deverão ter distribuição normal bivariada e, portanto, forma de elipse

a) Análise das distribuições uni e bivariadas

Para cada variável do conjunto de dados, a suposição de normalidade **univariada** pode ser verificada por meio de

- Métodos gráficos (PP e QQ *plots*)
- Testes de hipóteses formais (Shapiro-Wilk, Anderson-Darling, D'Agostino etc.)

Já a suposição de normalidade **bivariada** pode ser verificada pela construção de gráficos de dispersão $X_i \times X_j$, $i \neq j$:

- Todos os pares deverão ter distribuição normal bivariada e, portanto, forma de elipse

a) Análise das distribuições uni e bivariadas

Para cada variável do conjunto de dados, a suposição de normalidade **univariada** pode ser verificada por meio de

- Métodos gráficos (PP e QQ *plots*)
- Testes de hipóteses formais (Shapiro-Wilk, Anderson-Darling, D'Agostino etc.)

Já a suposição de normalidade **bivariada** pode ser verificada pela construção de gráficos de dispersão $X_i \times X_j$, $i \neq j$:

- Todos os pares deverão ter distribuição normal bivariada e, portanto, forma de elipse

Obs.: Os métodos gráficos são visuais e, portanto, limitados.

PP *plots*

As coordenadas do PP *plot* são as probabilidades acumuladas $p_1(q)$ e $p_2(q)$ para diferentes valores de q com

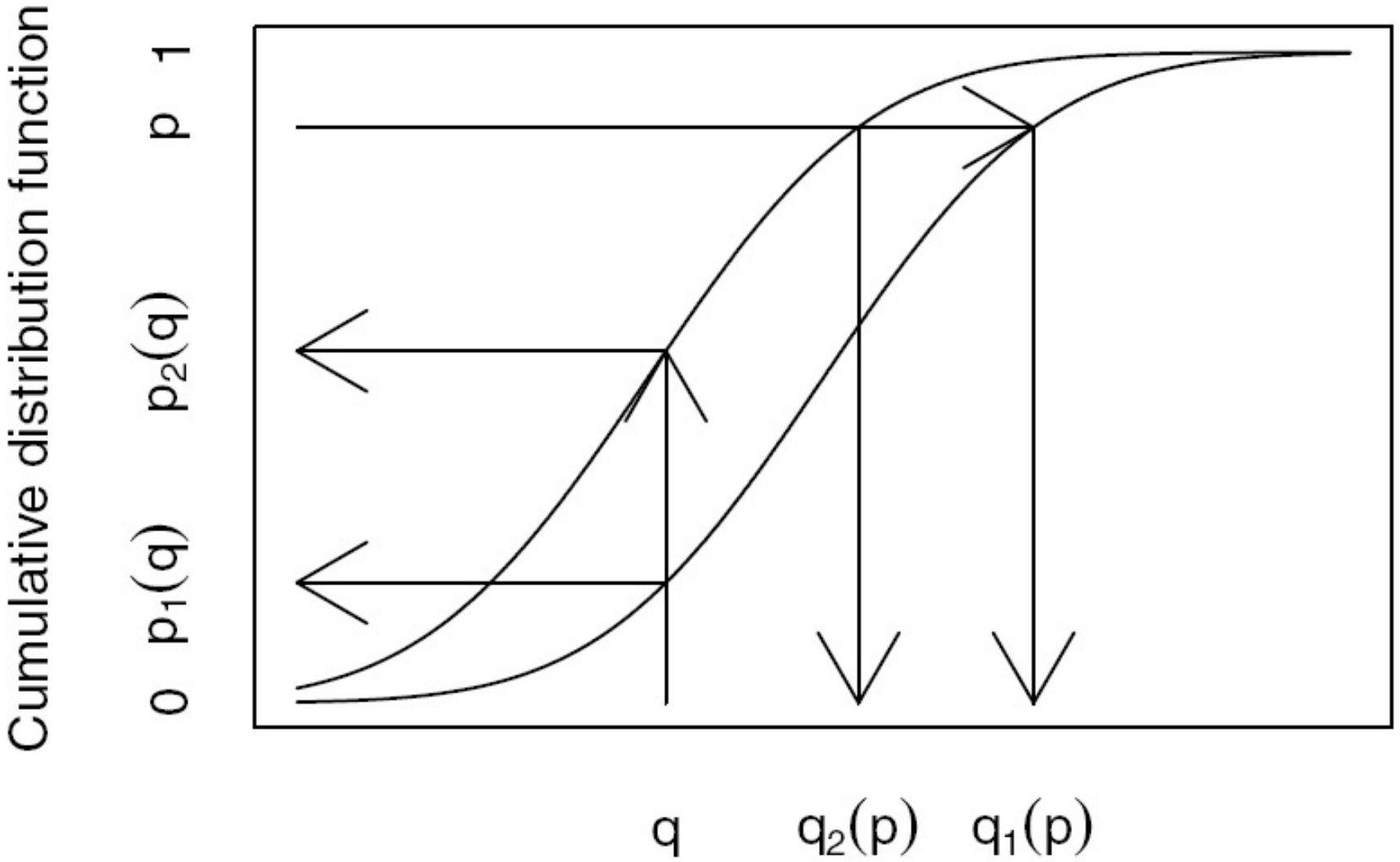
$$p_1(q) = P(X_1 \leq q),$$
$$p_2(q) = P(X_2 \leq q),$$

para as v.a.s X_1 e X_2.

QQ *plots*

As coordenadas do QQ *plot* são os quantis $q_1(p)$ e $q_2(p)$ para diferentes valores de p com

$$q_1(p) = p_1^{-1}(p),$$
$$q_2(p) = p_2^{-1}(p).$$


Cumulative distribution function
p 1
p2(q)
0 p1(q)
q
q2(p) q1(p)

Exemplo de obtenção de um QQ *plot*

102,13848	97,92823	97,95138	106,11683	104,39047
98,40948	94,98110	98,38730	97,93524	107,21028
94,38919	103,55796	105,81335	106,84641	104,06928
101,98353	100,69409	96,99118	87,54581	97,74081

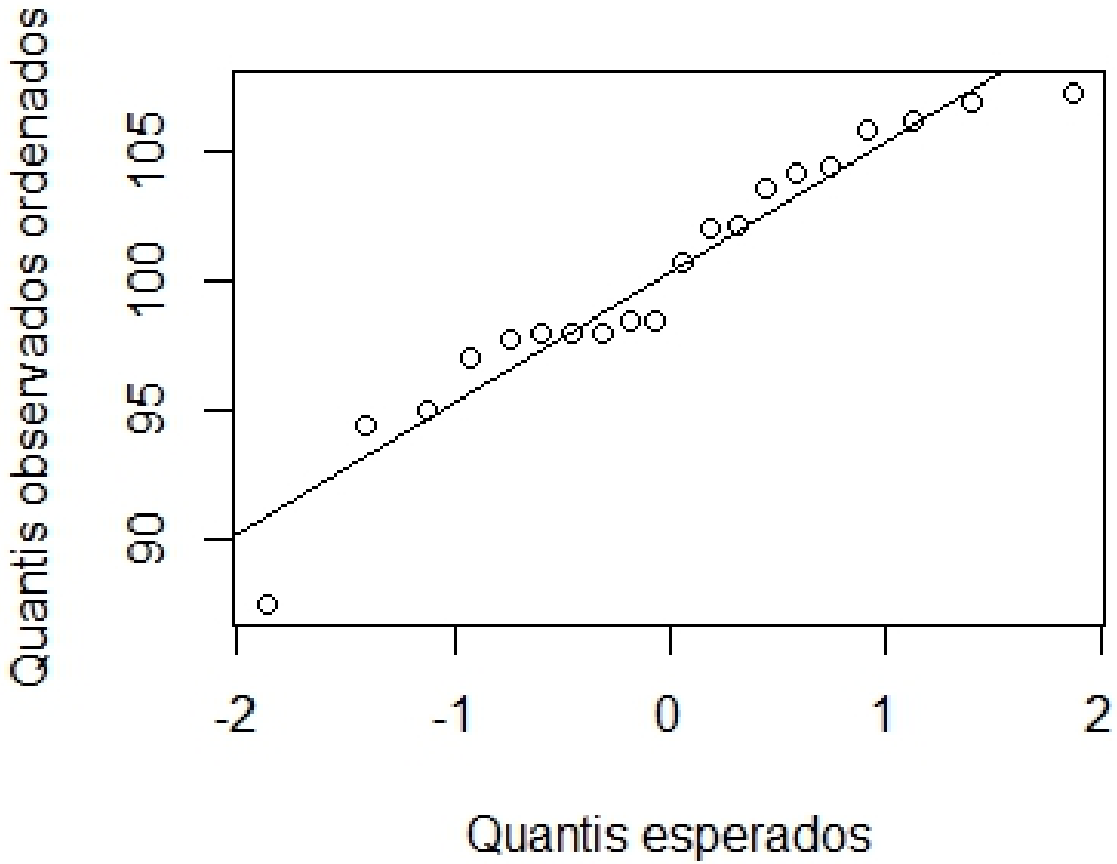
Passos:

1: Ordenar os dados amostrais
2: Calcular $P_i = (i - 0,375)/(n + 0,25)$
3: Obter os quantis da normal por $m_i = \Phi^{-1}[P_i]$
4: Plotar $m_i \times X_{(i)}$ (quantis esperados $\times$ quantis observados)

Valores obtidos - QQ *plot*

i	P_i	m_i	$X_{(i)}$
1	0,03086420	-1,86824165	87,54581
2	0,08024691	-1,40341264	94,38919
3	0,12962963	-1,12814365	94,98110
4	0,17901235	-0,91913552	96,99118
5	0,22839506	-0,74414274	97,74081
6	0,27777778	-0,58945580	97,92823
7	0,32716049	-0,44776752	97,93524
8	0,37654321	-0,31457229	97,95138
9	0,42592593	-0,18675612	98,38730
10	0,47530864	-0,06193162	98,40948
11	0,52469136	0,06193162	100,69409
12	0,57407407	0,18675612	101,98353
13	0,62345679	0,31457229	102,13848
14	0,67283951	0,44776752	103,55796
15	0,72222222	0,58945580	104,06928
16	0,77160494	0,74414274	104,39047
17	0,82098765	0,91913552	105,81335
18	0,87037037	1,12814365	106,11683
19	0,91975309	1,40341264	106,84641
20	0,96913580	1,86824165	107,21028

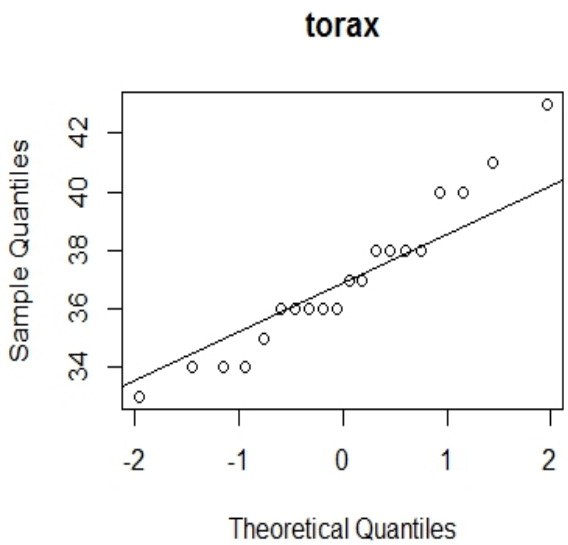
QQ *plot* obtido

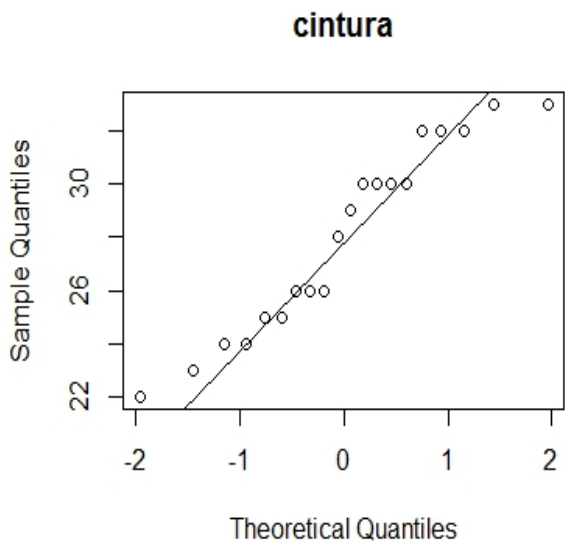


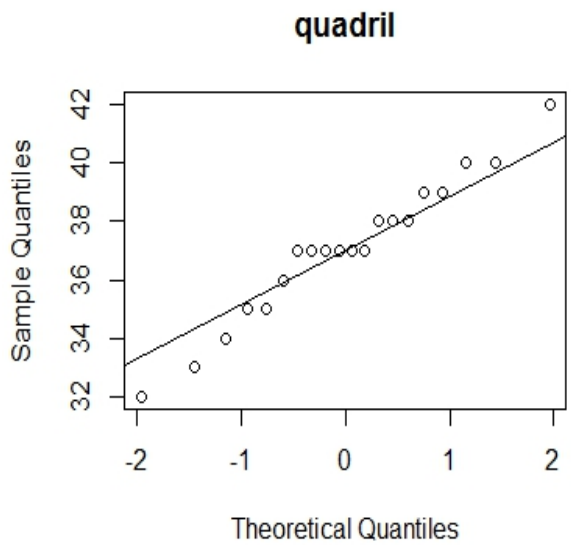
QQ e PP *plots*

- Os PP e QQ *plots* examinam cada variável separadamente
- Se a distribuição dos dados da variável em questão for uma normal, os dois métodos devem gerar gráficos com pontos sobre uma linha reta

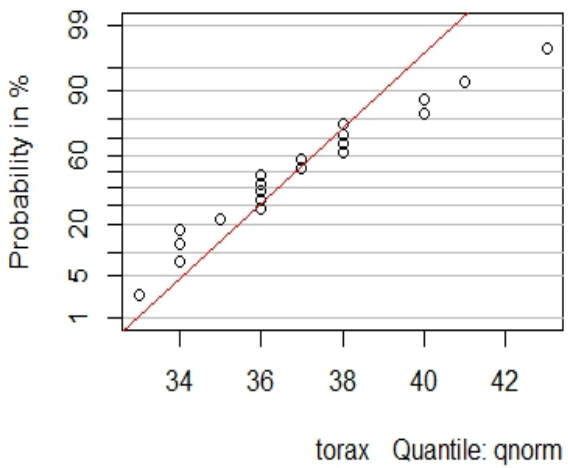
QQ *plots* - exemplo das medidas

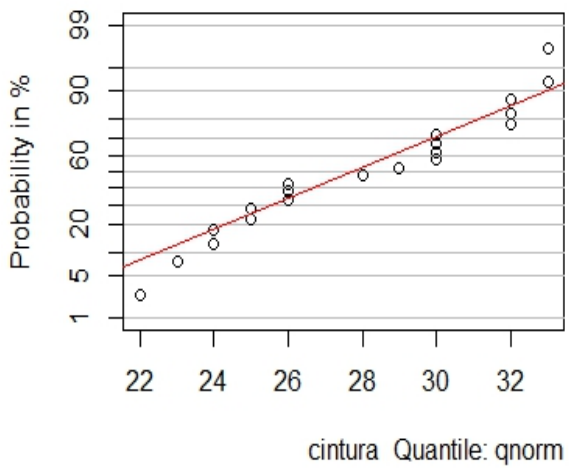


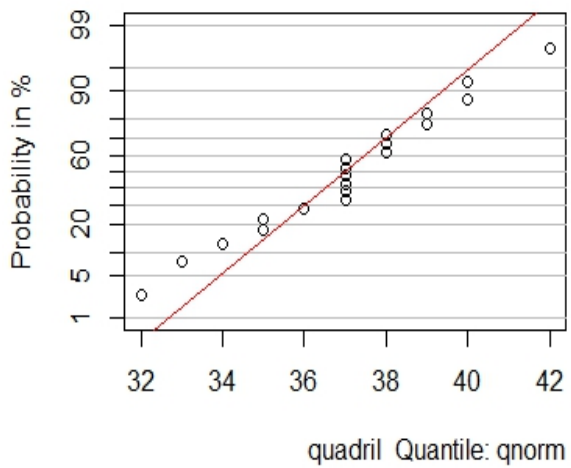




PP *plots* - exemplo das medidas







Testes de hipóteses univariados

Há vários testes de normalidade univariada na literatura:

- Shapiro-Wilk (poderoso)
- D'Agostino-Pearson
- Anderson-Darling
- Lilliefors
- outros

As hipóteses envolvidas são:

H_0 - os dados seguem distribuição normal

H_1 - os dados não seguem distribuição normal

Testes de hipóteses univariados

- Os programas estatísticos (como o R), geralmente, retornam o valor da estatística do teste e o valor-p.
- Se o valor-p for menor do que α (geralmente $0,05$), rejeitamos a hipótese de normalidade e os dados são considerados não normais.

b) Generalização dos QQ *plots* univariados (gráfico qui quadrado)

Quando n é grande, a variável

$$d_i^2 = (\mathbf{X}_{i\bullet} - \bar{\mathbf{X}})^T \mathbf{S}^{-1} (\mathbf{X}_{i\bullet} - \bar{\mathbf{X}}), \quad i = 1, \ldots, n \quad \sim \chi_p^2,$$

em que $\mathbf{X}_{i\bullet}$ representa os valores observados das p variáveis da i-ésima u.a., $\bar{\mathbf{X}}$ é o vetor de médias amostrais e $\mathbf{S}^{-1}$ é a inversa da matriz de covariâncias amostral.

b) Generalização dos QQ *plots* univariados (gráfico qui quadrado)

- Cada observação $\mathbf{X}_{i\bullet}$ é convertida em uma distância generalizada, dando uma medida da distância da observação ao vetor de médias, $d^2(i)$
- Se as observações vêm de uma normal multivariada, essas distâncias seguem, aproximadamente uma distribuição qui quadrado com p graus de liberdade (χ^2_p)
- Então, plotar as distâncias ordenadas *versus* os quantis da χ^2 deve levar a uma linha reta

Passos para obtenção do gráfico qui quadrado

1. Calcular as distâncias $d^2_{(i)}$ para todos os elementos da amostra
2. Ordenar os valores em ordem crescente:

$$d^2_{(1)} \leq d^2_{(2)} \leq \cdots \leq d^2_{(n)}.$$

   O valor $d^2_{(i)}$ representa a i-ésima observação ordenada
3. Plotar os pares $d^2_{(i)} \times \chi^2_p\left((i-1/2)/n\right)$, distâncias ordenadas $\times$ quantis χ^2_p, em que o valor $\chi^2_p\left((i-1/2)/n\right)$ corresponde a

$$P\left(\chi^2_p \leq \chi^2_p(i-1/2)/n\right) = (i-1/2)/n$$

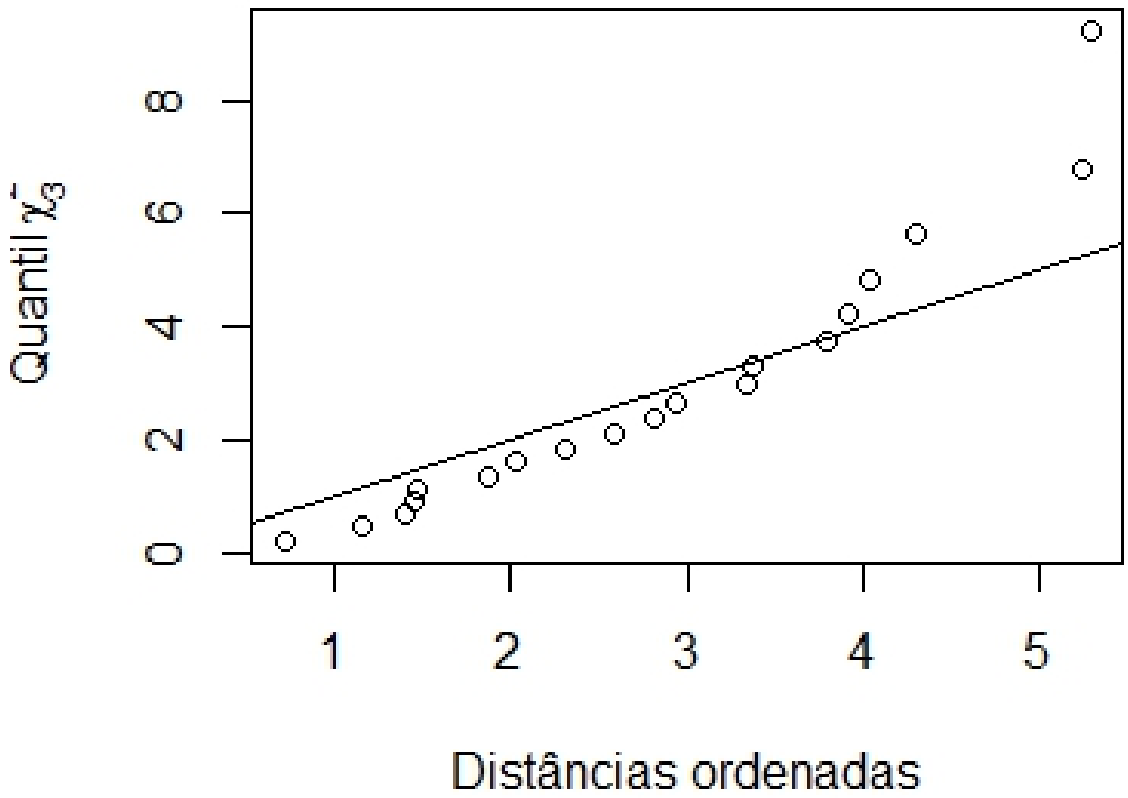
Exemplo: notas

As notas obtidas (de 0 a 25) em três provas de dezenove estudantes que participaram de uma disciplina foram:

Estudante	Nota 1	Nota 2	Nota 3
1	17,2	16,7	15,8
2	16,8	15,0	17,2
3	25,0	24,6	24,2
4	19,0	17,5	18,0
5	21,0	24,8	20,8
6	15,6	13,4	16,2
7	19,0	23,4	22,8
8	22,5	24,3	23,5
9	18,2	20,3	19,6
10	16,7	17,5	15,7
11	22,6	20,2	23,6
12	22,0	20,6	21,9
13	15,8	16,3	17,7
14	15,5	17,8	17,7
15	21,3	24,8	22,9
16	21,2	21,5	18,9
17	23,0	24,1	23,5
18	22,7	18,9	20,6
19	19,6	23,3	20,7

Valores obtidos - gráfico qui quadrado

i	X_1	X_2	X_3	d_i^2	$d_{(i)}^2$	$(i-1/2)/19$	χ_p^2
1	17,2	16,7	15,8	2,9280231	0,7155708	0,02631579	0,2236487
2	16,8	15,0	17,2	2,2989560	1,1584554	0,07894737	0,4901381
3	25,0	24,6	24,2	3,3742355	1,4000788	0,13157895	0,7202918
4	19,0	17,5	18,0	1,1584554	1,4579295	0,18421053	0,9399017
5	21,0	24,8	20,8	3,7939956	1,4698020	0,23684211	1,1577355
6	15,6	13,4	16,2	3,9050700	1,8697500	0,28947368	1,3787373
7	19,0	23,4	22,8	5,2346031	2,0334495	0,34210526	1,6065978
8	22,5	24,3	23,5	1,4698020	2,2989560	0,39473684	1,8446684
9	18,2	20,3	19,6	0,7155708	2,5860725	0,44736842	2,0964486
10	16,7	17,5	15,7	3,3301585	2,8196778	0,50000000	2,3659739
11	22,6	20,2	23,6	5,2837413	2,9280231	0,55263158	2,6582531
12	22,0	20,6	21,9	1,4000788	3,3301585	0,60526316	2,9798855
13	15,8	16,3	17,7	2,5860725	3,3742355	0,65789474	3,3400642
14	15,5	17,8	17,7	2,8196778	3,7939956	0,71052632	3,7523821
15	21,3	24,8	22,9	2,0334495	3,9050700	0,76315789	4,2384327
16	21,2	21,5	18,9	4,0397065	4,0397065	0,81578947	4,8359616
17	23,0	24,1	23,5	1,4579295	4,3007239	0,86842105	5,6209826
18	22,7	18,9	20,6	4,3007239	5,2346031	0,92105263	6,7886730
19	19,6	23,3	20,7	1,8697500	5,2837413	0,97368421	9,2357046


Quantil χ^2_3
0
2
4
6
8
1
2
3
4
5
Distâncias ordenadas

c) Testes de hipóteses multivariados

- Proposto por Mardia (1970): fundamentado nos coeficientes de assimetria e curtose da normal multivariada
- Shapiro-Wilk de Royston: generalização do teste de Shapiro-Wilk univariado
- Ver Ferreira (2008) para detalhes