MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Semestre IME/2019

Dados Multivariados

Banco de Dados:

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

$Y_{n\times p} = \left(y_{ij} \right)$: Matriz de Dados

↑ resposta do i-ésimo “indivíduo” na j-ésima variável

Espaço das unidades amostrais (indivíduos): Linhas de Y
Espaço das variáveis: Colunas de Y

Vetor de Variáveis Aleatórias Multidimensionais

Seja $Y_i = \left(Y_{i1}, Y_{i2}, ..., Y_{ip}\right)' \in \Re^p$ uma observação (multivariada) extraída de uma população com função densidade de probabilidades $f_Y(y \mid \theta)$, tal que a função de distribuição acumulada é dada por:

Y absolutamente contínuo

$$F_{Y_i}(y) = P\left(Y_{i1} \le y_1, ..., Y_{ip} \le y_p\right) = \int_{-\infty}^{y_1} ... \int_{-\infty}^{y_p} f_{Y_i}\left(y\right) dy_1 ... dy_p ; \qquad \int_{-\infty}^{\infty} f_{Y_i}\left(y\right) dy = 1$$

Espaço amostral

↑ sinal de integração p-dimensional

Note que: $Y_i : \Omega_Y \to \Re^p, \quad F : \Re^p \to [0,1]$

Seja: E: experimento aleatório

Ω: conjunto amostral de possíveis resultados de E

A: álgebra de subconjuntos de Ω, P: medida de probabilidade

Seja $Y_{n \times p} = \left(Y_1, Y_2, ..., Y_n\right)'$, $Y_i \in \Re^p$, uma amostra aleatória simples de n observações

de $f_Y(y \mid \theta)$ definidas no mesmo espaço de probabilidade (Ω, A, P). Neste caso:

$$F_Y(y) = \prod_{i=1}^{n} F_{Y_i}\left(y_i\right) = \prod_{i=1}^{n} P\left(Y_{i1} \le y_{i1}, ..., Y_{ip} \le y_{ip}\right)$$

↑

Variáveis Aleatórias Multidimensionais

- Vetor aleatório de variáveis contínuas: $Y = (Y_1, Y_2, ..., Y_p)' \in \Re^p$

Função de distribuição acumulada conjunta dada por: $F_Y(y)$

As variáveis $Y_1, Y_2, ..., Y_p$ são independentes, se e somente se, para qualquer $y \in \Re^p$

Função de distribuição acumulada de Y_j

$F_Y(y) = \prod_{j=1}^{p} F_{Y_j}(y_j)$. Analogamente, se e somente se, $f_Y(y) = \prod_{j=1}^{p} f_{Y_j}(y_j)$

- Seja $(Y_1, Y_2, ..., Y_q)' \in \Re^q$, $q < p$, um subconjunto de $Y \in \Re^p$. A distribuição marginal de do subconjunto de q variáveis dentre as p é dada por:

$$f_{(Y_1,...,Y_q)}(y_1,...,y_q) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_Y(y_1,...,y_q, y_{q+1},...y_p) dy_{q+1}...dy_p$$

- Seja $Y \in \Re^p$ tal que, $Y = (Y_1, ..., Y_p)' = (X, Z)'$ com *X* e *Z* partições de *Y*. Então, para qualquer $Z \in \Re^{p-q}$, com $f_Z(z) > 0$, a densidade condicional *q*-dimensional de X (q<p), quando *Z=z,* é dada por:

$$f_{X|Z}(x \mid z) = \frac{f_Y(y)}{f_Z(z)}; \quad x \in \Re^q$$

Vetor de Variáveis Aleatórias Multidimensionais

- Vetor aleatório da i-ésima observação: $Y_i = \left(Y_{i1}, Y_{i2}, ..., Y_{ip}\right)' \in \Re^p, i = 1,2,...,n$

- Matriz de n-observações p-dimensionais: $Y_{n\times p} = \left(Y_1, Y_2, ..., Y_n\right)' \in \Re^{n\times p}$

Amostra aleatória simples de *n* vetores aleatórios *p*-dimensionais (*AASn*):

$$f_Y(y) = \prod_{i=1}^{n} f_{Y_i}(y_i); \quad Y_i \in \Re^p$$

↑ distribuição multivariada

densidade conjunta: suposição de observações independentes

$$f_Y(y) = \prod_{i=1}^{n}\prod_{j=1}^{p} f_{Y_{ij}}(y_{ij})$$

↑ Função de densidade univariada

densidade conjunta: suposição de observações independentes avaliadas em p variáveis independentes

Variáveis Aleatórias Multidimensionais

Seja o vetor aleatório $Y = (Y_1, Y_2, ..., Y_p)' \in \Re^p$; $Y \sim (\mu; \Sigma)$.

- $E(Y) = \mu_{p\times 1} = (\mu_1, ..., \mu_p)' \in \Re^p$ é o vetor de médias populacional de *Y*, tal que:

.

$$\mu_j = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} y_j f_Y(y) dy_1 ... dy_p; \quad j = 1,2,..., p$$

Propriedade de linearidade: $E(AY + b) = AE(Y) + b$

- $E\left[(Y-\mu)(Y-\mu)'\right] = Cov(Y) = \Sigma_{p\times p}$ é a matriz de covariância populacional de Y.

Propriedades:

$\Sigma \geq 0$

$\Sigma = E(YY') - \mu\mu'$

$Cov(Y, X) = E\left[(Y - E(Y))(X - E(X))'\right]$

$Cov(X, Y) = Cov(Y, X)'$

$Cov(X + Y) = Cov(X) + Cov(Y) + 2Cov(X, Y)$

$Cov(a'Y) = a'\Sigma a$; a um vetor constante

$Cov(AY + b) = A\Sigma A'$; A matriz constante

$Cov(X + Z, Y) = Cov(X, Y) + Cov(Z, Y)$

Variáveis Aleatórias Multidimensionais

- Matriz aleatória (Gupta and Nagar, 2000):

Formulações alternativas

$$Y_{n\times p} = \left(Y_{ij}\right) \in \Re^{n\times p}; \quad Y_{n\times p} \sim N_{n,p}\left(M; \Psi \otimes \Sigma\right); \quad vec\left(Y\right)_{np\times 1} \sim N_{np}\left(vec(M); \Psi \otimes \Sigma\right)$$

$M_{n\times p} = 1_n \mu'_{p\times 1}$: matriz de médias

$vec\left(M\right)_{np\times 1} = 1_n \otimes \mu_{p\times 1}{}'$: vetor de médias de n observações em p variáveis

$\left(\Psi_{n\times n} \otimes \Sigma_{p\times p}\right)_{np\times np}$: matriz de covariâncias

Formulação flexível para diferentes modelagens

Matrizes de covariância Estruturadas: entre indivíduos (Ψ) e entre variáveis (Σ)

$\Psi = I_n; \quad \Sigma = I_p$

Observações e variáveis independentes

$\Psi = I_n; \quad \Sigma = (1-\alpha) I_p + \alpha 1_p 1_p{}'$

Observações independentes e correlação uniforme entre as variáveis

$\Psi = I_G \oplus \left[(1-\alpha) I_{n_g} + \alpha 1_{n_g} 1_{n_g}{}'\right]; \quad \Sigma = \left(\sigma_{jl}\right)$

Correlação uniforme entre observações agrupadas em G grupos

Correlação não estruturada entre variáveis

Distribuição Normal Multivariada

- Variável (escalar) Y $\in \Re$, com distribuição Normal univariada de média μ e variância σ^2 tem densidade dada por:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(y-\mu)/\sigma]^2/2} \quad -\infty < y < \infty, \quad \sigma^2 > 0$$

Notação: $Y \sim N_1(\mu; \sigma^2)$

Generalização multivariada para o vetor aleatório $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{ip})' \in \Re^p$

Distribuição Normal Multivariada com vetor de média μ e matriz de covariância Σ:

$$Y_{i\,p\times 1} \sim N_p(\mu_{p\times 1}; \Sigma_{p\times p}); \qquad f_{Y_i}(y) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(y-\mu)'\Sigma^{-1}(y-\mu)/2} \quad |\Sigma| > 0,\ y \in \Re^p,\ i = 1, ..., n$$

Mostre que: $\int_{\Re^p} f_Y(y)dy_1...dy_p = 1$

$$d_M^2 = (y-\mu)'\Sigma^{-1}(y-\mu) = c^2$$

a densidade é constante em superfícies onde a distância de Mahalanobis é constante.

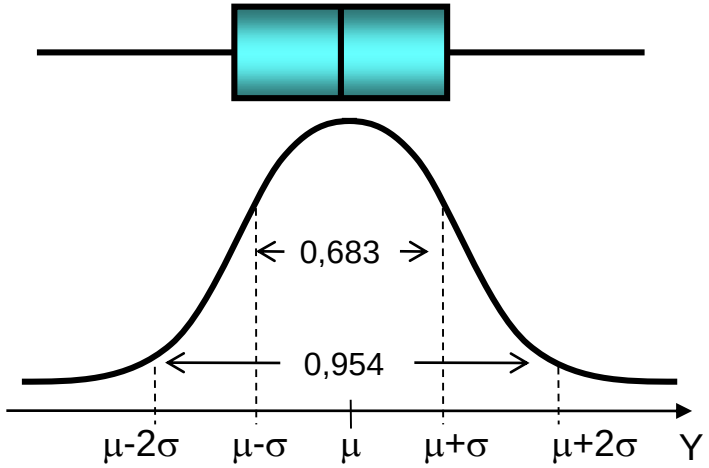
Use: $\Sigma^{-1/2}(Y_i - \mu) \sim N_p(0; I)$

Distribuição Normal Uni e Multivariada

Normal Univariada:

$$Y \sim N_1\left(\mu, \sigma^2\right)$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(y-\mu)/\sigma]^2/2} \qquad y \in \Re, \sigma^2 > 0$$



Normal Multivariada Bidimensional

$$Y_{p\times 2} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2\left(\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right);$$

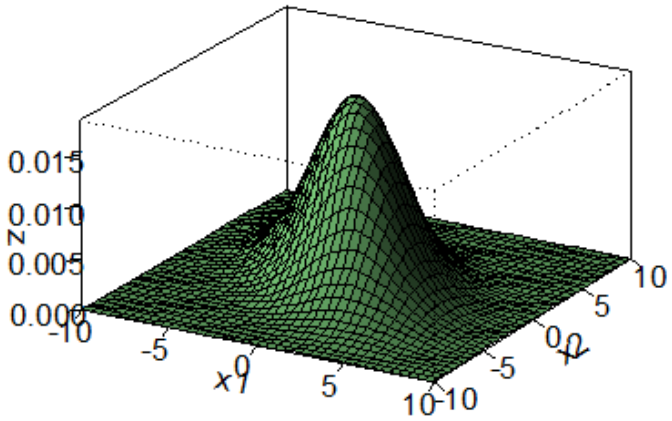
$$y \in \Re^2, |\Sigma| > 0, \sigma_{12} = \rho\sigma_1\sigma_2$$

$$f_Y(y) = \frac{1}{2\pi\sigma_1\sigma_2\left(1-\rho^2\right)^{1/2}}$$

$$\exp\left\{ \frac{-1}{2\left(1-\rho^2\right)} \left[\left(\frac{Y_1-\mu_1}{\sigma_1} \right)^2 + \left(\frac{Y_2-\mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{Y_1-\mu_1}{\sigma_1} \right) \left(\frac{Y_1-\mu_1}{\sigma_1} \right) \right] \right\}$$

Two dimensional Normal Distribution

$\mu_1 = 0, \mu_2 = 0, \sigma_{11} = 10, \sigma_{22} = 10, \sigma_{12} = 5, \rho = 0.5$



Distribuição Normal Multivariada

$$Y_{p\times 1} \sim N_p\left(\mu_{p\times 1}; \Sigma_{p\times p}\right); \qquad f_Y(y) = \frac{1}{(2\pi)^{p/2}\left|\Sigma\right|^{1/2}} e^{-(y-\mu)'\Sigma^{-1}(y-\mu)/2} \qquad y \in \Re^p; \left|\Sigma\right| > 0$$

- Definição por caracterização:

  Teorema de Cramér-Wold: A distribuição de um vetor aleatório $Y \in \Re^p$ está completamente determinada pelo conjunto de todas as distribuições uni-dimensionais de combinações lineares $a'Y$, com $a \in \Re^p$.

  $$Y_{p\times 1} \sim N_p\left(\mu_{p\times 1}; \Sigma_{p\times p}\right) \Leftrightarrow a'Y \sim N_1\left(a'\mu; a'\Sigma a\right)$$

- Função Geradora de Momentos da Normal Multivariada:

$$M_Y(t) = E\left(e^{t'Y}\right) = e^{t'\mu + \frac{1}{2}t'\Sigma t}$$

- Função Característica da Normal Multivariada:

$$\phi_Y(t) = E\left(e^{it'Y}\right) = E\left(e^{iX}\right) = \phi_X(1) = e^{it'\mu - \frac{1}{2}t'\Sigma t}, \quad t \in \Re^p, \quad X = t'Y$$

Distribuição Normal Multivariada

$$Y_{p\times 1} \sim N_p\left(\mu_{p\times 1}; \Sigma_{p\times p}\right); \qquad f_Y(y) = \frac{1}{(2\pi)^{p/2}\left|\Sigma\right|^{1/2}} e^{-(y-\mu)'\Sigma^{-1}(y-\mu)/2} \qquad y \in \Re^p; \left|\Sigma\right| > 0$$

Alguns Resultados:

- $Y_{p\times 1} \sim N_p\left(\mu_{p\times 1}; \Sigma_{p\times p}\right) \Rightarrow X = AY_{p\times 1} + b \sim N_q\left(A\mu + b\,;\, A\Sigma A'\right), \quad A_{q\times p} \in \Re^{q\times p}, b \in \Re^q$

- $Y_{p\times 1} \sim N_p\left(\mu_{p\times 1}; \Sigma_{p\times p}\right) \Rightarrow X = \Sigma^{-1/2}(Y-\mu) \sim N_p\left(0\,;\, I_p\right)$

  Distância populacional de Mahalanobis ao centroide em Y
  Distância Euclidiana em X

$$\Rightarrow \quad X'X = (Y-\mu)'\Sigma^{-1}(Y-\mu) = \sum_{j=1}^{p} X_j^2 \sim \chi_p^2$$

- $Y_{p\times 1} \sim N_p\left(\mu_{p\times 1}; \Sigma_{p\times p}\right)$; AY e BY são independentes $\Leftrightarrow A\Sigma B' = 0$

$Y_{p\times 1} \sim N_p\left(\mu_{p\times 1}; \Sigma = I_p\sigma^2\right)$, $G_{q\times p}$ é linha-ortonormal $\Rightarrow GY$ e $\left(\mathrm{I_p}\text{-}\mathrm{G'G}\right)Y$ são independentes

$$GG' = I_q$$

Distribuição Normal Multivariada

Alguns Resultados:

$$Y_{p\times 1}=\begin{pmatrix} Y_{1q\times 1} \\ Y_{2(p-q)\times 1}\end{pmatrix} \sim N_p\left(\mu_{p\times 1}=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}; \Sigma_{p\times p}=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\right).$$ Então:

- $Y_{1q\times 1} \sim N_q(\mu_1;\Sigma_{11})$; $\quad Y_{2(p-q)\times 1} \sim N_{(p-q)}(\mu_2;\Sigma_{22})$

distribuições marginais de Y são Normais

- Y_1 e $Y_{2.1}=(Y_2-\Sigma_{21}\Sigma_{11}^{-1}Y_1)$ são independentes, tal que,

$$Y_{2.1} \sim N_{p-q}\left(\mu_2-\Sigma_{21}\Sigma_{11}^{-1}\mu_1\,;\,\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

$$Y_1=\begin{bmatrix} I_q & 0\end{bmatrix}Y=AY;$$
$$Y_{2.1}=\begin{bmatrix}-\Sigma_{21}\Sigma_{11}^{-1} & I_{p-q}\end{bmatrix}\mathrm{Y}=\mathrm{BY};$$
$$\mathrm{A\Sigma B}=0$$

- $Y_2 \mid Y_1 \sim N_{(p-q)}\left(\mu_2+\Sigma_{21}\Sigma_{11}^{-1}(Y_1-\mu_1);\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$

$$Y_2=Y_{2.1}+\Sigma_{21}\Sigma_{11}^{-1}Y_1$$

Condicionado em Y1, este termo é constante

distribuições condicionais de Y são Normais

Teorema 3.2.4
Exemplo 3.2.1: $\Sigma=(1-\rho)I_p+\rho 1_p 1_p'$
(Mardia et al., 2003)

Distribuição Normal Multivariada

Distribuições Condicionais - Aplicação:

$$(X,Y,Z)' \sim N_3\left(\mu_{3\times1}=\begin{pmatrix}\mu_X\\ \mu_Y\\ \mu_Z\end{pmatrix};\Sigma_{3\times3}=\begin{pmatrix}\sigma_{XX} & \sigma_{XY} & \sigma_{XZ}\\ \sigma_{XY} & \sigma_{YY} & \sigma_{YZ}\\ \sigma_{XZ} & \sigma_{YZ} & \sigma_{ZZ}\end{pmatrix}\right)$$

Distribuição conjunta de (X,Y,Z)

$$(X,Y)|\,Z=z \sim N_2\left(\begin{pmatrix}\mu_X+\sigma_{XZ}\sigma_{ZZ}^{-1}(z-\mu_Z)\\ \mu_Y+\sigma_{YZ}\sigma_{ZZ}^{-1}(z-\mu_Z)\end{pmatrix};\Sigma_{XY|Z}\right)$$

Distribuição condicional de (X,Y)|Z

$$\Sigma_{XY|Z}=\begin{pmatrix}\sigma_{XX} & \sigma_{XY}\\ \sigma_{YX} & \sigma_{YY}\end{pmatrix}-\begin{pmatrix}\sigma_{XZ}\\ \sigma_{YZ}\end{pmatrix}\sigma_{ZZ}^{-1}\begin{pmatrix}\sigma_{XZ} & \sigma_{YZ}\end{pmatrix}$$

(Edwards, D.;2000)

$$X-\mu_X+\sigma_{XZ}\sigma_{ZZ}^{-1}(z-\mu_Z)$$
$$Y-\mu_Y+\sigma_{YZ}\sigma_{ZZ}^{-1}(z-\mu_Z)$$

Erros de predição de X e de Y dado Z=z

Distribuição Normal Multivariada

Distribuições Condicionais - Aplicação: (Edwards, D.;2000)

$$(X,Y,Z)' \sim N_3\left(\mu_{3\times1}=\begin{pmatrix}\mu_X\\ \mu_Y\\ \mu_Z\end{pmatrix};\Sigma_{3\times3}=\begin{pmatrix}\sigma_{XX} & \sigma_{XY} & \sigma_{XZ}\\ \sigma_{XY} & \sigma_{YY} & \sigma_{YZ}\\ \sigma_{XZ} & \sigma_{YZ} & \sigma_{ZZ}\end{pmatrix}\right)$$

Distribuição conjunta de (X,Y,Z)

$$(X,Y)|\,Z=z \sim N_2\left(\begin{pmatrix}\mu_X+\sigma_{XZ}\sigma_{ZZ}^{-1}(z-\mu_Z)\\ \mu_Y+\sigma_{YZ}\sigma_{ZZ}^{-1}(z-\mu_Z)\end{pmatrix};\Sigma_{XY|Z}\right)$$

Distribuição condicional de (X,Y)|Z

$$\Sigma_{XY|Z}=\frac{1}{\omega_{XX}\omega_{YY}-\omega_{XY}^2}\begin{pmatrix}\omega_{XX} & -\omega_{XY}\\ -\omega_{YX} & \omega_{YY}\end{pmatrix}=\begin{pmatrix}\omega_{XX} & \omega_{XY}\\ \omega_{YX} & \omega_{YY}\end{pmatrix}^{-1}$$

$$\Rightarrow \rho_{XY|Z}=\frac{-\omega_{XY}}{\sqrt{\omega_{XX}}\sqrt{\omega_{YY}}}$$

Correlação parcial de (X,Y)|Z

$$\Sigma^{-1}=\begin{pmatrix}\sigma_{XX} & \sigma_{XY} & \sigma_{XZ}\\ \sigma_{XY} & \sigma_{YY} & \sigma_{YZ}\\ \sigma_{XZ} & \sigma_{YZ} & \sigma_{ZZ}\end{pmatrix}^{-1}=\begin{pmatrix}\omega_{XX} & \omega_{XY} & \omega_{XZ}\\ \omega_{XY} & \omega_{YY} & \omega_{YZ}\\ \omega_{XZ} & \omega_{YZ} & \omega_{ZZ}\end{pmatrix}$$

Matriz de precisão

Função de Verossimilhança

$Y = (Y_1, Y_2, ..., Y_n)' \in \Re^{n \times p}$: Amostra aleatória simples de tamanho n da $N_p(\mu;\Sigma)$

AASn da $N_p(\mu;\Sigma)$

$$L(\mu, \Sigma \mid Y) = \prod_{i=1}^{n} f_{Y_i}(y_i) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(y_i-\mu)'\Sigma^{-1}(y_i-\mu)/2}$$

$$= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right\}$$

Como, $\left\langle Y_i' A Y_i = tr\left(Y_i A Y_i'\right) = tr\left(A Y_i Y_i'\right);\quad \Sigma = (\sigma_{jl});\quad \Sigma^{-1} = (\omega_{jl}) \right\rangle$, a menos de um termo constante:

$$= \left|(\omega_{jl})\right|^{n/2} \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}\sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)'\right]\right\}$$

Como, $\left\langle \sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)' = \sum_{i=1}^{n}(y_i-\bar{Y})(y_i-\bar{Y})' + n(\bar{Y}-\mu)(\bar{Y}-\mu)' = nS + n(\bar{Y}-\mu)(\bar{Y}-\mu)' \right\rangle$, tem-se:

$$L(\mu, \Sigma \mid Y) = \left|\omega_{jl}\right|^{n/2} \exp\left\{-\frac{1}{2}\left[tr\left(\Sigma^{-1} nS\right) + n\left(\bar{Y}-\mu\right)'\Sigma^{-1}\left(\bar{Y}-\mu\right)\right]\right\}$$

Função de Verossimilhança

$$L(\mu,\Sigma\,|\,Y)=\left|\omega_{jl}\right|^{n/2}\exp\left\{-\frac{1}{2}\left[tr\left(\Sigma^{-1}nS\right)+n\left(\bar{Y}-\mu\right)'\Sigma^{-1}\left(\bar{Y}-\mu\right)\right]\right\}$$

Assim, o logaritmo de L, a menos de um termo constante, é:

$$\log L(\mu,\Sigma\,|\,Y)=\frac{n}{2}\log\left|\left(\omega_{jl}\right)\right|-\frac{1}{2}\sum_{j=1}^{p}\sum_{l=1}^{p}\left(\omega_{jl}ns_{jk}\right)-n\left(\bar{Y}-\mu\right)'\Sigma^{-1}\left(\bar{Y}-\mu\right) \qquad *(1)$$

Alternativamente:

$$\log L(\mu,\Sigma\,|\,Y)=\frac{n}{2}\log\left|\left(\omega_{jl}\right)\right|-\frac{1}{2}\sum_{j=1}^{p}\sum_{l=1}^{p}\omega_{jl}\left[ns_{jk}-n\left(\bar{Y}_j-\mu_j\right)'\left(\bar{Y}_l-\mu_l\right)\right] \qquad *(2)$$

$N_p(\mu,\Sigma)$ - Estimadores

Diferenciando *(1) com respeito a μ e igualando a zero, obtém-se o EMVS de μ:

$$\Sigma^{-1}\left(\bar{Y}-\mu\right)=0 \quad \Rightarrow \quad \hat{\mu}=\bar{Y}_{p\times 1}; \quad \hat{\mu}_j=\bar{Y}_j, \quad j=1,2,\ldots,p$$

Diferenciando *(2) com respeito a ω_{jl} e igualando a zero, obtém-se o EMVS de Σ:

$$n\left(\omega_{jl}^{-1}\right)-ns_{jk}=0 \quad \Rightarrow \quad \hat{\sigma}_{jl}=s_{jk} \quad \Rightarrow \quad \hat{\Sigma}=S$$

Mostre que estes estimadores maximizam logL:

$$\log L\left(\hat{\mu},\hat{\Sigma}\mid Y\right)-\log L\left(\mu,\Sigma\mid Y\right)\geq 0$$

Ainda, como, $f_{Y_i}\left(y_i\mid\mu,\Sigma\right)=\left|\left(2\pi\right)\right|^{-p/2}\left|\Sigma\right|^{-1/2}\exp\left\{-\frac{1}{2}tr\Sigma^{-1}\left[S+\left(\bar{Y}-\mu\right)\left(\bar{Y}-\mu\right)'\right]\right\}$

$\bar{Y}_{p\times 1}$ e $S_{p\times p}$ são estatísticas conjuntamente suficientes para μ e Σ, respectivamente.

Os estimadores não viciados de μ e Σ são, respectivamente:

$$\bar{Y}_{p\times 1} \quad \text{e} \quad S_{u\,p\times p}$$

Distribuição Amostral

$Y = (Y_1, Y_2, ..., Y_n)' \in \Re^{n \times p}$ é *AASn* da $N_p(\mu;\Sigma)$

$$Y_{n\times p} = (Y_{ij}) \in \Re^{n\times p}; \qquad Y \sim N_{n,p}(1_n \mu\,; I_n \otimes \Sigma); \qquad vec(Y)_{np\times 1} \sim N_{np}(1_n \otimes \mu\,; I_n \otimes \Sigma)$$

Distribuição amostral de $\bar{Y} = \frac{1}{n} Y' 1_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$; $Y_i \in \Re^p$

$$Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu\,;\Sigma) \quad \Rightarrow \quad \bar{Y} \sim N_p\left(\mu\,; \frac{1}{n}\Sigma\right)$$

Ganho em precisão

Distribuição Amostral

$Y = (Y_1, Y_2, ..., Y_n)' \in \Re^{n \times p}$ é *AASn* da $N_p(\mu;\Sigma)$

Distribuição amostral de $S_{p\times p} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})' = \frac{1}{n} Y'HY$

Definição: Se $M_{p\times p} = X'X; \quad X_{n\times p} \sim N_{n,p}(0_{n\times p}; I_n \otimes \Sigma)$. Então *M* tem distribuição de Wishart com *n* graus de liberdade e matriz de escala Σ, tal que: $M_{p\times p} \sim W_p(n;\Sigma)$

→ $W_1(n;\sigma^2) = \sigma^2 \chi^2_n$

Teorema: Se $Y_{n\times p} \sim N_{n,p}(1_n \mu; I_n \otimes \Sigma)$ e $C_{n\times n}$ é matriz simétrica, idempotente com as linhas somando 0, então, $Y'CY \sim W_p(r;\Sigma); \quad r = trC$

$nS_{p\times p} \sim W_p(n-1;\Sigma)$

H=H', H=HH', trH=n-1

Distribuição Amostral

Teorema: Se $Y_{n\times p} \sim N_{n,p}\left(1_n \mu\,; I_n \otimes \Sigma\right)$. Então, $X = AYB$ e $Z = CYD$ são independentes se e somente se $B'\Sigma D = 0$ ou $AC' = 0$.

$$\bar{Y} \sim N_p\left(\mu\,; \frac{1}{n}\Sigma\right), \qquad nS_{p\times p} \sim W_p\left(n-1; \Sigma\right)$$

$$\bar{Y} = \frac{1}{n} Y'1_n \quad \Rightarrow \quad \bar{Y}' = \frac{1}{n} 1_n' Y$$

$$S = \frac{1}{n} Y'HY\,; \qquad H = I_n - \frac{1}{n} 1_n 1_n'$$

$$\left(\frac{1}{n} 1_n'\right) H = 0$$

Logo:

$\bar{Y}$ e S são independentes.

Distâncias de Mahalanobis

Definição: Considere $\delta \sim N_p(0;\Sigma)$, $M \sim W_p(n;\Sigma)$, com δ e M variáveis independentes.

Então: $n\delta' M^{-1}\delta \sim T^2(p;n)$ Distribuição T^2 de Hotelling

$$Y_{i\,p\times 1} \sim N_p(\mu;\Sigma)$$

$$\bar{Y} \sim N_p(\mu;\Sigma/n); \quad \sqrt{n}(\bar{Y}-\mu) \sim N_p(0;\Sigma)$$

$$nS \sim W_p(n-1;\Sigma)$$

$$(n-1)S_u \sim W_p(n-1;\Sigma)$$

Distância generalizada

$$(n-1)(\bar{Y}-\mu)' S^{-1}(\bar{Y}-\mu) \sim T^2(p;n-1)$$

$$n(\bar{Y}-\mu)' S_u^{-1}(\bar{Y}-\mu) \sim T^2(p;n-1)$$

Teorema 3.5.2
(Mardia et al., 2003) $T^2(p;n-1) = \dfrac{(n-1)p}{n-p} F_{(p;n-p)}$

Regiões de Confiança para $\mu \in \Re^p$

Inferências sobre o Vetor de Médias μ

$$Y_{i\,p\times 1} \overset{iid}{\sim} N_p\left(\mu_{p\times 1}; \Sigma_{p\times p}\right); \qquad f_{Y_i}(y) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(y-\mu)'\Sigma^{-1}(y-\mu)/2} \qquad |\Sigma| > 0,\ y \in \Re^p\ , i = 1,\ldots,n$$

Uma Região de Confiança R(Y) para o vetor de médias da $N_p(\mu;\Sigma)$ é uma região de valores prováveis de $\mu \in \Re^p$, com base na amostra.

Antes da amostra ser coletada, esta região deve satisfazer:

$$P\left(\mu \in \Re^p;\quad \mu \subset R(Y)\right) = 1-\alpha$$

Uma construção (natural) desta região, com base na amostra, é usar medidas de distância tais como:

$$n\left(\bar{Y}-\mu\right)' S_u^{-1} \left(\bar{Y}-\mu\right) \quad \leq \quad c^2$$

São todos os valores de μ "próximos" à evidência amostral $\bar{Y}$

Regiões de Confiança para $\mu \in \Re^p$

Amostra aleatória da Distribuição Normal Multivariada:

$$Y_{n\times p}: \quad AAS_n \quad N_p(\mu;\Sigma); \quad |\Sigma| > 0$$

$$\bar{Y}_{p\times 1} \sim N_p(\mu;\Sigma/n); \quad \sqrt{n}(\bar{Y}-\mu) \sim N_p(0;\Sigma); \quad (n-1)S_u \sim W_p(n-1;\Sigma)$$

$$\Rightarrow d_M^2 = n(\bar{Y}-\mu)' S_u^{-1}(\bar{Y}-\mu) \sim T^2(p;n-1) = \frac{(n-1)p}{n-p} F_{(p;n-p)}$$

$$(\bar{Y}-\mu)'(S_u/n)^{-1}(\bar{Y}-\mu)$$

Região (elipsóides) de Confiança para o Centróide da população

$$P\left(\bar{Y} \in \Re^p;\ d_M^2(\bar{Y};\mu) = (\bar{Y}-\mu)'(S/n)^{-1}(\bar{Y}-\mu) \le c^2\right) = 1-\alpha$$

Regiões de Confiança para $\mu \in \Re^p$

Inferências sobre o vetor μ

$$R(Y) = \left\{ \mu \in \Re^p ;\ n\left(\bar{Y} - \mu\right)' S_u^{-1} \left(\bar{Y} - \mu\right) \ \leq \ c^2 = \frac{(n-1)\,p}{(n-p)} F_{p,(n-p)}(\alpha) \right\}$$

$\Rightarrow$ Para determinar se algum ponto $\boldsymbol{\mu}_0$ cai na região R(**Y**) basta calcular a distância generalizada ao quadrado e compará-la com o valor crítico dado em função da distribuição F e do nível de significância α, isto é,

$$n\left(\bar{\mathbf{Y}} - \boldsymbol{\mu}_0\right)' \mathbf{S}_u^{-1} \left(\bar{\mathbf{Y}} - \boldsymbol{\mu}_0\right) \ \leq \ c^2 = \frac{(n-1)\,p}{(n-p)} F_{p,(n-p)}(\alpha)$$

$\Rightarrow$ Regiões de Confiança correspondem a Regiões de Aceitação em testes de hipóteses sobre o vetor μ.

Regiões de Confiança para $\mu \in \Re^p$

Inferências sobre o vetor μ

$$R(Y) = \left\{ \mu \in \Re^p ;\ n\left(\bar{Y} - \mu\right)' S_u^{-1} \left(\bar{Y} - \mu\right) \ \leq \ c^2 = \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \right\}$$

$\Rightarrow$ Para determinar se algum ponto $\boldsymbol{\mu}_0$ cai na região R(**Y**) basta calcular a distância generalizada ao quadrado e compará-la com o valor crítico dado em função da distribuição F e do nível de significância α, isto é,

$$n\left(\bar{\mathbf{Y}} - \boldsymbol{\mu}_0\right)' \mathbf{S}_u^{-1} \left(\bar{\mathbf{Y}} - \boldsymbol{\mu}_0\right) \ \leq \ c^2 = \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha)$$

$\Rightarrow$ Regiões de Confiança correspondem a Regiões de Aceitação em testes de hipóteses sobre o vetor μ.

Regiões de Confiança e Testes de Hipóteses

Taxas de açucar, sódio e potássio sanguíneas em 20 mulheres adultas

Indiv.	Açucar	Sódio	Potássio
1	3,7	48,5	9,3
2	5,7	65,1	8
3	3,8	47,2	10,9
4	3,2	53,2	12
5	3,1	55,5	9,7
6	4,6	36,1	7,9
7	2,4	24,8	14
8	7,2	33,1	7,6
9	6,7	47,4	8,5
10	5,4	54,1	11,3
11	3,9	36,9	12,7
12	4,5	58,8	12,3
13	3,5	27,8	9,8
14	4,5	40,2	8,4
15	1,5	13,5	10,1
16	8,5	56,4	7,1
17	4,5	71,6	8,2
18	6,5	52,8	10,9
19	4,1	44,1	11,2
20	5,5	40,9	9,4
Média	4,64	45,4	9,97
S	2,879		
	10,002	199,798	
	-1,81	-5,627	3,628

$$R(Y)=\left\{\mu;\, n\left(\bar{Y}-\mu\right)' S^{-1}\left(\bar{Y}-\mu\right)\leq \underbrace{\frac{(20-1)p}{(20-3)}F_{3,(17)}(\alpha)}\right\}$$

$$\alpha=0{,}10 \Rightarrow 8{,}18$$

$$\alpha=0{,}05 \Rightarrow 10{,}72$$

Suponha o interesse na seguinte hipótese:

$$H_0:\mu=\mu_0=(4,50,10)'$$

$$T^2=n\left(\bar{Y}-\mu_0\right)' S^{-1}\left(\bar{Y}-\mu_0\right)=9{,}74$$

Conclusão: $\alpha=0{,}10 \Rightarrow T^2 \notin R(Y);\quad rej\ H_0$

$\alpha=0{,}05 \Rightarrow T^2 \in R(Y);\quad não\ rej\ H_0$

Regiões de Confiança para o Vetor μ
Uma Única População

Morfometria cefálica para os dois primeiros filhos de 25 famílias (Everitt, 2007)

	1° Filho		2° Filho	
Família	Comprimento	Perímetro	Comprimento	Perímetro
1	191	155	179	145
2	195	149	201	152
3	181	148	185	149
4	183	153	188	149
5	176	144	171	142
6	208	157	192	152
7	189	150	190	149
8	197	159	189	152
9	188	152	197	159
10	192	150	187	151
11	179	158	186	148
12	183	147	174	147
13	174	150	185	152
14	190	159	195	157
15	188	151	187	158
16	163	137	161	130
17	195	155	183	158
18	186	153	173	148
19	181	145	182	146
20	175	140	165	137
21	192	154	185	152
22	174	143	178	147
23	176	139	176	143
24	197	167	200	158
25	190	163	187	150

$$Y_{25\times 4} = \left(Y_1,\ldots,Y_{25}\right)'; \; Y_{i_{4\times 1}} \overset{iid}{\sim} N_4\left(\mu;\Sigma\right)$$

Estatísticas Descritivas:

$$\bar{Y} = \begin{pmatrix} 185{,}72 & 151{,}12 & 183{,}84 & 149{,}24 \end{pmatrix}'$$

$$S_u = \begin{pmatrix} 91{,}481 & 50{,}753 & 66{,}875 & 44{,}267 \\ & 52{,}186 & 49{,}259 & 33{,}651 \\ & & 96{,}775 & 54{,}278 \\ & & & 43{,}222 \end{pmatrix}$$

Regiões de Confiança para o Vetor μ
Uma Única População

Morfometria cefálica para os dois primeiros filhos de 25 famílias (Everitt, 2007)

	1° Filho		2° Filho	
Família	Comprimento	Perímetro	Comprimento	Perímetro
1	191	155	179	145
2	195	149	201	152
3	181	148	185	149
4	183	153	188	149
5	176	144	171	142
6	208	157	192	152
7	189	150	190	149
8	197	159	189	152
9	188	152	197	159
10	192	150	187	151
11	179	158	186	148
12	183	147	174	147
13	174	150	185	152
14	190	159	195	157
15	188	151	187	158
16	163	137	161	130
17	195	155	183	158
18	186	153	173	148
19	181	145	182	146
20	175	140	165	137
21	192	154	185	152
22	174	143	178	147
23	176	139	176	143
24	197	167	200	158
25	190	163	187	150

$$Y_{25\times 4} = \left(Y_1,\ldots,Y_{25}\right)'; \; Y_{i_{4\times 1}} \overset{iid}{\sim} N_4\left(\mu;\Sigma\right)$$

Distribuição marginal:

$$Y_{n\times 2}; Y_{i_{2\times 1}} = \left(Y_{i1}, Y_{i3}\right) \overset{iid}{\sim} N_2\left(\mu = \begin{pmatrix}\mu_1\\ \mu_3\end{pmatrix}; \Sigma = \begin{pmatrix}\sigma_{11} & \sigma_{13}\\ \sigma_{13} & \sigma_{33}\end{pmatrix}\right)$$

$$R(Y) = \left\{\mu \in \Re^2; T^2 \leq \frac{(25-1)2}{(25-2)} F_{2;23}(\alpha)\right\}$$

$$\alpha = 0{,}10 \Rightarrow 5{,}3196$$

Suponha o interesse na seguinte hipótese:

$$\left\{ H_0 : \mu = \begin{pmatrix}182\\ 182\end{pmatrix} \right. \qquad T^2 = 4{,}186$$

Conclusão?

Intervalos de Confiança - Regiões de Confiança

(Everitt, 2007)

Caso univariado

$$I.C(\mu_k) \text{ a } 100(1-\alpha)\%$$

$$= \left(\bar{Y}_k \mp t_{n-1}(\alpha/2)\sqrt{\frac{s_{kk}}{n}} \right)$$

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

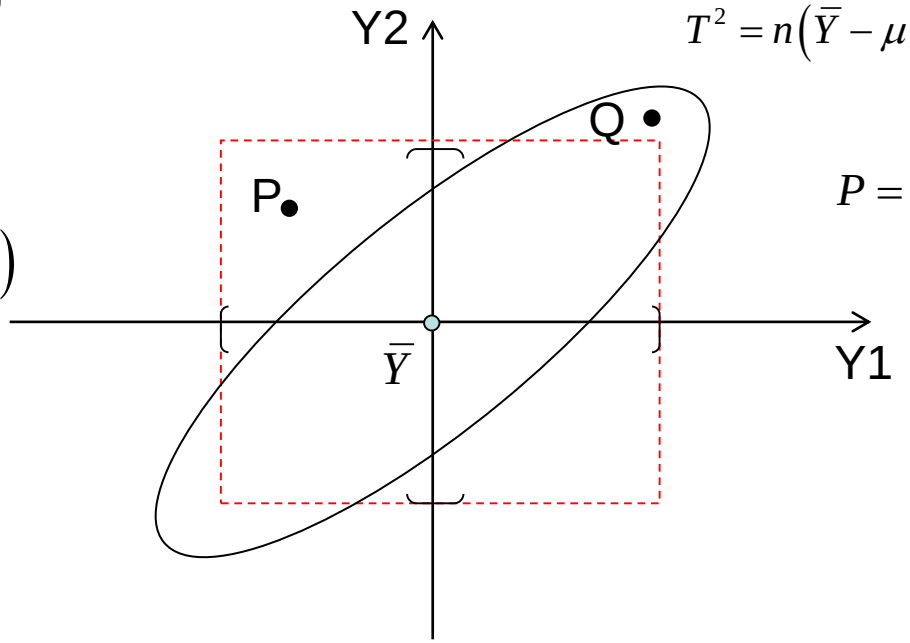
$$= n(\bar{Y}-\mu)(s^2)^{-1}(\bar{Y}-\mu)$$

$$t^2_{(n-1)} = F_{1,(n-1)}$$

Caso multivariado

$$R(Y) = \left\{ \mu \in \Re^p;\ n(\bar{Y}-\mu)' S_u^{-1} (\bar{Y}-\mu) \le c^2 = \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \right\}$$

$$T^2 = n(\bar{Y}-\mu)' S_u^{-1} (\bar{Y}-\mu) \sim \frac{(n-1)p}{n-p} F_{(p;n-p)}$$



$$P = (\mu_{P1}, \mu_{P2}); \quad Q = (\mu_{Q1}, \mu_{Q2})$$

⇒ Com base na evidência amostral em R(Y), valores de $\mu_0 \in \Re^2$ iguais a Q (P) estão na região de aceitação (rejeição) de possíveis valores do parâmetro μ.

⇒ Os Intervalos de Confiança univariados podem dar decisões diferentes da região de confiança.

⇒ Como representar no gráfico a elipse de concentração de pontos amostrais?

Distâncias de Mahalanobis

$$Y_{i\,p\times 1} \overset{iid}{\sim} N_p(\mu;\Sigma); \quad \sqrt{n}(\bar{Y}-\mu) \sim N_p(0;\Sigma); \quad (n-1)S_u \sim W_p(n-1;\Sigma)$$

$$n(\bar{Y}-\mu)' S_u^{-1}(\bar{Y}-\mu) \sim T^2(p;n-1)$$

Distância generalizada

Teorema 3.5.2 (Mardia et al., 2003): $T^2(p;n-1) = \dfrac{(n-1)p}{n-p} F_{(p;n-p)}$

Resultado: (Johnson and Whichern, 2008)

$Y_{i\,p\times 1} \in \Re^p, i = 1,2,...,n$ é AASn tal que, $E(Y_i) = \mu, \quad Cov(Y_i) = \Sigma, \ |\Sigma| > 0.$

Então, para *(n-p)* suficientemente grande,

$$n(\bar{Y}-\mu)' S_u^{-1}(\bar{Y}-\mu) \sim \chi^2_p$$

Região de Confiança para μ

Elipsóide de Confiança

Distâncias de Mahalanobis

$$Y_{i\,p\times 1} \overset{iid}{\sim} N_p(\mu;\Sigma) \quad \begin{cases} (Y_i-\mu)\sim N_p(0;\Sigma) \\ (Y_i-\mu)'\,\Sigma^{-1}(Y_i-\mu)\sim \chi^2_p \end{cases}$$

$$P\left(Y_i\in\Re^p;\ d^2_M(Y_i;\mu)=(Y_i-\mu)'\,\Sigma^{-1}(Y_i-\mu)\le c^2\right)=1-\alpha$$

$$c^2=\chi^2_p(\alpha)$$

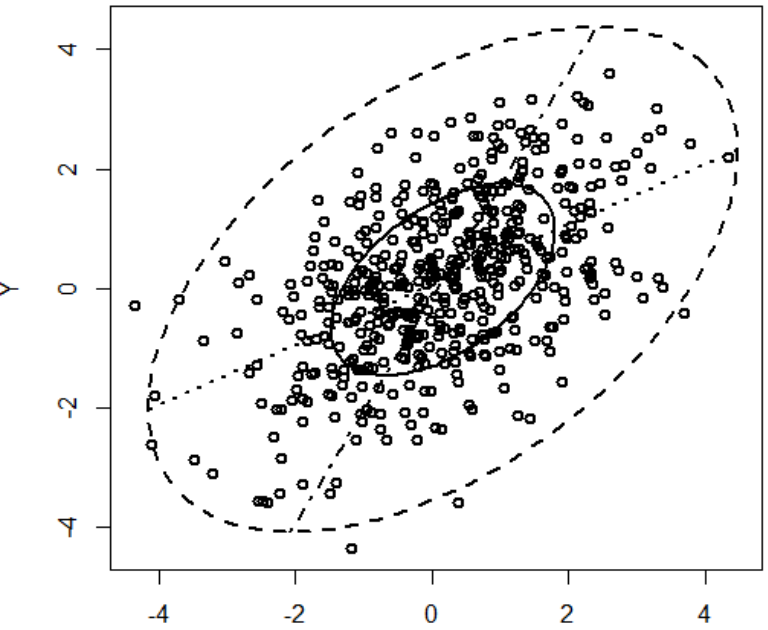
Resultado: (Johnson and Wichern, 2008). Sob Normalidade dos dados, para *(n-p)* suficientemente grande,

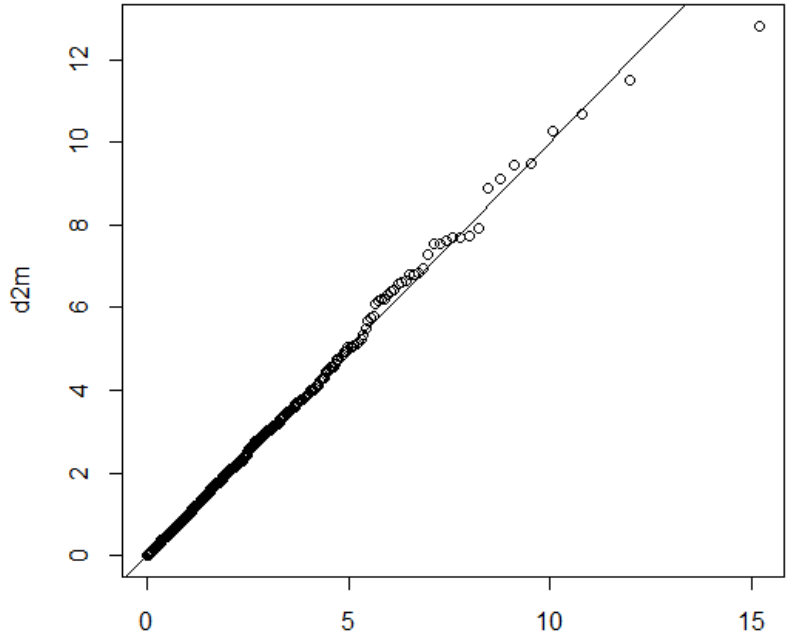
$$(Y_i-\bar{Y})'\,S_u^{-1}(Y_i-\bar{Y})\sim\chi^2_p$$

Diagnóstico de observações atípicas

Averiguação da hipótese de Normalidade dos dados via o Gráfico Chi-Quadrado (Q-Q Plot das Observações).

Distâncias de Mahalanobis






```
library(MASS)
mu<-c(0,0)
sigma<-matrix(c(2,1,1,2),ncol=2)
n<-500
y<-mvrnorm(n,mu,sigma)
mi<-colMeans(y)
s<-cov(y)
par(mfrow=c(1,2))
bivbox(y, method ="O")
# Copy Everitt's bivbox function
d2m<-mahalanobis(y,mi,s)
quantis <- qchisq(ppoints(length(y)),df=2)
qqplot(quantis, d2m)
abline(0,1)
```

Regiões de Confiança - Duas Populações

Amostras Pareadas

Amostra Pareada $\Rightarrow$ respostas multivariadas são avaliadas na mesma unidade amostral em "duas" condições diferentes (Ex.: Antes e Depois de uma intervenção)

Duas Populações

$$\mathrm{Y}_{1\mathrm{n}\times p};\ \mathrm{Y}_{1i\ p\times 1}=\left(Y_{1i1},Y_{1i2},\ldots,Y_{1ip}\right)' \qquad \mathrm{Y}_{2\mathrm{n}\times p};\ \mathrm{Y}_{2i\ p\times 1}=\left(Y_{2i1},Y_{2i2},\ldots,Y_{2ip}\right)' \qquad i=1,2,\ldots,n$$

$$\mathrm{Y}_{1i\ p\times 1}\overset{iid}{\sim}N_p\left(\mu_1;\Sigma_1\right) \qquad \mathrm{Y}_{2i\ p\times 1}\overset{iid}{\sim}N_p\left(\mu_2;\Sigma_2\right)$$

Observações Pareadas $\Rightarrow$ Uma Única População de Diferenças

$$D_{ij}=Y_{1ij}-Y_{2ij} \qquad j=1,2,\ldots,p, \quad i=1,2,\ldots,n$$

$$D_{i\ p\times 1}=\left(D_{i1},D_{i2},\ldots,D_{ip}\right)'\overset{iid}{\sim}N_p\left(\mu_D=\mu_1-\mu_2;\Sigma_D\right) \quad i=1,2,\ldots,n$$

$$T^2=n\left(\bar{D}-\mu_D\right)'S_D^{-1}\left(\bar{D}-\mu_D\right) \sim \frac{(n-1)p}{(n-p)}F_{p,n-p}$$

(S_D: S_u)

Elipse de Confiança:

$$R\left(Y_1,Y_2\right)=\left\{\mu_D\in\Re^p;\ n\left(\bar{D}-\mu_D\right)'S_D^{-1}\left(\bar{D}-\mu_D\right)\le c_\alpha^2\right\}$$

Regiões de Confiança – Duas Populações

Caso Multivariado - Amostras Independentes - Homocedasticidade:

$$Y_{1n_1\times p};Y_{1i} \overset{iid}{\sim} N_p\left(\mu_1;\Sigma_1\right); \qquad Y_{2n_2\times p};Y_{2i} \overset{iid}{\sim} N_p\left(\mu_2;\Sigma_2\right); \quad \Sigma_1=\Sigma_2=\Sigma$$

$$\Rightarrow \bar{D}_{p\times 1}=\bar{Y}_1-\bar{Y}_{12} \sim N_p\left(\mu_D=\mu_1-\mu_2;\Sigma\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)$$

Matriz de covariâncias comum aos grupos (com denominador n-1)

$$S_c=\frac{\sum_{i=1}^{n_1}\left(Y_{1i}-\bar{Y}_1\right)\left(Y_{1i}-\bar{Y}_1\right)'+\sum_{i=1}^{n_2}\left(Y_{2i}-\bar{Y}_2\right)\left(Y_{2i}-\bar{Y}_2\right)'}{n_1+n_2-2}=\frac{(n_1-1)S_{u1}+(n_2-1)S_{u2}}{n_1+n_2-2};$$

$$T^2=\left(\bar{D}-\mu_D\right)'\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_c\right]^{-1}\left(\bar{D}-\mu_D\right) \sim \frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)}F_{(p,(n_1+n_2-p-1))}$$

Elipse de Confiança:

$$R\left(Y_1,Y_2\right)=\left\{\mu_D\in\Re^2;\left(\bar{D}-\mu_D\right)'\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_c\right]^{-1}\left(\bar{D}-\mu_D\right) \leq c_\alpha^2\right\}$$