Técnicas Multivariadas em Saúde

Lupércio França Bessegato
Dep. Estatística/UFJF

Roteiro

1. Introdução
2. Distribuições de Probabilidade Multivariadas
3. Representação de Dados Multivariados
4. Testes de Significância c/ Dados Multivariados
5. Análise de Componentes Principais
6. Análise Fatorial
7. Análise de Correlação Canônica
8. Análise de Conglomerados
9. Análise Discriminante
10. Análise de Correspondência
11. Referências

Vetores Aleatórios

Definições Principais

- Vetores aleatórios:
  - √ Cada componente é uma variável aleatória

$$\mathbf{X}' = [X_1, X_2, \ldots, X_p]$$

- Vetor de médias:

$$\boldsymbol{\mu} = \begin{bmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \\ \vdots \\ \mathrm{E}(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

  - √ A média μ_i é uma das medidas mais utilizadas para sintetizar a informação de tendência central de X_i.

- Variância e desvio padrão:

$$\mathrm{Var}(X_i) = \sigma_i^2 = \sigma_{ii}$$
$$\mathrm{DP}(X_i) = \sqrt{\sigma_{ii}} = \sigma_i$$

√ Informa sobre a disposição dos valores da variável aleatória X_i em relação a μ_i.

- Covariância (σ_{ij}):

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)] \\ &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\,\mathrm{E}(X_j)\end{aligned}$$

√ Mede o grau de relacionamento linear entre duas variáveis aleatórias

- Matriz de covariâncias ($\boldsymbol{\Sigma}$):

√ Variâncias e covariâncias de um vetor aleatório

$$\boldsymbol{\Sigma}_{p\times p} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}.$$

√ É simétrica ($\sigma_{ij} = \sigma_{ji}$)

√ É não negativa definida

$$\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} \geq 0,\ \forall \boldsymbol{a} \in \mathbb{R}^p$$

– Autovalores de $\boldsymbol{\Sigma}$ são não negativos

$$\lambda_i \geq 0,\ \forall i = 1, 2, \ldots, p$$

√ Algumas matrizes são positivas definidas

$$\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} > 0,\ \forall \boldsymbol{a} \in \mathbb{R}^p$$

– Os autovalores são todos positivos

– A matriz $\boldsymbol{\Sigma}$ é não singular e seu determinante é maior que zero

– Assim, a matriz $\boldsymbol{\Sigma}$ terá inversa ($\boldsymbol{\Sigma}^{-1}$)

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} = \boldsymbol{I}_p$$

√ Uma matriz que não tenha a propriedade de simetria e que não seja não negativa definida não poderá ser uma matriz de covariâncias

- Coeficiente de correlação:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

√ É adimensional

√ $-1 \leq \rho_{ij} \leq 1$

√ Medida mais adequada para avaliar o grau de relacionamento linear entre duas variáveis quantitativas

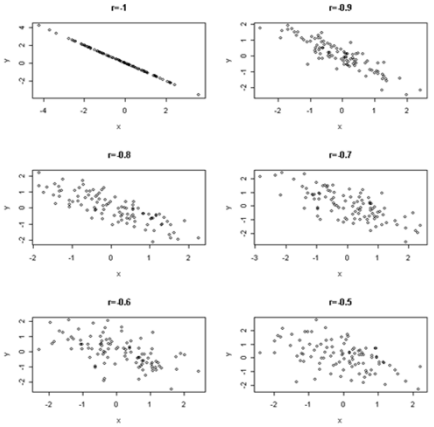
√ Quanto mais próximo $|\rho_{ij}|$ de 1, maior a indicação de que existe um relacionamento linear entre X_i e X_j.

√ Correlação linear próxima de zero é uma indicação de não relacionamento linear entre as variáveis

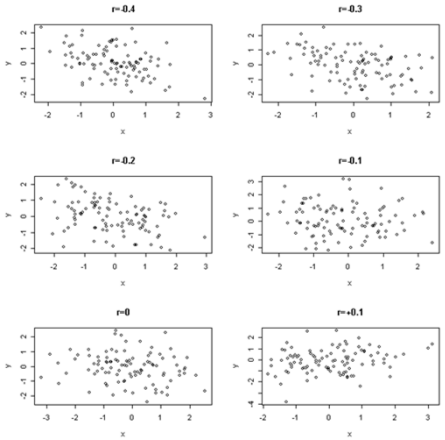
- Matriz de correlação do vetor aleatório **X**:

$$\mathrm{P}_{p\times p} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}.$$

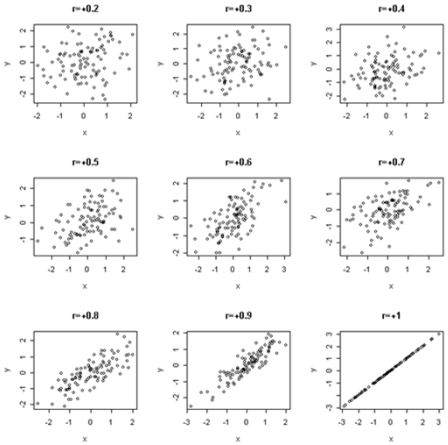
Diagramas de Dispersão (1)

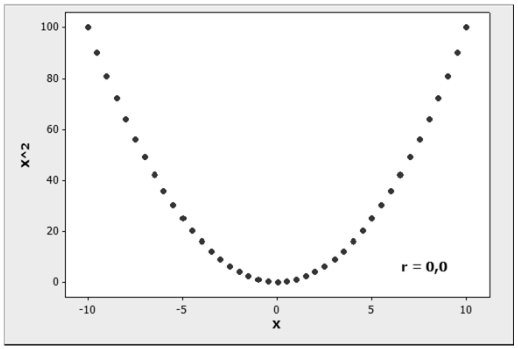


Diagramas de Dispersão (2)



Diagramas de Dispersão (3)





Existe uma relação NÃO –LINEAR entre as variáveis.

- Matrizes de covariâncias de dois vetores aleatórios
  - √ Vetor **X** com matriz de covariâncias $\mathbf{\Sigma_X}$ (pxp)
  - √ Vetor Y com matriz de covariâncias $\mathbf{\Sigma_Y}$ (qxq)

$$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y}) = \begin{matrix} p \\ q \end{matrix}\begin{bmatrix} \overset{p}{\boldsymbol{\Sigma}_X} & \overset{q}{\boldsymbol{\Sigma}_{XY}} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_Y \end{bmatrix}$$

  - √ $\mathbf{\Sigma_{XY}}$: matriz das covariâncias entre as variáveis aleatórias de X e de Y
    - Dimensão pxq
  - √ $\mathbf{\Sigma_{YX}} = \mathbf{\Sigma_{XY}}'$
  - √ Mesmo procedimento é válido para a matriz de correlações

- Variância total:

$$\mathrm{tr}(\boldsymbol{\Sigma}) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp}$$

  - √ Forma de sintetização da variância global da distribuição multivariada
  - √ Altos valores de variâncias totais indicam uma maior dispersão global das variáveis

- Variância generalizada:

$$\det(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|$$

  - √ Também fornece uma noção da dispersão global da distribuição multivariada
  - √ Distribuuições com maiores variabilidades globais apresentam maiores valores de variâncias generalizadas
  - √ A variância generalizada é influenciada pelas covariâncias (ou correlações) entre as variáveis

- Exemplo:

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \text{ e } \boldsymbol{\Sigma}_2 = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$$

√ Variâncias totais e generalizadas
- Tr($\boldsymbol{\Sigma}_1$) = 4 e $|\boldsymbol{\Sigma}_1|$ = 3
- Tr($\boldsymbol{\Sigma}_2$) = 4 e $|\boldsymbol{\Sigma}_2|$ = 2
  - Mesma variância total e variâncias generalizadas diferentes

√ É possível encontrar matrizes de covariâncias diferentes com mesma variância total e mesma variância generalizada

√ Há sempre um risco na comparação usando apenas uma destas duas medidas

- Combinações lineares:

√ Vetor de coeficientes: $\mathbf{c}' = (c_1, c_2, \ldots, c_p)$

√ Vetor aleatório: $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$

√ Combinação linear: $Y = \mathbf{c}'\mathbf{X} = c_1X_1 + c_2X_2 + \cdots + c_pX_p$

√ Média de Y:

$$\begin{aligned} \mu_Y = \mathbf{c}'\boldsymbol{\mu}_\mathbf{X} &= (c_1, c_2, \ldots, c_p) \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \\ &= c_1\mu_1 + c_2\mu_2 + \cdots + c_p\mu_p \end{aligned}$$

√ Variância de Y:

$$\begin{aligned} \sigma_Y^2 &= \mathrm{Var}(Y) = \mathrm{Cov}(\mathbf{c}'\boldsymbol{\Sigma}_\mathbf{X}) \\ &= \mathbf{c}'\boldsymbol{\Sigma}_\mathbf{X}\mathbf{c} \end{aligned}$$

Exemplo

- Combinação linear:

$$Y = c_1 X_1 + c_2 X_2$$

√ Média de Y:

$$\begin{aligned} \mu_Y = \boldsymbol{c}'\boldsymbol{\mu} &= (c_1, c_2) \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ &= c_1\mu_1 + c_2\mu_2 \end{aligned}$$

√ Variância de Y:

$$\begin{aligned} \mathrm{Var}(Y) = \sigma_Y^2 &= \boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c} \\ &= [c_1, c_2] \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \\ &= c_1^2\sigma_{11} + c_2^2\sigma_{22} + 2c_1c_2\sigma_{12} \end{aligned}$$

- Caso geral:

√ q combinações lineares de p variáveis aleatórias

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_q \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1p} \\ c_{21} & c_{22} & \ldots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{q1} & c_{q2} & \ldots & c_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \mathbf{C}\mathbf{X}$$

√ Vetor de médias do vetor de combinações lineares:

$$\boldsymbol{\mu}_\mathbf{Z} = \mathrm{E}(\mathbf{Z}) = \mathrm{E}(\mathbf{C}\mathbf{X}) = \mathbf{C}\boldsymbol{\mu}_\mathbf{X}$$

√ Matriz de covariâncias de Z

$$\boldsymbol{\Sigma}_\mathbf{Z} = \mathrm{Cov}(\mathbf{Z}) = \mathrm{Cov}(\mathbf{C}\mathbf{X}) = \mathbf{C}\boldsymbol{\Sigma}_\boldsymbol{X}\mathbf{C}'$$

Exemplo

- Combinações lineares:

$$Z_1 = X_1 - X_2$$
$$Z_2 = X_1 + X_2$$

√ Forma matricial:

$$\boldsymbol{Z} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \boldsymbol{CX}$$

√ Vetor de médias de **Z** :

$$\boldsymbol{\mu_Z} = \mathrm{E}(\mathbf{Z}) = \mathbf{C}\boldsymbol{\mu_X} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_2 + \mu_2 \end{bmatrix}$$

√ Matriz de covariâncias de **Z**:

$$\boldsymbol{\Sigma_Z} = \mathrm{Cov}(\mathbf{Z}) = \mathbf{C}\boldsymbol{\Sigma_X}\mathbf{C}' = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} \sigma_{11} - 2\sigma_{12} + \sigma_{22} & \sigma_{11} - \sigma_{22} \\ \sigma_{11} - \sigma_{22} & \sigma_{11} + 2\sigma_{12} + \sigma_{22} \end{bmatrix}$$

- $Cov(Z_1, Z_2)$ não depende de $Cov(X_1, X_2)$
- Se $\sigma_{11} = \sigma_{22}$, as variáveis Z_1 e Z_2 são não correlacionadas [$Cov(Z_1, Z_2) = 0$]

Teorema da Decomposição Espectral

- $\boldsymbol{\Sigma}_{pxp}$ é uma matriz de covariâncias

√ Existe uma matriz ortogonal $\mathbf{O}_{pxp}$ tal que

$$\mathbf{O}'\boldsymbol{\Sigma}\mathbf{O} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} = \boldsymbol{\Lambda}$$

√ Autovalores ordenados de **Σ**: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$

√ Matriz ortogonal:

- $\mathbf{O}' = \mathbf{O}^{-1}$
- $\mathbf{O}'\mathbf{O} = \mathbf{O}\mathbf{O}' = \mathbf{I}_{pxp}$

- A matriz **Σ** é similar à matriz **Λ**

√ Traço de **Σ**: $\mathrm{tr}(\boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} \lambda_i$

√ Determinante de **Σ**: $\det(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}| = |\boldsymbol{\Lambda}| = \prod_{i=1}^{p} \lambda_i$

- Autovetores de **Σ**:

√ $\mathbf{e}_i$: autovetor normalizado correspondente a λ_i

√ i-ésima coluna de O $\quad \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{bmatrix}$

- Matriz ortogonal **O**: $\mathbf{O} = [\mathbf{e}_1 \quad \mathbf{e}_2 \ldots \mathbf{e}_p]$
- Teorema da decomposição espectral:

  como $\mathbf{O}'\boldsymbol{\Sigma}_{p\times p}\mathbf{O} = \boldsymbol{\Lambda}$ e $\mathbf{O}^{-1} = \mathbf{O}'$

  então $\boldsymbol{\Sigma}_{p\times p} = \mathbf{O}\boldsymbol{\Lambda}\mathbf{O}' = \sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i'$

  √ $\mathbf{e}_i$ tem comprimento unitário

  $$\|\mathbf{e}_i\| = \sqrt{\mathbf{e}_{i1}^2 + \mathbf{e}_{i2}^2 + \cdots + \mathbf{e}_{ip}^2} = 1$$

  √ Produto escalar dos autovetores:

  – Projeção do vetor $\mathbf{e}_i$ em $\mathbf{e}_j$

  $$\mathbf{e}_i'\mathbf{e}_j = \begin{cases} 1 & , \text{se } i = j \\ 0 & , \text{se } i \neq j \end{cases}$$

- Equação característica:

  √ Os autovalores são solução da equação característica:

  $$|\boldsymbol{\Sigma}_{p\times p} - \lambda \mathbf{I}_{p\times p}| = 0$$

Estimação de Parâmetros

- Parâmetros populacionais precisam ser estimados
- Amostra aleatória de observações multivariadas

  √ n vetores aleatórios independentes e identicamente distribuídos

  $$\mathbf{X}_1 = \begin{bmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1p} \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2p} \end{bmatrix}, \ldots, \quad \mathbf{X}_n = \begin{bmatrix} X_{n1} \\ X_{n2} \\ \vdots \\ X_{np} \end{bmatrix},$$

  √ Matriz de dados:

  $$\mathbf{X}_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix}$$

- Vetor de médias amostrais ($\bar{\mathbf{X}}$)

  √ É estimador não viciado de $\boldsymbol{\mu}_X$.

  $$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix}.$$

  √ $\bar{X}_i$: média amostral da i-ésima variável

- Matriz de covariâncias amostrais (**S**)
  - √ É estimador não viciado de $\mathbf{\Sigma_X}$.

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{12} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1p} & S_{2p} & \dots & S_{pp} \end{bmatrix}.$$

  - √ Elementos da matriz de covariâncias amostral:

$$S_{ij} = \begin{cases} \dfrac{\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)^2}{n-1} & , \text{para } i = j \\ \dfrac{\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{n-1} & , \text{para } i \neq j \end{cases}.$$

- Matriz de correlações amostrais (**R**)

$$\mathbf{R}_{p\times p} = \begin{bmatrix} 1 & R_{12} & \dots & R_{1p} \\ R_{12} & 1 & \dots & R_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ R_{1p} & R_{2p} & \dots & 1 \end{bmatrix}.$$

  - √ Elementos da matriz de covariâncias amostral:

$$R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}, \text{ para } i \neq j.$$

    - R_{ij} é a covariância amostral das variáveis padronizadas

- Relação entre **S** e **R**:
  - √ $D^{½}$: matriz dos desvios padrão amostrais

$$\mathbf{D}^{1/2}_{p\times p} = \begin{bmatrix} \sqrt{S_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{S_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{S_{pp}} \end{bmatrix}.$$

  - √ Matriz de correlações amostrais: $\mathbf{R}_{p\times p} = \mathbf{D}^{-1/2}\mathbf{S}\,\mathbf{D}^{-1/2}$.
  - √ Matriz de covariâncias: $\mathbf{S}_{p\times p} = \mathbf{D}^{1/2}\mathbf{R}\,\mathbf{D}^{1/2}$.
  - √ Observação padronizada: $Z_{ik} = \dfrac{X_{ik} - \bar{X}_i}{\sqrt{S_{kk}}}$.

Teste de Hipóteses para ρ

- Suposição:
  - √ X_i e X_j têm distribuição normal univariada.
- Hipóteses:
  - √ H_0: $\rho_{ij} = 0$ vs. H_1: $\rho_{ij} \neq 0$
- Estatística de teste: $T = R_{ij}\sqrt{\dfrac{n-2}{1-R_{ij}^2}}$
  - √ R_{ij}: correlação amostral observada entre X_i e X_j.
- Distribuição amostral: $T = R_{ij} \overset{H_0}{\sim} t_{n-2}$
  - √ Rejeita-se H_0 se $|T| > t_{(n-2),\alpha/2}$

- Estimação das matrizes de covariâncias e de correlação
  - √ Tamanho da amostra (n) deve ser maior que p
  - √ É adequado trabalhar com um número maior que p+1

Distribuição Normal Multivariada

Normal Multivariada

- Suponha que tenhamos *p* variáveis X_1, X_2, ..., X_p
  - √ Vetor de componentes $\mathbf{X}' = [X_1, X_2, \ldots, X_p].$
  - √ Vetor de médias: $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_p].$
  - √ Matriz de variâncias e covariâncias

$$\Sigma_X = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}.$$

  - √ Variância da variável aleatória X_i: $\mathrm{Var}(X_i) = \sigma_{ii} = \sigma_i^2$
  - √ Covariância entre Variáveis X_i e X_j: $\mathrm{Covar}(X_i, X_j) = \sigma_{ij}$

Função de Densidade de Probabilidade

- Distribuição Normal Univariada:

distância quadrática padronizada

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

- Distribuição Normal Multivariada:

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

Padronização volume sob superfície

distância generalizada quadrática padronizada

√ Distância de Mahalanobis do vetor **x** ao vetor de média **μ**.

$$(\mathbf{x}-\boldsymbol{\mu})'\,\Sigma^{-1}\,(\mathbf{x}-\boldsymbol{\mu})$$

– Distância padronizada ou distância estatística

√ Função de densidade da normal p-variada pode ser expressa como:

$$f_{\mathbf{X}}(\mathbf{x}) = k\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\left(\sum_{i=1}^{p}\frac{1}{\lambda_i}\mathbf{e}_i\mathbf{e}_i'\right)(\mathbf{x}-\boldsymbol{\mu})\right\}$$
$$= k\exp\left\{-\frac{1}{2}\sum_{i=1}^{p}\frac{1}{\lambda_i}\left[\mathbf{e}_i'(\mathbf{x}-\boldsymbol{\mu})\right]^2\right\}$$

√ Função de densidade da normal p-variada pode ser expressa como:

$$f_{\mathbf{X}}(\mathbf{x}) = k\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\left(\sum_{i=1}^{p}\frac{1}{\lambda_i}\mathbf{e}_i\mathbf{e}_i'\right)(\mathbf{x}-\boldsymbol{\mu})\right\}$$
$$= k\exp\left\{-\frac{1}{2}\sum_{i=1}^{p}\frac{1}{\lambda_i}\left[\mathbf{e}_i'(\mathbf{x}-\boldsymbol{\mu})\right]^2\right\}$$

– onde $k = \dfrac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}$

√ Função de densidade da normal p-variada pode ser expressa como:

$$f_{\mathbf{X}}(\mathbf{x}) = k\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\left(\sum_{i=1}^{p}\frac{1}{\lambda_i}\mathbf{e}_i\mathbf{e}_i'\right)(\mathbf{x}-\boldsymbol{\mu})\right\}$$
$$= k\exp\left\{-\frac{1}{2}\sum_{i=1}^{p}\frac{1}{\lambda_i}\mathbf{e}_i'(\mathbf{x}-\boldsymbol{\mu})^2\right\}$$

– onde $k = \dfrac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}$

√ Para todos os vetores x e para C constante, tais que

$$C^2 = (\mathbf{x}-\boldsymbol{\mu})'\,\Sigma^{-1}\,(\mathbf{x}-\boldsymbol{\mu}) = \sum_{i=1}^{p}\frac{1}{\lambda_i}\left[\mathbf{e}_i'(\mathbf{x}-\boldsymbol{\mu})\right]^2$$

√ A função de densidade assume o mesmo valor numérico

√ Curva de mesma densidade tem formato de elipsóide

– Eixo principal: direção correspondente à variável de maior variabilidade

– (maior autovalor)

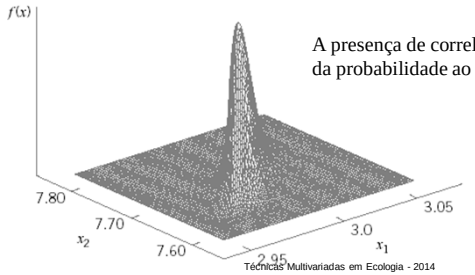
– Segundo eixo: relacionado com a variável de segunda maior variância

– (segundo auto valor e assim por diante)

Normal Bivariada

- Função de densidade de probabilidade (p=2)

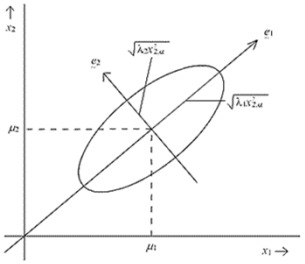
$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}$$



A presença de correlação causa concentração da probabilidade ao longo de uma linha

- Curvas de nível de uma normal bivariada

$$C^2 = (\mathbf{x}-\boldsymbol{\mu})'\,\Sigma^{-1}\,(\mathbf{x}-\boldsymbol{\mu}) = \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]$$



√ Maior semi-eixo: $\frac{C}{\sqrt{\lambda_1}}$

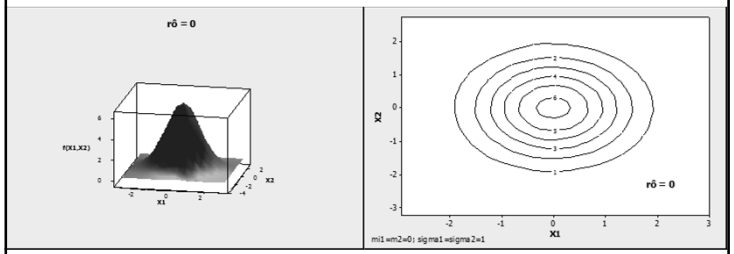
√ Menor semi-eixo: $\frac{C}{\sqrt{\lambda_2}}$

- Variáveis não correlacionadas ($\rho_{12} = 0$)

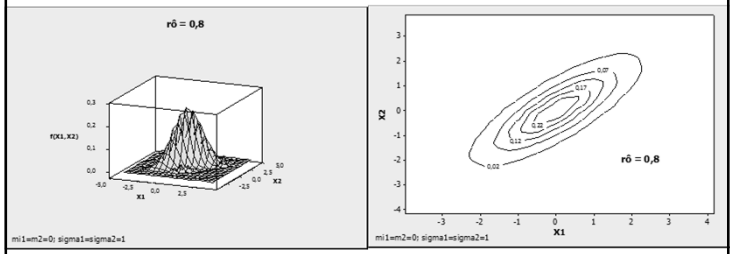
$$\begin{aligned} f_{X_1,X_2}(x_1,x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right\} \\ &\times \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\} \\ &= f_{X_1}(x_1) f_{X_2}(x_2) \end{aligned}$$

√ X_1 e X_2 serão independentes se elas forem não correlacionadas

- X1 e X2 independentes

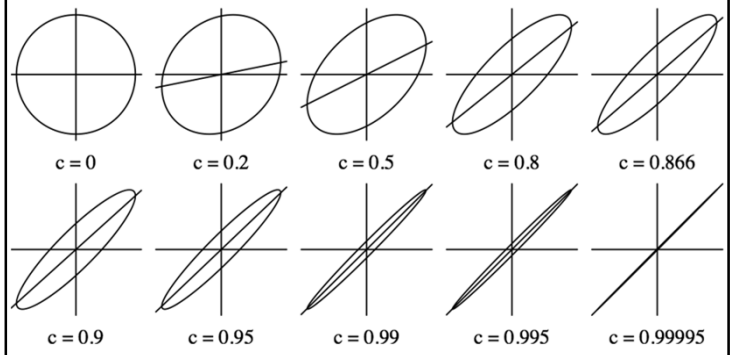


- Corr(X_1, X_2) = 0,8



√ A presença de correlação causa concentração da probabilidade ao longo de uma linha

Efeito Correlação



- Variância generalizada – Normal bivariada

$$|\Sigma| = \sigma_{11}\sigma_{22}(1-\rho^2)$$

√ À medida que ρ tende a zero a superfície fica mais dispersa em torno da média
(variância generalizada maior)

√ Quanto maior o valor de |ρ| menor será a variância generalizada

Propriedades da Normal Multivariada

Seja o vetor aleatório **X** ~ Normal p-variada

√ Se cov(X_1, X_2) = 0, então X_1 e X_2 são independentes

√ As densidades marginais são normais
$X_i \sim N(\mu_i, \sigma_{ii})$

√ As combinações lineares construídas com componentes de **X** são normais

√ Qualquer conjunto de k variáveis de **X**, k < p, tem distribuição normal k-variada

√ As distribuições condicionais envolvendo subconjuntos de variáveis aleatórias de **X** são normais

√ Combinações lineares de vetores aleatórios que tenham distribuição normal multivariada também são normalmente distribuídas

Verificação da Hipótese de Normalidade

Métodos de Verificação – Normal Multivariada

- Análise das distribuições univariadas e bivariadas auxiliam na verificação da suposição de normalidade p-variada
  - √ Demonstrar que distribuições univariadas e bivariadas são normais não implica que o vetor aleatório seja normal multivariado
  - √ Na prática, é muito grande a chance de o vetor ser normal, quando aas distribuições normais e bivariadas são normais

Distribuições Univariadas – Verificação da Normalidade

- Avaliação gráfica
  - √ Verificação de simetria
    - Histograma (n>25)
    - Gráfico de pontos (n pequenos)
  - √ Gráficos de probabilidade normal
- Testes de hipóteses
  - √ Ryan-Jones
  - √ Shapiro-Wilk
  - √ Anderson-Darling

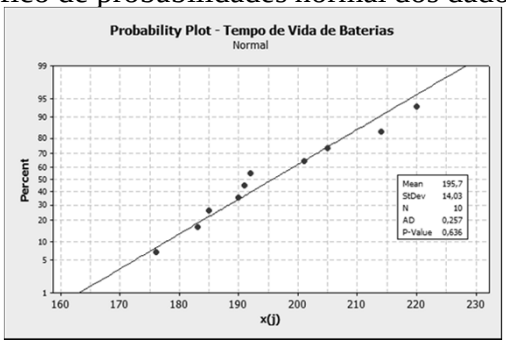
Gráficos de Probabilidade Normal

√ Pontos próximos da reta indicam que a hipótese de normalidade permanece defensável
- Há muita variabilidade na linearidade para amostras pequenas.
- Em geral, não são informativos a menos que o tamanho amostral seja moderadamente grande ($n \geq 20$)
- Linearidade do gráfico de probabilidade pode ser medida pelo coeficiente de correlação entre os pontos

√ Padrões de desvio podem fornecer pistas sobre a natureza da não normalidade

Técnicas Multivariadas em Ecologia - 2014

- Gráfico de probabilidades normal dos dados:



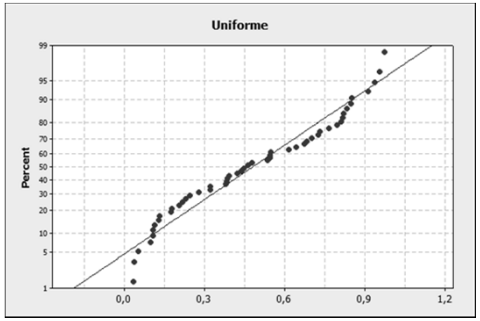
√ Ser mais influenciado pelos pontos do meio que pelos dos extremos

√ Eixo y com escala de probabilidades (escala z)

Gráfico de Probabilidades Normal

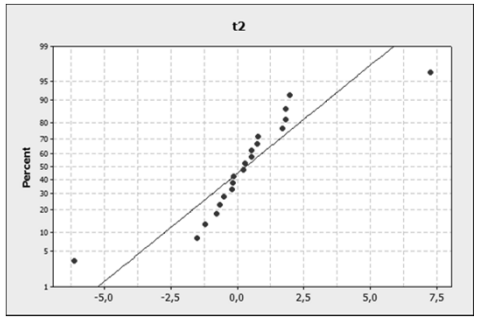
- Pode ser útil na identificação de distribuições que sejam simétricas mas que tenham caudas mais pesadas (ou mais leves) que a normal

- Distribuição de cauda leve



√ Pontos à esquerda tendem a ficar abaixo da linha e à direita tendem a ficar acima
- As menores e maiores observações não serão tão extremas como se esperaria de uma normal

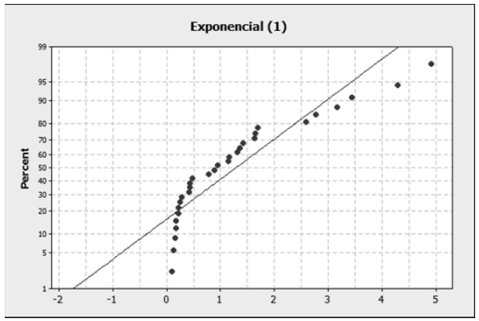
- Distribuição de cauda pesada



√ Pontos à esquerda tendem a ficar acima da linha e à direita tendem a ficar abaixo

√ Gráfico em forma de S

- Distribuição assimétrica



√ Pontos de ambas as extremidades tendem a estar abaixo da linha

√ Gráfico tem forma curvada

Distribuições Bivariadas – Verificação da Normalidade

- Diagrama de dispersão de pares de variáveis:
  - √ Observações provenientes de normal mulvariada:
    - cada distribuição bivariada será normal
    - plot dos pontos bivariados observados devem exibir padrão global aproximadamente elíptico

Distâncias Quadráticas Generalizadas

- Método mais formal para julgar a normalidade
  - √ Distância estatística de cada ponto amostral ao centróide de todas as observações

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}), \quad j = 1, 2 \ldots n.$$

  - √ Pode ser usada para $p \geq 2$
  - √ $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$: observações amostrais

- Se população for normal multivariada e n e $(n - p)$ forem suficientemente grandes
  - √ Cada uma das distâncias quadráticas deveria se comportar como uma variável aleatória χ^2
  - √ Embora essas distâncias não sejam independentes ou exatamente distribuídas como uma χ^2 é útil plotá-las como se fossem

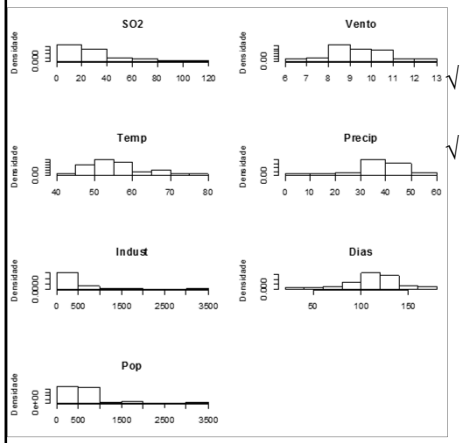
Q-Q Plot

- Procedimento:
  - √ Ordenar as distâncias quadráticas:
    - $d_{(1)}^2 \leq d_{(2)}^2 \leq ... \leq d_{(n)}^2 \leq$
  - √ Plotar os pares $(q_{c,p}((j - ½)/n), d(j)^2)$
  - √ $q_{c,p}((j - ½)/n)$ é o $100(j - ½)/n$ percentil superior de uma χ^2_p
- Em um gráfico de χ^2 os pontos deveriam estar próximos da linha reta

Exemplo

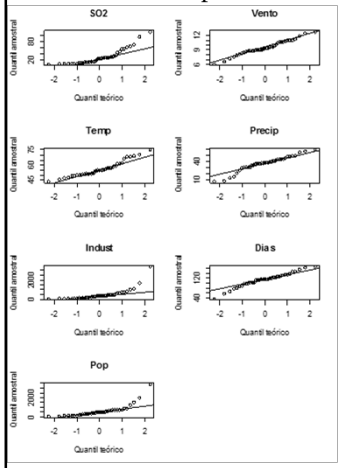
- Estudo poluição do ar
  - √ Amostra: 41 cidades americanas
  - √ Variáveis:
    - SO2: concentração no ar (mg/m3)
    - Temp: temperatura
    - Popul: população, em milhares (censo 1970)
    - Vento: velocidade média anual (milhas/hora)
    - Precip: precipitação média anual (pol)
    - Dias: número médio anual de dias de chuva
  - √ Dados: USairpollution (pacote: HSAUR2)

- Histogramas das variáveis (univariada)



- √ SO2 e Precipitação:
  - Forte assimetria
- √ Indústrias e População:
  - Indícios de outliers

- Gráficos de probabilidade univariados



√ Gráficos para SO2 e Precipitação desviam-se consideravelmente da linearidade

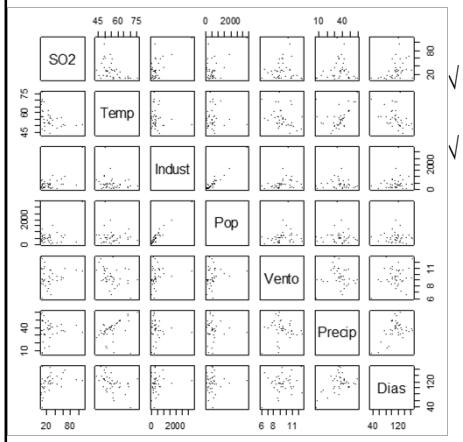
√ Gráficos para Indústrias e População evidenciam outliers

- Teste de normalidade
  - √ H_0: dados se ajustam à distribuição normal
  - √ H_1: dados não se ajustam à distribuição normal

```
> teste<- sapply(colnames(dados), function(x)shapiro.test(dados[[x]])$p.value)
> as.matrix(teste)
                [,1]
SO2     9.723376e-06
Temp    2.214972e-02
Indust 2.781101e-09
Pop     3.622798e-08
Vento   6.972580e-01
Precip 3.725311e-02
Dias    2.419457e-01
```

√ Não se rejeita a hipótese de normalidade para as variáveis Vento e Dias

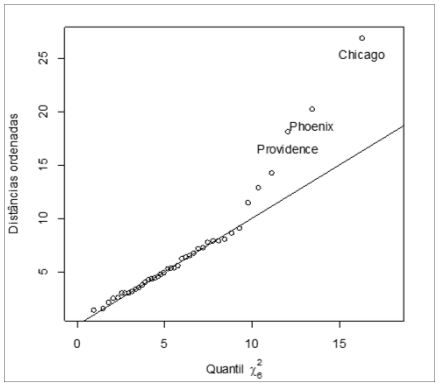
- Diagramas de dispersão (bivariada)



√ SO2 e Precipitação:
  - Formato não elíptico

√ Indústrias e População:
  - Indício de formação elíptica

- Gráfico das distâncias generalizadas (χ^2):



√ Gráfico detectou possíveis outliers nos dados multivariados
  - Desvio da variabilidade natural dos dados

Comentários

- É difícil construir um bom teste global de normalidade conjunta em mais de duas dimensões
- Hipótese de normalidade aparenta estar violada
  - √ As marginais aparentam ser normais? E aalgumas combinações lineares de componentes X_i?
  - √ Diagramas de dispersão de diferentes características têm aparência elíptica?
  - √ Há outliers que devessem ser verificados?

- Hipótese de normalidade individual é menos crucial em situações em que o tamanho amostral é grande e as técnicas dependem da média amostral (ou de distâncias envolvendo essa média)

Teste de Hipóteses para a Matriz de Correlação

- Suposição:
  - √ Vetor aleatório é normal p-variado
- Hipóteses:
  - √ H_0: $\mathbf{P}_{pxp} = \mathbf{I}_p$ vs. H_1: $\mathbf{P}_{pxp} \neq \mathbf{I}_p$
  - √ Equivale a testar se as variáveis são independentes
    - (matriz de covariâncias diagonal)

- Estatística de teste: $T = -\left[n - \frac{1}{6}(2p+11)\right] \sum_{j=1}^{p} \ln(\hat{\lambda}_i)$
  - √ λ_i: autovalores da matriz de correlação
- Distribuição amostral: $T \overset{H_0}{\sim} \chi^2_{gl}$, com $gl = \frac{1}{2}p(p-1)$

Exemplo

- Matriz de correlação amostral (n=40):

$$\mathbf{R}_{4\times4} = \begin{bmatrix} 1 & 0,8 & 0,6 & 0,7 \\ 0,8 & 1 & 0,4 & 0,5 \\ 0,6 & 0,4 & 1 & 0,3 \\ 0,7 & 0,5 & 0,3 & 1 \end{bmatrix}.$$

  - √ Autovalores de R:
    - $\lambda_1 = 2,687$; $\lambda_2 = 0,714$; $\lambda_3 = 0,486$ e $\lambda_4 = 0,113$
  - √ Estatística de teste:

$$T = -\left[40 - \frac{1}{6}(2(4)+11)\right][\ln(2,687) + \ln(0,714) + \ln(0,486) + \ln(0,113)] = 82,81$$

    - Graus de liberdade: p(p – 1)/2=6
  - √ Valor crítico $\chi^2_{6;\,0,05} = 12,59$
  - √ Conclusão:
    - as 4 variáveis não são mutuamente independentes

- Comentários
  - √ Existem outros coeficientes que medem a associação entre variáveis que independem da suposição de normalidade:
    - Teste não paramétrico de Kendall
    - Teste não paramétrico de Spearman
  - √ Transformar os dados originais é uma alternativa quando os dados não provém de normal multivariada
    - Classe de transformação muito utilizada: Box-Cox
    - Nem sempre é possível obter uma transformação adequada

Referências

Bibliografia Recomendada

- MANLY, B. J. F. *Métodos Estatísticos Multivariados: uma Introdução.* Bookman, 2008.
- JOHNSON, R. A.; WINCHERN, D. W. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2007
- MINGOTI, D.C. *Análise de Dados através de Métodos de Estatística Multivariada*. Ed. UFMG, 2005.
- EVERITT, B.; HOTHORN, T. *An Introduction to Applied Multivariate Analysis with R*. Springer, 2011.