NOTAS DE AULA

Análise exploratória de dados: outra estratégia

Prof.: Idemauro Antonio Rodrigues de Lara

Medidas de posição: separatrizes

Definição. Separatrizes são índices ou medidas de posição que dividem um conjunto ordenado de valores em k partes iguais quanto ao número de elementos. Se $k = 2$, tem-se a mediana do conjunto. Além da mediana, outras separatrizes usuais são os quartis, decis e percentis.

1. **Quartis** são medidas que dividem um conjunto ordenado de dados em 4 partes iguais quanto ao número de elementos.

   Notação: $Q_i \quad i = 1, 2, 3.$

2. **Decis** são medidas que dividem um conjunto ordenado de dados em 10 partes iguais quanto ao número de elementos.

   Notação: $D_i \quad i = 1, 2, 3, \ldots, 9.$

3. **Percentis** são medidas que dividem um conjunto ordenado de dados em 100 partes iguais quanto ao número de elementos.

   Notação: $P_i \quad i = 1, 2, 3, \ldots, 99.$

Medidas de posição: separatrizes

Determinação das sepatrizes

Considere $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$ as estatísticas de ordem da amostra.

1. **Método geral**

$$Q_i = x_{[\frac{i}{4}n]}, \qquad i = 1, 2, 3.$$

$$D_i = x_{[\frac{i}{10}n]}, \qquad i = 1, 2, 3, \ldots, 9.$$

$$P_i = x_{[\frac{i}{100}n]}, \qquad i = 1, 2, 3, \ldots, 99.$$

2. Método da interpolação linear simples

Este método é demonstrado geometricamente e, parte da hipótese que existe uma relação linear entre a ordem dos valores no rol e o percentil correspondente ao respectivo valor.

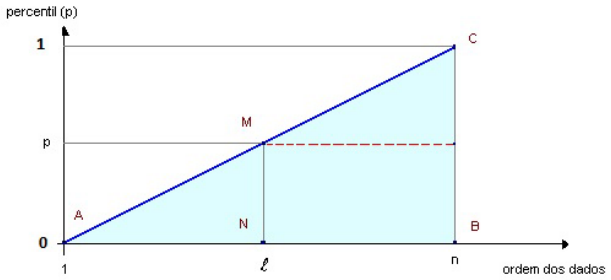

Figura 1: Geometria e estatísticas de ordem.

Pela figura 1 temos que o $\Delta ABC \sim \Delta AMN$, então:

$$\frac{\overline{AB}}{\overline{BC}} = \frac{\overline{AN}}{\overline{MN}} \Rightarrow \frac{n-1}{1} = \frac{\ell - 1}{p}$$

o que nos leva a:

$$\ell = p(n-1) + 1 \qquad (0 < p < 1)$$

Deve-se observar que, por hipótese, a relação é linear, assim, a separatriz desejada sai por interpolação linear simples:

$$Sp = x_{[\mathrm{int}(\ell)]} + \{x_{[\mathrm{int}(\ell+1)]} - x_{[\mathrm{int}(\ell)]}\}\mathrm{dec}(\ell)$$

em que:

$\mathrm{int}(\ell)$ denota a parte inteira do número ℓ;
$\mathrm{dec}(\ell)$ denota a parte decimal do número ℓ.

Exemplo: Considere os dados sobre a produção diária de leite, em kg, de 10 produtores rurais.

9,80	9,90	9,95	10,00	8,78
9,90	9,34	10,34	11,75	15,00

Calcular e interpretar os quartis da distribuição.

$$Q_1 =$$

$$Q_2 =$$

$$Q_3 =$$

Técnicas descritivas

Em uma análise descritiva visamos resumir a apresentação dos dados. São elementos importantes na descrição:

- **Tabela de frequências**
- **Gráficos**, em particular, o **histograma** para dados contínuos;
- **Medidas descritivas: média, variância, assimetria.**

Outras técnicas descritivas

Tukey (1977), em sua obra "Exploratory data analysis", apresenta outros elementos muito úteis para a descrição dos dados.

- **Diagrama de ramo e folhas**
- **Resumo dos cinco números**
- **Gráfico de Caixa** (*boxplot*)

Diagrama de ramo e folhas

O diagrama de ramo e folhas é um dispositivo para apresentação dos dados, que pode, em alguns casos, substituir a tabela de frequências e o histograma simultaneamente. Não existem regras rígidas para sua construção, a ideia básica consiste na segmentação dos valores dos dados em duas partes: os "ramos" e as "folhas".

Exemplo. Para ilustrar vamos considerar os dados referemtes ao comprimento das asas, em mm, de uma amostra de um tipo de inseto.

4,2	4,3	4,3	4,4	4,4	4,1	4,1	4,1	4,1	4,1
7,7	8,3	8,5	11,3	13,8	6,0	6,5	6,9	7,1	7,2
5,8	6,0	6,0	6,0	6,0	4,6	4,9	5,0	5,0	5,3
3,5	3,5	3,8	3,8	3,9	2,2	2,3	2,5	2,6	3,0

Dispositivo de ramo e folhas referente aos comprimentos das asas, em mm.

```
 2 | 2356
 3 | 0055889
 4 | 011111234469
 5 | 0038
 6 | 0000059
 7 | 127
 8 | 35
 9 |
10 |
11 | 3
12 |
13 | 8
```

Resumo dos cinco números

De acordo com Tukey (1977) a caracterização da distribuição dos dados fica bem descrita pelo "resumo dos cinco números".

Mediana	
1^0 Quartil	3^0 Quartil
Mínimo	Máximo

Este resumo associa o limite inferior e superior do rol aos quartis, dando uma ideia razoável da:

- **Tendência central** $\Rightarrow$ mediana;
- **Dispersão** $\Rightarrow$ amplitude total, dispersão inferior (md $-$ min), dispersão superior (max $-$ md), o desvio quartil $(Q_3 - Q_1)$ e o desvio semi quartil $dQ = \frac{1}{2}(Q_3 - Q_1)$;
- **Forma da distribuição** $\Rightarrow$ para uma distribuição simétrica, espera-se que md $-$ min $\simeq$ max $-$ md; md $- Q_1 \simeq Q_3 -$ md e $Q_1 -$ min $\simeq$ max $- Q_3$.

Observação: Coeficiente de Curtose de Keley

$$C(k) = \frac{Q_3 - Q_1}{2(D_9 - D_1)} = \frac{dQ}{D_9 - D_1}$$

- $C(k) = 0,263 \Rightarrow$ distribuição mesocúrtica.
- $C(k) < 0,263 \Rightarrow$ distribuição leptocúrtica.
- $C(k) > 0,263 \Rightarrow$ distribuição platicúrtica.

Gráfico de caixa

Para completar a "descrição" existem os limites discrepantes, c_1 e c_2, que delimitam os valores "fora do padrão", a saber:

$$c_1 = Q_1 - 1,5(Q_3 - Q_1) = Q_1 - 3dQ \quad \text{e} \quad c_1 = Q_3 + 1,5(Q_3 - Q_1) = Q_3 + 3dQ$$

Assim, os **valores discrepantes** em uma distribuição, caso existam, estarão localizados **abaixo** de c_1 ou **acima** de c_2.

Com estes sete pontos pode-se construir o gráfico de "caixa".

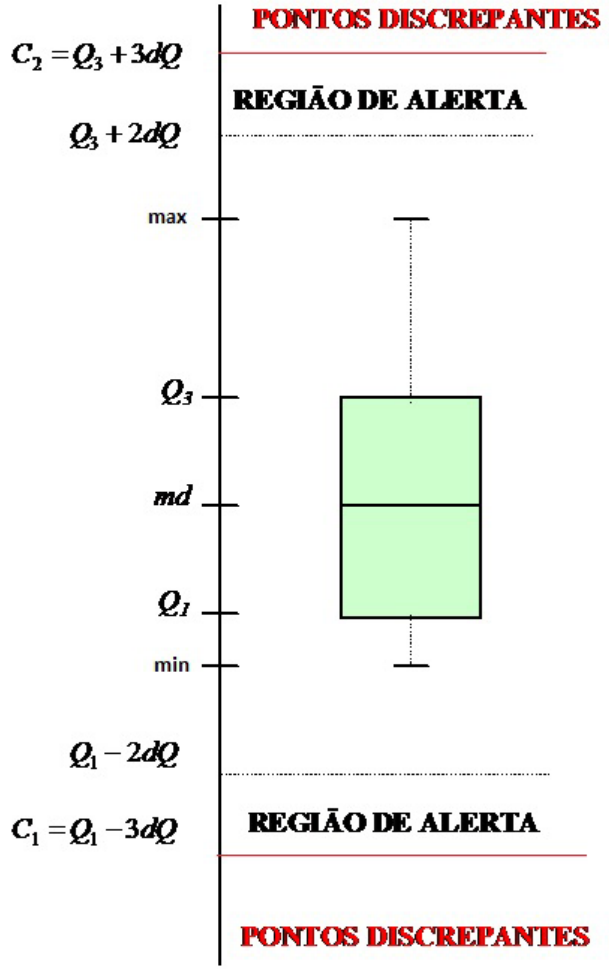

Figura 2: Esquema de um gráfico de caixa (*boxplot*).

Exemplo. Considere os dados referemtes ao comprimento das asas, em mm, de uma amostra de um tipo de inseto. Pede-se: resumo dos cinco números, construção do gráfico de caixa, estudar a assimetria e a existência ou não de valores discrepantes.

2,2	2,3	2,5	2,6	3,0	3,5	3,5	3,8	3,8	3,9
4,1	4,1	4,1	4,1	4,1	4,2	4,3	4,3	4,4	4,4
4,6	4,9	5,0	5,0	5,3	5,8	6,0	6,0	6,0	6,0
6,0	6,5	6,9	7,1	7,2	7,7	8,3	8,5	11,3	13,8

Mediana =

1^0 Quartil = 3^0 Quartil =

Mínimo = Máximo =

$c_1 =$

$c_2 =$

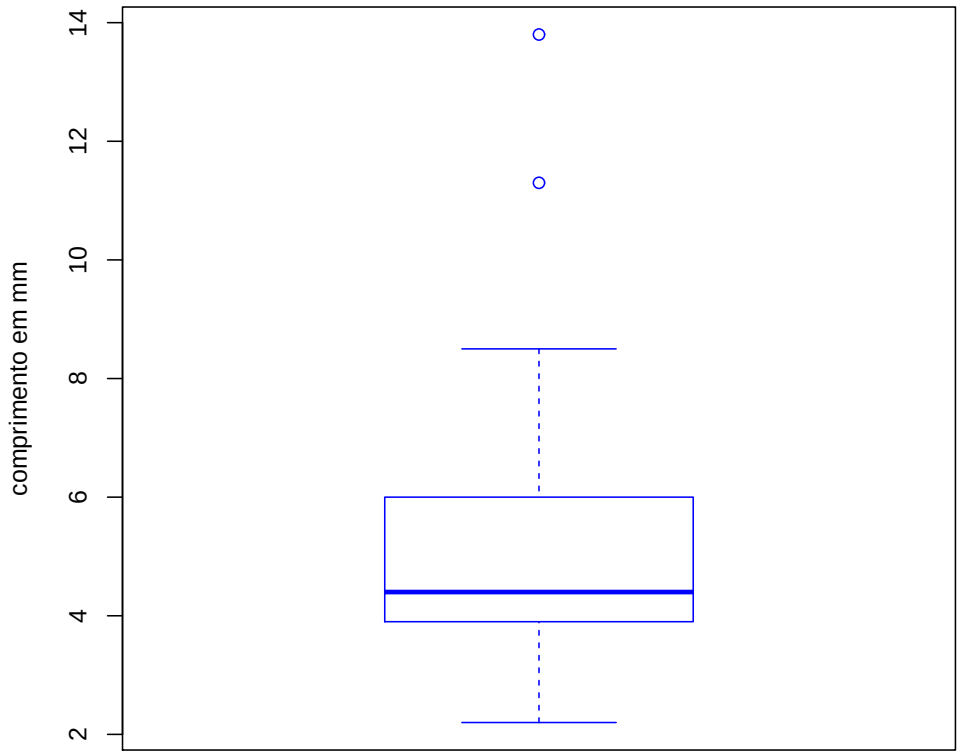

Figura 3: Gráfico de caixa referente ao comprimento das asas de uma espécie de inseto.

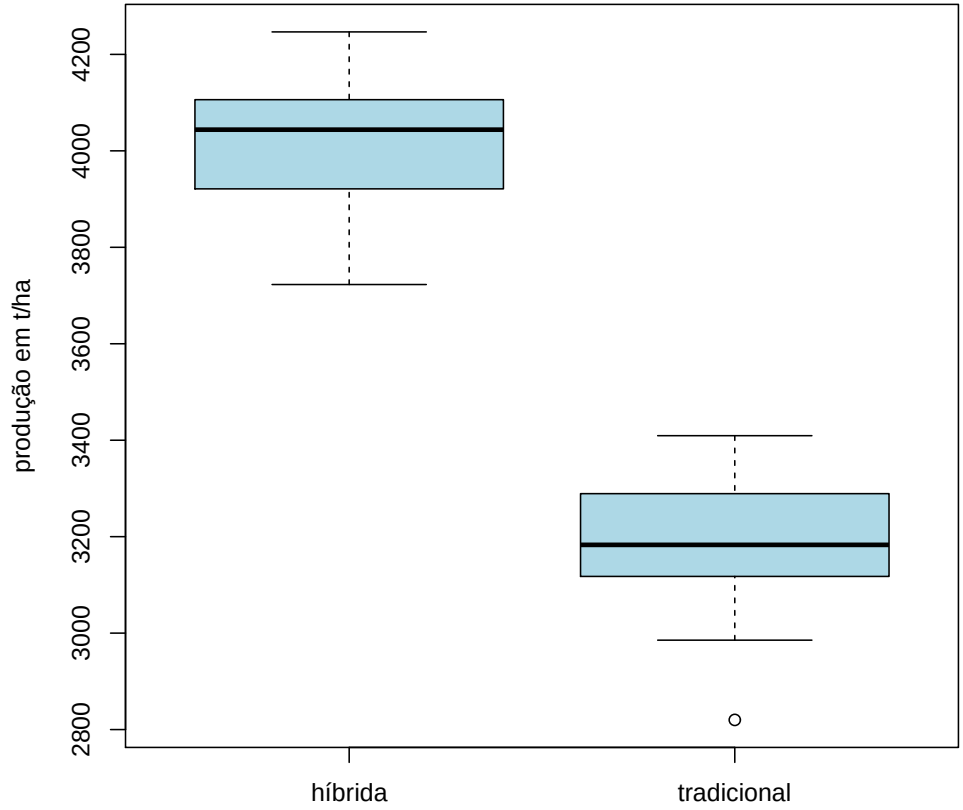

Figura 4: Produção de duas variedades de milho.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ANDRADE, D.F.; OGLIARI, P.J. **Estatística para as Ciências Agrárias e Biológicas com noções de experimentação**. Editora da UFSC, 2007.
2. BUSSAB, W.O. ; P.A. MORETIN, **Estatística Básica**, 5^a edição. Editora Saraiva, 2002.
3. IEMMA, A. F. **Estatística Descritiva.** Piracicaba. $\varphi\sigma\rho$ Publicações, 1992.
4. TUKEY,P. **Exploratory Data Analysis**. Addinson Wesley, 1977.