Elementos de Estatística

Lupércio F. Bessegato & Marcel T. Vieira

UFJF – Departamento de Estatística
2013

Medidas Resumo

Medidas Resumo

- Medidas que sintetizam informações contidas nas variáveis em um único número
- Tipos:
  - √ Medidas de tendência central
  - √ Medidas de dispersão
  - √ Quartis, Decis e Percentis
  - √ Medidas de assimetria
  - √ Medidas de curtose

Medidas de Tendência Central

Medidas de Tendência Central

- Em geral, podem ser interpretadas como o ponto ao redor do qual os dados são distribuídos
- Algumas medidas de posição (tendência central):
  - √ Média
  - √ Mediana
  - √ Moda

Média

- Tendência central dos dados caracterizada pela média aritmética simples;
  - √ Média amostral
  - √ Média populacional

Média Amostral

- Os dados em geral são provenientes de uma amostra de observações selecionada de uma população
- Definição:

  Se n observações em uma amostra forem denotadas por x_1, x_2, ..., x_n, a média amostral será:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Exemplo – Peso

- Peso (kg)
- $n = 50$ indivíduos
- Média amostral

$$\bar{x} = \frac{3.046,4}{50} = 60,93\,\text{kg}$$

Média Populacional

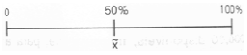
- Valor médio de todas as observações em uma população:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- A média amostral é um '***bom***' estimador da média populacional

Mediana

- Valor que divide a distribuição dos dados em duas partes de igual tamanho



- 50% das observações ficam acima da mediana e 50%, abaixo

- Determinação da mediana:
  - √ Quantidade ímpar de observações:

1	4	7	9	10	12	14

Mediana

  - √ Quantidade par de observações

11	13	15	16	19	21	22	25

Mediana

Procedimento

- Ordenar os dados
- Se n for ímpar:
  - √ A mediana é o valor do elemento central
  - √ Elemento de ordem $\frac{n+1}{2}$
- Se n for par:
  - √ A mediana é o valor médio entre os dois elementos centrais
  - √ Elementos de ordem $\frac{n}{2}$ e $\frac{n}{2}+1$

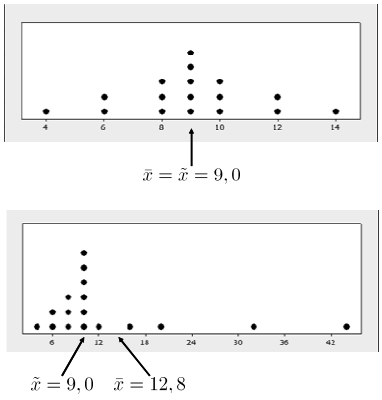
Exemplo – Peso (kg)

- Peso (kg)
- $n = 50$ indivíduos
- Valor médio entre o 25º e o 26º valores ordenados

  $x_{(25)} = 58$; $x_{(26)} = 58$
- Mediana

$$\tilde{x} = \frac{58 + 58}{2} = 58\,\text{kg}$$

Média & Mediana



Média e Mediana

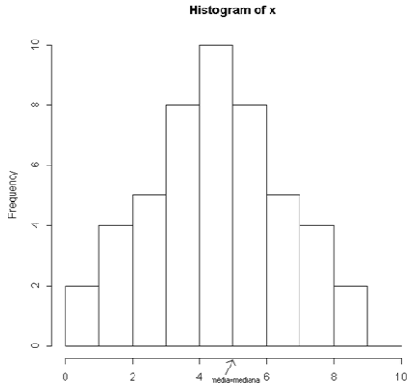
- Valores atípicos (muito grandes ou muito pequenos) causam grandes variações na média
- Em geral, a mediana não é afetada da mesma forma que a média
- A mediana é uma medida mais robusta (menos afetada pro valores atípicos)

Média *vs.* Mediana

Média	Mediana
• fácil de ser manipulada algebricamente;	• difícil de ser manipulada algebricamente;
• representa o "centro de massa" dos dados (ponto de equilíbrio no histograma).	• valor da posição central dos dados ordenados;
• afetada grandemente por valores extremos .	• não é afetada por valores extremos.

Média *vs.* Mediana (2)

- Para distribuições muito assimétricas, a mediana é uma medida mais apropriada para caracterizar um conjunto de dados.
- Se a distribuição é aproximadamente simétrica, então média e mediana são aproximadamente iguais.
  - √ Em distribuições perfeitamente simétricas média = mediana.



Média – Dados em Tabelas de Frequência

- Para dados disponíveis apenas em tabela de frequências
- Para calcular a média em tabela com k classes:

Ponto Médio	Freqüência
x_1	f_1
x_2	f_2
⋮	⋮
x_k	f_k

$$n = \sum_{i=1}^{k} f_i$$

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{n} = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i f_i$$

- Exemplo - Tabela de Frequências – Peso

Peso (kg)	Ponto Médio (x_i)	Freq. Absoluta (f_i)	$x_i \cdot f_i$
40 ⊢ 50	45	8	360
50 ⊢ 60	55	22	1210
60 ⊢ 70	65	8	520
70 ⊢ 80	75	6	450
80 ⊢ 90	85	5	425
90 ⊢ 100	95	1	95
Total			3060

$$\bar{x}_{tab.} = \frac{3060}{50} = 61,20\,\text{kg}$$

$$\bar{x}_{exata} = 60,93\,\text{kg}$$

Moda

- É o valor mais frequente da distribuição.
- No histograma, ou na tabela de frequências, a classe modal é a classe de maior frequência e a moda são aproximadas pelo ponto médio da classe.

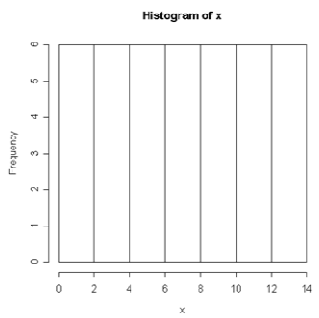
Exemplo: Peso

- Classe Modal: [50; 60)
  √ Maior frequência = 22 observações
- Moda: 55 kg

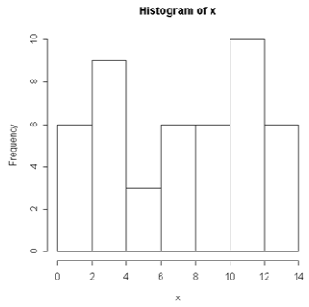
Moda (2)

- Uma distribuição pode não possuir moda (amodal – distribuição "achatada").
- Uma distribuição pode possuir mais de uma moda (multimodal).
- Uma distribuição pode possuir apenas uma moda (unimodal).

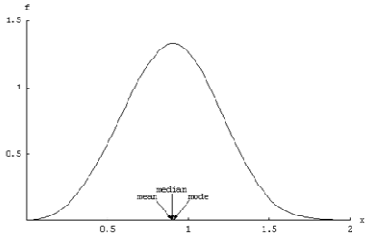
Distribuição "Achatada"



Distribuição Multimodal

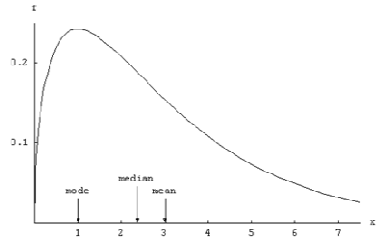


Medidas de Posição – Distribuições Simétricas



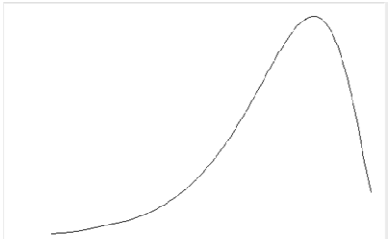
média = mediana = moda

Medidas de Posição – Distribuições Assimétricas à Direita



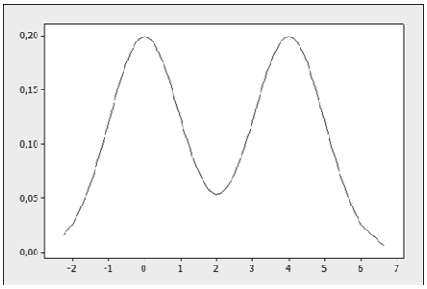
média > mediana > moda

Medidas de Posição – Distribuições Assimétricas à Esquerda



média < mediana < moda

Distribuições Bimodais



média = mediana ≠ moda

Medidas de Dispersão

Comparação entre Grupos de Dados

Stem-and-Leaf Display: grupo_1

```
Stem-and-leaf of grupo_1  N  = 10
Leaf Unit = 0,10

(10)  5  0000000000
```

Stem-and-Leaf Display: grupo_2

```
Stem-and-leaf of grupo_2  N  = 10
Leaf Unit = 0,10

 4  2  0000
 5  3  0
 5  4
 5  5
 5  6
 5  7  0
 4  8  0000
```

Stem-and-Leaf Display: grupo_3

```
Stem-and-leaf of grupo_3  N  = 10
Leaf Unit = 0,10

 3  4  000
(4) 5  0000
 3  6  000
```

Stem-and-Leaf Display: grupo_4

```
Stem-and-leaf of grupo_4  N  = 10
Leaf Unit = 0,10

 1  1  0
 2  2  0
 3  3  0
 4  4  0
(2) 5  00
 4  6  0
 3  7  0
 2  8  0
 1  9  0
```

Stem-and-Leaf Display: grupo_5

```
Stem-and-leaf of grupo_5  N  = 10
Leaf Unit = 0,10

 1  3  0
 3  4  00
(4) 5  0000
 3  6  00
 1  7  0
```

Média e Mediana

- Todos os conjuntos têm média e mediana iguais a 5
- Podemos afirmar que a distribuição dos dados é a mesma?

Comentários

- Há grandes diferenças entre os grupos;
  - √ Grupo 1: Todos os valores são iguais a 5.
  - √ Grupo 2: Nenhum valor igual a 5;
  - √ Grupo 3: Valores concentrados entre 4 e 6.
  - √ Grupo 4: Valores espalhados entre 1 e 9
  - √ Grupo 5: Valores dispersos entre 3 e 7
- Além da média e da mediana, é necessário outro tipo de medida para caracterizar os grupos!

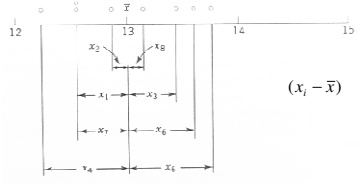
Medidas de Dispersão

- Informações importantes sobre os dados:
  - √ Valor em torno do qual os dados se **concentram**
  - √ Valor do grau de dispersão dos dados
- Medidas de dispersão mais comuns:
  - √ Amplitude amostral
  - √ Variância amostral (Desvio-padrão amostral)
  - √ Distância interquartílica (ou desvio interquartílico)

Amplitude Amostral - *r*

- É a mais simples das medidas de dispersão.
- É definida como: $r = \max(x_i) - \min(x_i)$
- Desvantagem:
  - ✓Omite toda a informação entre o mínimo e o máximo
  - ✓Em geral, quando $n < 10$, esta perda de informações não será muito séria

Construção de uma Medida de Dispersão



- Quanto maior a variabilidade dos dados, maior o valor absoluto de alguns desvios
- Valor absoluto complica o tratamento matemático
- A soma dos desvios é zero
- Uma solução: considerar o quadrado dos desvios

Variância Amostral

- É a média dos desvios quadráticos em relação à média.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Tem unidade diferente dos dados.
- Por questões técnicas (Inferência), adota-se *n-1* no denominador da média.
  - ✓Torna-se o 'melhor' estimador

Desvio-padrão Amostral (*s*)

- É a raiz quadrada da variância amostral
  - √ A unidade de medida é a mesma dos dados!

- Conjunto de dados:

5 2 3 4 8

$$\begin{aligned} s^2 &= \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_5 - \bar{x})^2}{5-1} \\ &= \frac{(5-4,4)^2 + (2-4,4)^2 + (3-4,4)^2 + (4-4,4)^2 + (8-4,4)^2}{4} \\ &= \frac{21,2}{4} = 5,3 \end{aligned}$$

$$s = \sqrt{s^2} = \sqrt{5,3} = 2,30$$

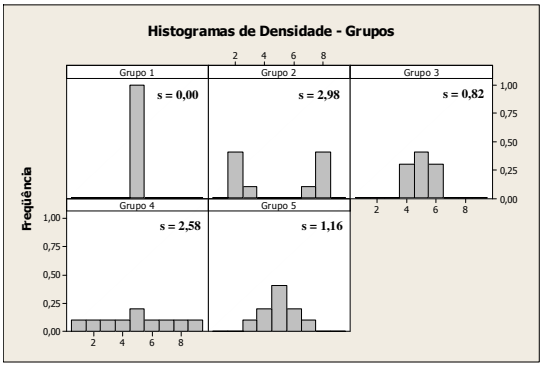
Cálculo Alternativo

- Variância: $s^2 = \frac{1}{n-1}\left[\sum x_i^2 - n(\bar{x})^2\right]$

x_i	x_i^2
5	25
2	4
3	9
4	16
8	54
22	**118**

$$s^2 = \frac{1}{5-1}\left[118 - 5(4,4)^2\right] = \frac{21,2}{4} = 5,3$$

$$s = \sqrt{5,3} = 2,30$$



Coeficiente de Variação

- Medida relativa de dispersão: $cv = \frac{s}{\bar{x}} \cdot 100$
- Medida adimensional
- Fornece medida de homogeneidade dos dados
  - √ Quanto menor o *cv*, maior a homogeneidade
- Utilidades:
  - √ Comparação grau de concentração (dispersão) em torno da média
  - √ Comparação entre variáveis (ou grupos)

Exemplo – Peso

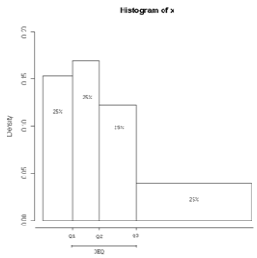
- Peso (kg)
- *n = 50* indivíduos
- Variância: $s^2 = 148,33$
- Desvio-padrão: $s = \sqrt{148,33} = 12,18$
- Média: $\bar{x} = 60,93$
- Coeficiente de variação: $cv = \frac{s}{\bar{x}} = \frac{12,18}{60,93} = 19,99\%$

Atividade nº 5

Quartis e Percentis

Quartis

- Dividem o conjunto de dados em 4 partes iguais



- 1º Quartil (Q_1):

  25% dos dados estão abaixo (75% acima)
- 3º Quartil (Q_3):

  75% dos dados estão abaixo (25% acima)
- 2º Quartil:

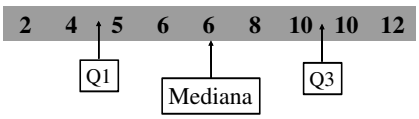
  É a mediana!

Procedimento para Determinação dos Quartis

- Várias definições são usadas na literatura e por diferentes pacotes computacionais
  - √ As diferentes definições dão respostas muito parecidas
- Regra que adotaremos:
  - √ O primeiro quartil (Q_1) é a mediana de todas as observações com posição estritamente abaixo da posição da mediana
  - √ O terceiro quartil (Q_3) é a mediana das observações que estão estritamente acima da posição da mediana.

- Determinação da mediana:

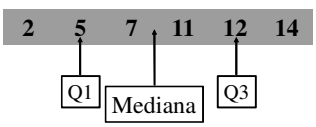
√ $n = 9$



$$Q_1 = \frac{4+5}{2} = 4,5$$

$$Q_3 = \frac{10+10}{2} = 10$$

√ $n = 6$



$$Q_1 = 5$$

$$Q_3 = 12$$

Exemplo – Peso

- Peso (kg)

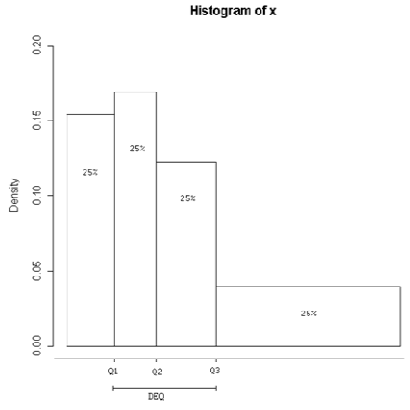
Mínimo	44,0	$x_{(1)} = 44,0$
Q_1	52,0	$x_{(13)}=52,0$
Q_2 = Mediana	58,0	$x_{(25)}=58,0$ $x_{(26)}=58,0$
Q_3	68,5	$x_{(38)}=68,5$
Máximo	95,0	$x_{(50)}=95,0$

Distância Interquartílica

- Medida de variabilidade dada por .

$$DI = Q_3 - Q_1$$

- Menos sensível a valores extremos que a amplitude e a variância (desvio-padrão)
- É uma medida um pouco mais refinada que a amplitude amostral.



Exemplo: Peso

- Peso (kg)

Q_1	52,0	$x_{(13)}=52,0$
Q_2 = Mediana	58,0	$x_{(25)}=58,0$ $x_{(26)}=58,0$
Q_3	68,5	$x_{(38)}=68,5$
Distância Interquartílica	**16,50**	$Q_3 - Q_1$

Box-plot

Esquema dos 5 Números

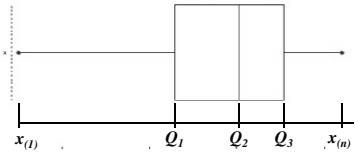
- São os cinco valores importantes para se ter uma boa ideia da assimetria dos dados.
- São as seguintes medidas da distribuição:
- $x_{(1)}$, Q_1, Q_2, Q_3 e $x_{(n)}$.

Esquema dos 5 Números (2)

- Para uma distribuição aproximadamente simétrica, tem-se:
  - √ $Q_2 - x_{(1)} \cong x_{(n)} - Q_2$;
  - √ $Q_2 - Q_1 \cong Q_3 - Q_2$;
  - √ $Q_1 - x_{(1)} \cong x_{(n)} - Q_3$;
  - √ distâncias entre mediana e Q1, mediana e Q3 menores do que distâncias entre os extremos e Q1 e Q3.

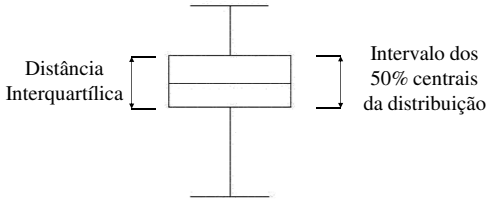
Box Plot

- A informação do esquema dos cinco números pode ser expressa num diagrama, conhecido como *box plot (gráfico-caixa).*
- Descreve várias características dos dados:
  - √ Centro, dispersão, simetria e valores atípicos



Box Plot (2)

- O retângulo é traçado de maneira que suas bases têm alturas correspondentes Q_1 e Q_3.
- Corta-se o retângulo por segmento paralelo às bases, na altura correspondente Q_2.
- O retângulo do *boxplot* corresponde aos 50% valores centrais da distribuição.

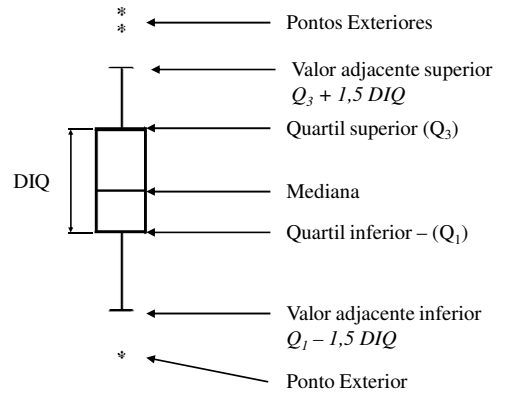


Região de Observações Típicas

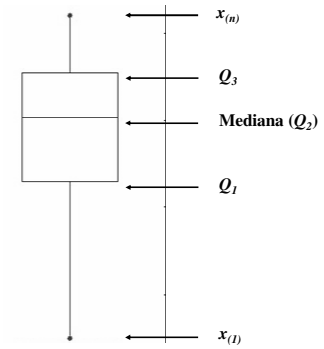
- Delimita-se a região que vai da base superior do retângulo até o maior valor observado que NÃO supere o valor de $Q_3+1,5 \times DIQ$.
- Procedimento similar para delimitar a região que vai da base inferior do retângulo, até o menor valor que NÃO é menor do que $Q_1-1,5 \times DIQ$.

Região de Observações Atípicas

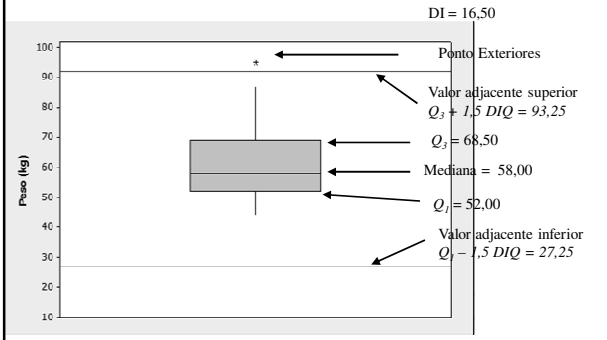
- Observações são representadas por asteriscos e situam-se:
  - √ ou, acima do Valor adjacente superior (*Q3 + 1,5 DIQ*)
  - √ ou, abaixo do Valor adjacente inferior (*Q1 – 1,5 DIQ*)
- Estes pontos exteriores são denominados *outliers* ou valores atípicos.



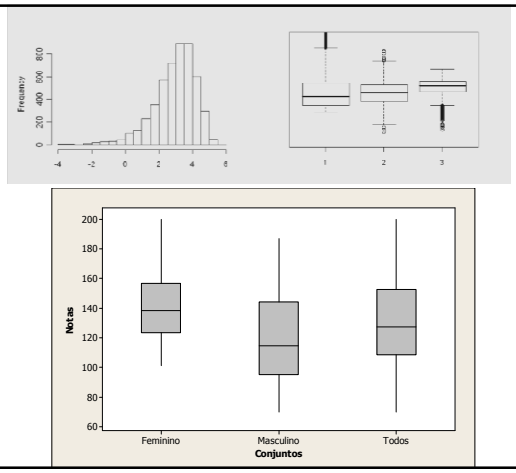
- Se não houver pontos exteriores:



Exemplo - Peso



Atividade nº 6



Referências

Bibliografia

- Wild, C.J. e Seber, G.A.F. (LTC) *Encontros com o Acaso: um Primeiro Curso de Análise de Dados e Inferência*
- Moore, D.S. e McCabe, G.P. (LTC) *Introdução à Prática da Estatística*
- Agresti, A. e Finlay, B. (Penso) *Métodos Estatísticos para as Ciências Sociais*