Análise Multivariada

Lupércio França Bessegato
Dep. Estatística/UFJF

Roteiro

Distribuição Normal Multivariada

Normal Multivariada

- Suponha que tenhamos *p* variáveis X_1, X_2, ..., X_p
  - √ Vetor de componentes $\mathbf{X}' = [X_1, X_2, \ldots, x_p](\underset{\sim}{X}')$
  - √ Vetor de médias: $\boldsymbol{\pi}' = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_p](\underset{\sim}{\pi}')$
  - √ Matriz de variâncias e covariâncias

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

  - √ 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:

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

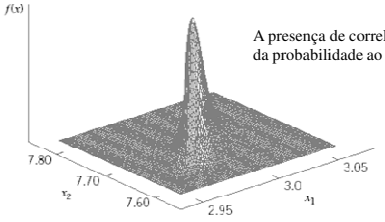
Padronização volume sob superfície

distância generalizada quadrática padronizada

Normal Bivariada

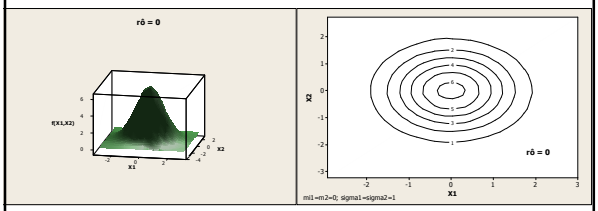
- Função de densidade de probabilidade (*p*=2)

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

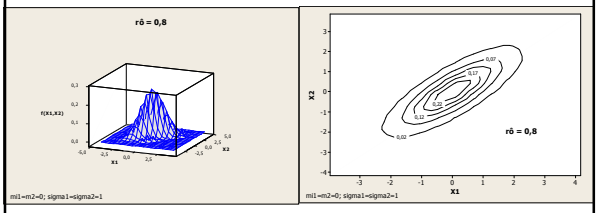


A presença de correlação causa concentração da probabilidade ao longo de uma linha

- X1 e X2 independentes

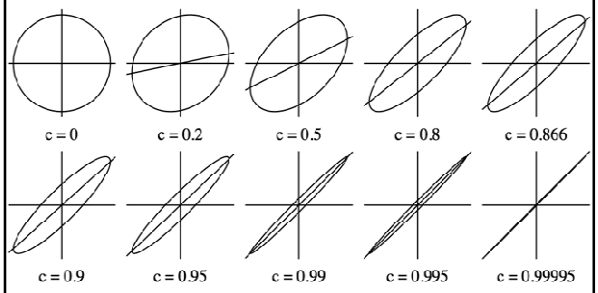


- Corr(X_1, X_2) = 0,8



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

Efeito Correlação



Vetor de Média Amostral

- Suponha uma amostra aleatória de uma distribuição normal multivariada $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$

  $\boldsymbol{x}_i$: i-ésimo vetor amostral

  $\boldsymbol{x}_i' = [\, x_{i1}, x_{i2}, ..., x_{ip}]$
- Vetor de médias amostrais

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

Matriz de Variâncias e Covariâncias Amostral

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

- Variâncias Amostrais (diagonal de S)

$$S_{jj} = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \quad j = 1, 2, \ldots, p$$

- Covariâncias amostrais

$$S_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_k)$$

- $\bar{\mathbf{X}}$ e $\mathbf{S}$ são estimadores não-viciados de m e S, respectivamente

$$E[\bar{\mathbf{X}}] = \boldsymbol{\mu}$$

$$E[\mathbf{S}] = \Sigma$$

Verificação da Hipótese de Normalidade

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
- O conjunto de observações bivariadas tais que $(\boldsymbol{x}-\boldsymbol{\mu})\mathbf{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \leq \chi_2^2(0,5)$ tem probabilidade 0,5
  - √ Devemos esperar a mesma percentagem (50%) de valores amostrais estejam internos (ou sobre) a elipse

Exemplo 4.12 – Dados Empresas

- Dados sobre 10 maiores indústrias americanas (na época)
  - √ Variáveis:
    - X1: vendas
    - X2: lucros
  - √ Banco de dados: *BD_multivariada.xls/industrias_usa*

- Vetor de médias amostral

Descriptive Statistics: x1=sales; x2=profits; x3=assets

Variable	Mean
x1=sales	62309
x2=profits	2927

- Matriz de covariâncias amostral

Covariances: x1=sales; x2=profits

	x1=sales	x2=profits
x1=sales	1000509114	
x2=profits	25575600	1430020

- Está dentro (ou sobre) a curva estimada de 50% qualquer ponto que satisfaça a relação:

$$\begin{bmatrix} x_1 & - & 62.309 \\ x_2 & - & 2.927 \end{bmatrix}' \begin{bmatrix} 0,000184 & -0,003293 \\ -0,003293 & 0,128831 \end{bmatrix} \times 10^{-5} \begin{bmatrix} x_1 & - & 62.309 \\ x_2 & - & 2.927 \end{bmatrix} \leq 1,39$$

- Comandos Minitab para cálculo das distâncias:

```
# Dados nas colunas C1 C2

Let K1 = 2     # Quantidade variáveis (p)
Let K2 = 10    # Tamanho amostra (n)

        # Calculo medias

Mean C2 K3   # Media X1
Mean C3 K4   # Media X2

        # Cálculo Matriz S

Covariance C2-C3 M1
Name M1 "S"
Print M1

        # Calculo Inversa de S

Invert M1 M2
Name M2 "Sinv"
```

```
        # Calculo Distancia generalizada

Let C5 = C2 - K3
Let C6 = C3 - K4
Name C5 "X1 - med1"
Name C6 "X2 - med2"
Copy C5 - C6 M3
Multiply M3 M2 M4
Copy M4 C7 - C8

Name C9 "dj^2"
Let C9 = C7*c5 + C8*c6
```

- Cálculo da distância quadrática generalizada

Company	x1–sales	x2–profits	dj^2
General Motors	126974	4224	4,34
Ford	96933	3835	1,20
Exxon	86656	3510	0,59
IBM	69438	3758	0,83
General Eletric	55264	3939	1,88
Mobil	50976	1809	1,01
Philip Morris	39069	2946	1,02
Chrysler	36156	359	5,33
Du Pont	35209	2480	0,81
Texaco	32416	2413	0,97

√ 70% dos dados estão dentro da curva de 50%

√ Importante: a amostra é muito pequena para tirar conclusões

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:
    - $d_{(1)}^2 \leq d_{(2)}^2 \leq \ldots \leq d_{(n)}^2 \leq$
  - √ 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

Exemplo 4.13 – Construindo QQ-Plot

- Gráfico qui-quadrado das distâncias generalizadas calculadas no Exemplo 4.12
  √ Uso de Macro Global do Minitab

Macro Global – Características

- Age diretamente na planilha em uso (planilha global);
- Ao escrever a macro, deve-se saber quais colunas, constantes e matrizes serão usadas ao executar a macro;
- Não executam macros locais
- Alguns comandos dos menus são macros locais

Macro Global – Uso

- Usar quando:
- A tarefa é simples;
- É possível saber os estado da planilha;
- Não necessitar de comandos que são macros locais;

Criando uma Macro

- Use um processador de texto para escrevê-la;
- Salve-a no sub-diretório MACROS, com extensão .MAC
- Para executar a macro entre com: *% nome_da_macro*
- Se julgar mais conveniente, execute interativamente os comandos e copie-os da janela *Session* ou *History*

Criando uma Macro (2)

- Para mostrar uma linha em branco use o comando *NOTE*
- Se a extensão do arquivo não for .MAC, digite o nome e a extensão do arquivo
- Caso o arquivo tenha sido salvo em outro diretório, especifique seu caminho
- (*% caminho_do_arquivo\nome_da_macro*)

Estrutura Macro Global

- Uma macro global segue a seguinte estrutura:

GMACRO	Início da macro global
Template	Nome da macro
Body of the macro	Comandos da macro
ENDMACRO	Final da macro global

- Macro Global – Minitab

```
GMACRO
DISTANCIA
# Macro Global
# Cálculo de Distância Generalizada
# Dados nas colunas C1 C2

Let K1 = 2      # Quantidade variáveis
(p)
Let K2 = 10   # Tamanho amostra (n)

        # Calculo medias

Mean C2 K3   # Media X1
Mean C3 K4   # Media X2

        # Cálculo Matriz S

Covariance C2-C3 M1
Name M1 "S"
Print M1

        # Calculo Inversa de S

Invert M1 M2
Name M2 "Sinv"
```

```
        # Calculo Distancia generalizada

Let C5 = C2 - K3
Let C6 = C3 - K4
Name C5 "X1 - med1"
Name C6 "X2 - med2"
Copy C5 - C6 M3
Multiply M3 M2 M4
Copy M4 C7 - C8

Name C9 "dj^2"
Let C9 = C7*c5 + C8*c6

        # Ordenacao distancias

Name C10 "d(j)^2"
Sort C9 C10
```

```
        # Cálculo percentis qui-quadrado

Set C11
1(1:10/1)1
End
Let C11 = (C11 - 1/2)/K2

Name C12 "qc.p((j-1/2)/10)"
InvCDF c11 c12;
ChiSquare K1.

        # Plot Qui-Quadrado

Plot c12*c10;
Title "Q-Q Plot das Distâncias
Ordenadas";
Symbol.
ENDMACRO
```

Execução da Macro

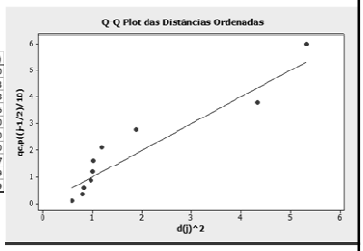
- Para executar, digite *%nome_da_macro;*
  - √ Ex. Entre *%analise*
- Se a extensão do arquivo não for MAC deve-se digitá-la também;
- Ao ser executada, o Minitab procura a macro primeiro no diretório atual e depois no subdiretório MACROS
  - √ Caso a macro não esteja salva em um desses diretório, deve-se especificar o caminho até ela.

Exemplos

Template	Nome arquivo	Execução
MyMacro	MyMacro.MAC	%MYMACRO
Analise	TEST.MAC	%TEST
Analise2	TESTE2.TXT	%TESTE2.TXT
Analise	Analise.MAC	%C:\pasta1\ ... \analise

- QQ-Plot das distâncias Generalizadas:

j	d(j)^2	(j-1/2)/n	$q_{c,p}((j-1/2)/10)$
1	0,59	0,05	0,10
2	0,81	0,15	0,33
3	0,83	0,25	0,58
4	0,97	0,35	0,86
5	1,01	0,45	1,20
6	1,02	0,55	1,60
7	1,20	0,65	2,10
8	1,88	0,75	2,77
9	4,34	0,85	3,79
10	5,33	0,95	5,39



√ Pontos não estão dispostos em torno de linha reta
- Menores distâncias aparentam ser grandes, médias distâncias aparentam ser pequenas

√ Dados não parecem ser normais (embora *n* seja pequeno)

Referências

Bibliografia Recomendada

- JOHNSON, R. A.; WINCHERN, D. W. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1998
- MINGOTI, D.C. *Análise de Dados através de Métodos de Estatística Multivariada*. Ed. UFMG, 2005.
- LATTIN, J.; CARROLL, J. D.; GREEN, P. E. *Análise de Dados Multivariados*. Cengage, 2011.