Análise Multivariada

Lupércio França Bessegato
Dep. Estatística/UFJF

Roteiro

Introdução

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

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

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

6

Análise Multivariada - 2015

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

7

Análise Multivariada - 2015

√ Distância de Mahalanobis do vetor **x** ao vetor de média $\boldsymbol{\mu}$.

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

- Distância padronizada ou distância estatística

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

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

8

Análise Multivariada - 2015

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

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

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

9

Análise Multivariada - 2015

√ 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)

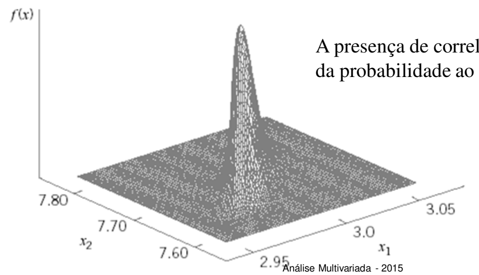
– Segundo eixo: relacionado com a variável de segunda maior variância

– (segundo auto valor e assim por diante)

Normal Bivariada

- Função de densidade de probabilidade (*p*=2)

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

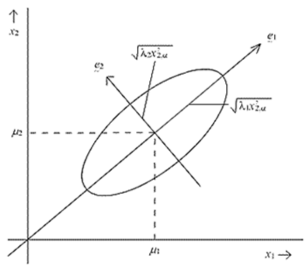


A presença de correlação causa concentração da probabilidade ao longo de uma linha

- 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: $\dfrac{C}{\sqrt{\lambda_1}}$

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

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

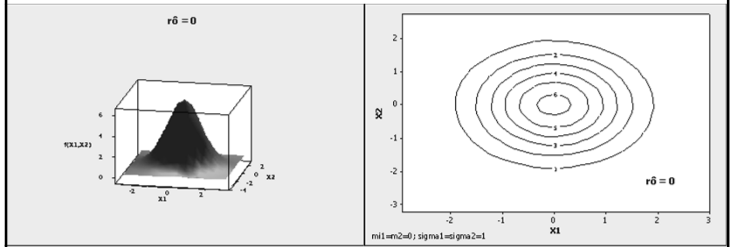
$$
\begin{aligned}
f_{X_1,X_2}(x_1,x_2) = {} & \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \right.\right. \\
& \left.\left. + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\rig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

√ X_1 e X_2 serão independentes se elas forem não correlacionadas

15

Análise Multivariada - 2015

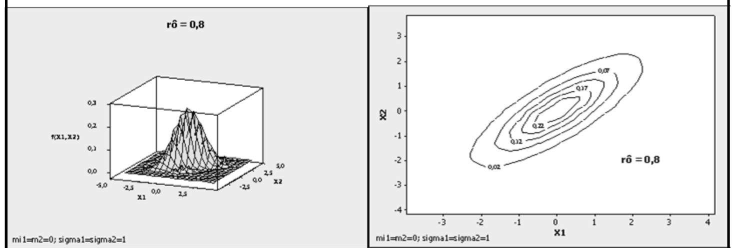
- X1 e X2 independentes



16

Análise Multivariada - 2015

- Corr(X_1, X_2) = 0,8

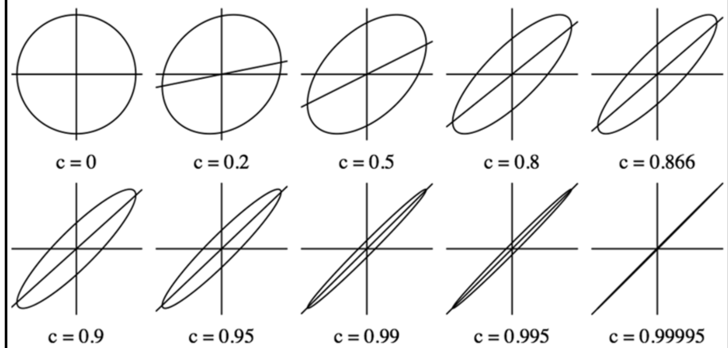


√ A presença de correlação causa concentração da probabilidade ao longo de uma linha

17

Análise Multivariada - 2015

Efeito Correlação



19

Análise Multivariada - 2015

- 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

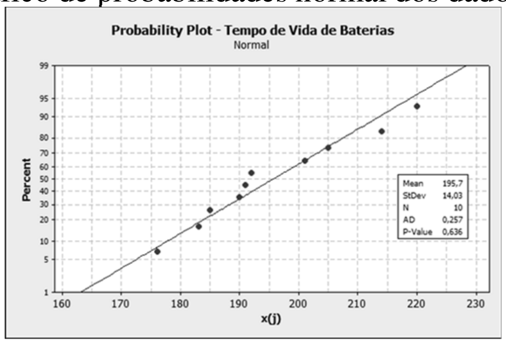
Distribuições Univariadas – Verificação da Normalidade

- 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

- Gráfico de probabilidades normal dos dados:

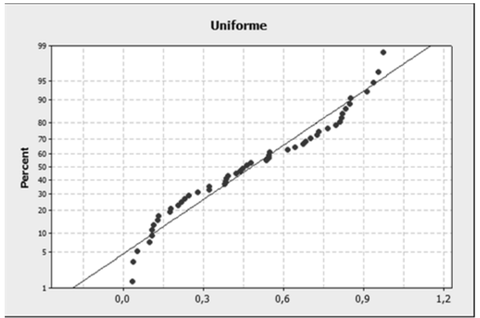


  - √ Ser mais influenciado pelos pontos do meio que pelos dos extremos
  - √ Eixo y com escala de probabilidades (escala z)

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

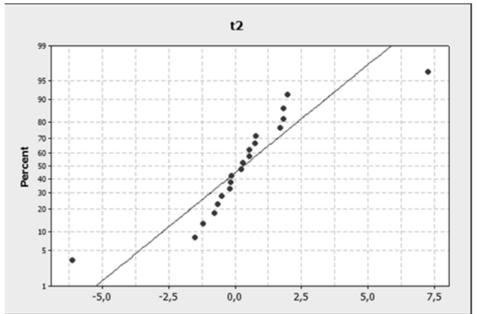
- Distribuição de cauda leve



√ 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

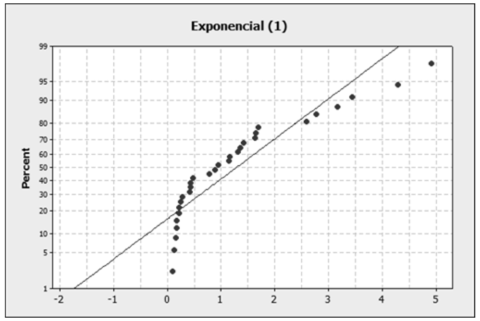
- Distribuição de cauda pesada



√ 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

37

Análise Multivariada - 2015

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

38

Análise Multivariada - 2015

- 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

39

Análise Multivariada - 2015

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

40

Análise Multivariada - 2015

- Gráfico de χ^2
  - √ Distâncias quadráticas generalizadas

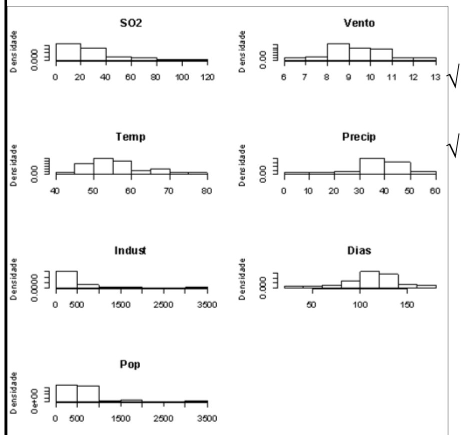
$$d_j^2 = (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}})$$

  - √ Os pontos d_j^2 de uma normal p-variada tem distribuição χ^2_p
    - Para amostras grandes
  - √ Em um gráfico de χ^2 os pontos deveriam estar próximos da linha reta

Exemplo

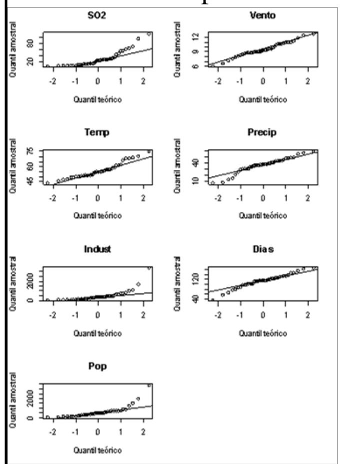
- Estudo poluição do ar
  - √ Amostra: 41 cidades americanas
  - √ Variáveis:
    - SO2: concentração no ar (mg/m3)
    - Temp: temperatura
    - Popul: população, em milhares (censo 1970)
    - Vento: velocidade média anual (milhas/hora)
    - Precip: precipitação média anual (pol)
    - Dias: número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

- Histogramas das variáveis (univariada)



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

- Gráficos de probabilidade univariados



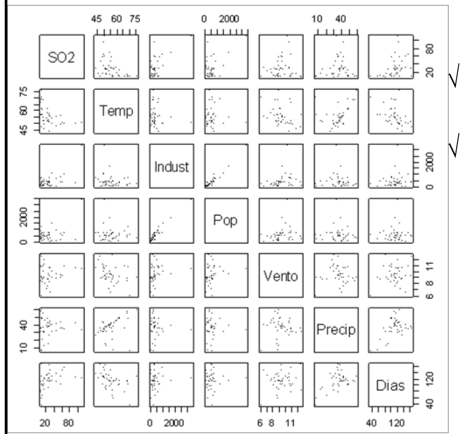
  - √ Gráficos para SO2 e Precipitação desviam-se consideravelmente da linearidade
  - √ Gráficos para Indústrias e População evidenciam outliers

- 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

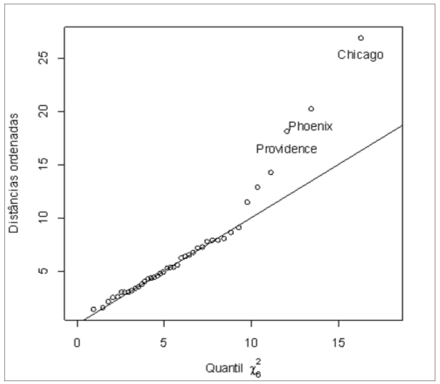
- Diagramas de dispersão (bivariada)



√ 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)

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.