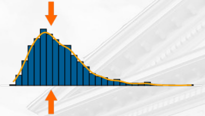
Medidas de posição

Prof. Walmes M. Zeviani

Departamento de Estatística
Universidade Federal do Paraná

Conteúdo

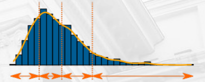
Medidas de posição



Expressam:
→ Posição central ocupada pela variável.
→ Valores que dividem a amostra em partes iguais.
→ Valores típicos onde há maior densidade.

São elas:
→ Média artimética.
→ Mediana.
→ Moda.
→ Média geométrica.
→ Média harmônica.
→ Média aparada.

Medidas de posição relativa



Expressam:
→ Pontos no domínio da variável que definem porções com frequências conhecidas.

São elas:
→ Quartis.
→ Decis.
→ Percentis.
→ Máximo e mínimo.

Figura 1. Medidas de posição usadas em análise descritiva de dados.

Medidas de posição

Média aritmética

- A **média aritmética** é soma de todos os valores dividida pela quantidade de valores.
- Tem interpretação física de **centro de gravidade**.
- A média é
$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$
- $\mu = \overline{y}$ é valor que minimiza a soma do quadrado dos desvios
$$\text{SQD}(\mu) = \sum_{i=1}^{n} (y_i - \mu)^2.$$
- É uma medida influenciada por **valores extremos** (*outliers*).

Cálculo e representação gráfica

Considere a seguinte amostra de dados

4	9	10	12	15
7	10	11	12	15
8	10	11	13	18
8	10	12	14	24

e determine a média.

Fazendo os cálculos:

- $\sum_{i=1}^{20} y_i = 233$.
- $\overline{y} = 233/20 = 11.65$.

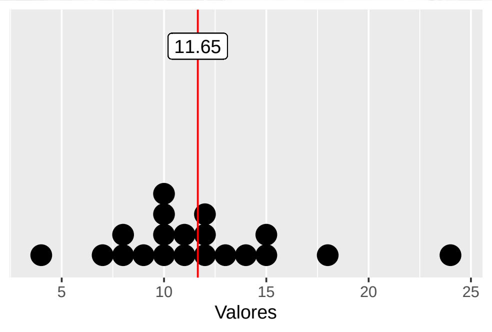


Figura 2. Gráfico de pontos empilhados de uma variável cujos valores estão na tabela ao lado.

Média aritmética ponderada (dados agrupados)

- Quando os dados estão **agrupados**, ou seja, quando se possuem as **frequências relativas** (f_r) de valores individuais (que se repetem) ou de classes, obtém-se a média considerando a ponderação pela frequência.
- A média ponderada é
$$\bar{y} = \frac{\sum_{i=1}^{k} f_i \cdot y_i}{\sum_{i=1}^{k} f_i},$$
em que f_i é a frequência da classe i (absoluta ou relativa) e k é o número de classes ($k \leq n$).
- No caso em que os valores foram agrupados em classe, usa-se como y_i o **ponto médio da classe**. Exemplo: na classe [10, 15] o ponto médio é 12,5.
- Note que, para o caso de dados individuais, $f_i = 1/n$ para todos os valores da variável y, e retorna-se à primeira expressão apresentada.

Outros tipos de média · **média geométrica**

- Definida como a n-ésima raíz do produto de n números, ou seja

$$m_g = \left(\prod_{i=1}^{n} y_i\right)^{1/n} = \sqrt[n]{y_1 y_2 \cdots y_n}.$$

- Tem relação com logaritmo.
- Usada para remover o efeito de escalas para comparar valores médios entre grupos.
- Muito usada para cálculo de retorno médio de **juros compostos**.
- Exemplo: um fundo de investimento apresentou as seguintes taxas de juros mensais: 0.643%, 0.487%, 0.797%, 0.327%, 0.487%. Qual é a taxa de juros média do período?

$$m_g = (1.00643 \cdot 1.00487 \cdot 1.00797 \cdot 1.00327 \cdot 1.00487)^{1/5} = 1.00548.$$

Portanto a rentabilidade média é de 0.548%

Outros tipos de média · **média harmônica**

- A **média harmônica** é a recíproca da média aritmética dos recíprocos, definida por

$$m_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{y_i}}, \text{ sendo a versão ponderada } m_h = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{y_i}}.$$

- Ela é usada para calcular médias sobre valores que são taxas ou tem **relação divisiva**.
- Exemplo: 3 amigos dirigem 100 km cada um mantendo a velocidade de 50, 65 e 75 km/h em cada trecho. Qual é a velocidade média da viagem?

$$m_h = \frac{100 + 100 + 100}{\frac{100}{50} + \frac{100}{65} + \frac{100}{75}} = 60.$$

Outros tipos de média · **média aparada**

- A **média aparada** é usada para evitar o efeito dos **valores extremos**.
- Uma média 10% aparada é obtida ao se descartar 5% dos valores em cada extremidade.
- Se a amostra é de tamanho 100, significa descartar os 5 menores e os 5 maiores valores e com os 90 restantes, calcular a média.
- A média aparada considera, portanto, o conceito de **medidas de posição relativa**.

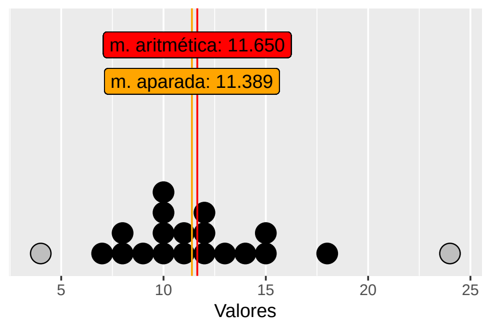


Figura 3. Gráfico de pontos empilhados indicando as observações usadas para o cálculo da média 10% aparada e comparação com a média aritmética.

Mediana

- A mediana é o número que ocupa a **posição intermediária** quando os valores são ordenados.
- Separa o conjunto de valores em **duas partes de mesmo tamanho**. Assim, se todos os valores na amostra forem distintos, metade dos valores é menor que a mediana e metade é maior que ela.
- Indica-se que a amostra está ordenada usando a notação de parênteses no índice

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n-1)} \leq y_{(n)}.$$

- O valor de y na k-ésima posição, $y_{(k)}$, é chamado de k-ésima estatística de ordem.
- A mediana é calculada por

$$md = \begin{cases} y_{((n+1)/2)}, & \text{se } n \text{ for ímpar} \\ (y_{(n/2)} + y_{(n/2+1)})/2, & \text{se } n \text{ for par.} \end{cases}$$

Mediana

n ímpar

mediana

$$y_{((n+1)/2)} = y_{(6)}$$

$$\underbrace{y_{(1)} \quad y_{(2)} \quad y_{(3)} \quad y_{(4)} \quad y_{(5)}}_{1^{\underline{\text{a}}}\text{ metade}} \quad y_{(6)} \quad \underbrace{y_{(7)} \quad y_{(8)} \quad y_{(9)} \quad y_{(10)} \quad y_{(11)}}_{2^{\underline{\text{a}}}\text{ metade}}$$

n par

mediana

$$(y_{(n/2)} + y_{(n/2+1)})/2 = (y_{(6)} + y_{(7)})/2$$

$$\underbrace{y_{(1)} \quad y_{(2)} \quad y_{(3)} \quad y_{(4)} \quad y_{(5)} \quad y_{(6)}}_{1^{\underline{\text{a}}}\text{ metade}} \quad \underbrace{y_{(7)} \quad y_{(8)} \quad y_{(9)} \quad y_{(10)} \quad y_{(11)} \quad y_{(12)}}_{2^{\underline{\text{a}}}\text{ metade}}$$

Figura 4. Cálculo da mediana para as 2 situações possíveis conforme o tamanho da amostra.

Cálculo e interpretação gráfica

Considere a seguinte amostra de dados

4	9	10	12	15
7	10	**11**	12	15
8	10	**11**	13	18
8	10	12	14	24

e determine a mediana.

Fazendo os cálculos:

- $n = 20$ é par.
- $md = (y_{(10)} + y_{(11)})/2 = (11 + 11)/2 = 11.$

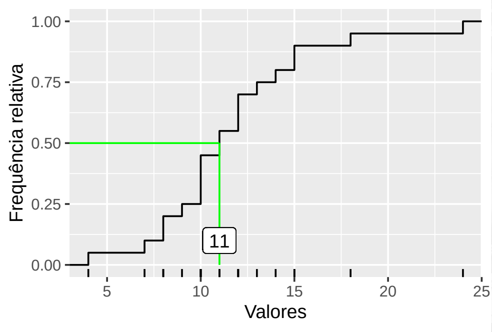


Figura 5. Gráfico de frequência relativa acumulada da variável cujos valores estão na tabela ao lado.

Moda

- Moda é o valor ou classe que ocorre com **maior frequência** (ou densidade) na amostra.
- A moda representa o valor mais típico, ou seja, **o que mais se repete**.
- Para variáveis onde todos os valores são distintos, a moda fica **indefinida** já que a frequência é $1/n$ para todos os valores de y.
- Pode-se agrupar os dados em classe e reportar a **classe modal**.
- Mas pode-se determinar a moda como sendo o valor que corresponde ao **máximo da densidade empírica**.

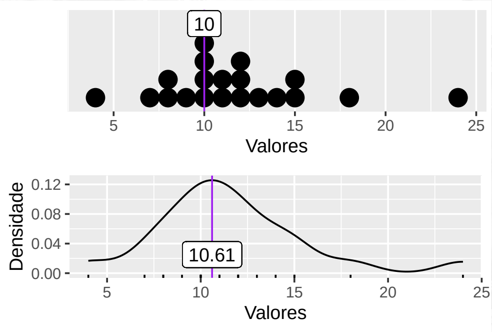


Figura 6. Gráfico de pontos empilhados da variável e gráfico de densidade empírica. Ambos indicam o valor da moda para a mesma amostra.

Média, moda e mediana em relação à assimetria

- Assimetria à direita:

$$\text{moda} < \text{mediana} < \text{média}.$$

- Assimetria à esquerda:

$$\text{média} < \text{mediana} < \text{moda}.$$

- Para memorizar.
  - A moda está na região de maior densidade.
  - Como a média é "puxada" pelos valores extremos, encontra-se para o lado da cauda longa.
  - A mediana está entre a moda e a média.

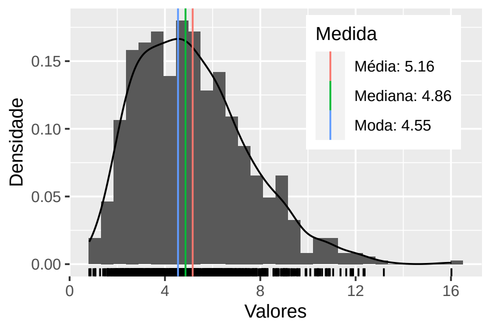


Figura 7. Histograma de frequência da variável e o ordenamento nos valores das medidas de posição média, mediana e moda.

Quando usar cada medida de posição

Candidato: Quanto ganha um funcionário da empresa?

Entrevistador: Você quer saber o que exatamente?

- O salário **médio**?
- O salário **intermediário**?
- Ou o salário **típico**?

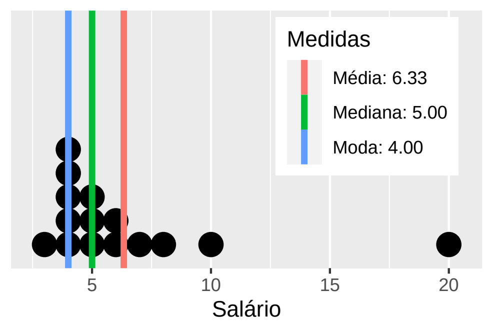


Figura 8. Salários dos funcionários de uma empresa. $n = 15$.

Quando usar cada medida de posição

- **Média**: distribuição unimodal simétrica e sem valores extremos.
- **Mediana**: distribuição assimétrica ou com presença de valores extremos.
- **Moda**: quando valores se repetem, estão agrupados em classe ou é variável qualitativa.
- As três medidas:
  - Perdem significado em distribuições multimodais.
  - Aproximam-se em distribuições unimodais simétricas.
- **Sempre faça gráficos!**

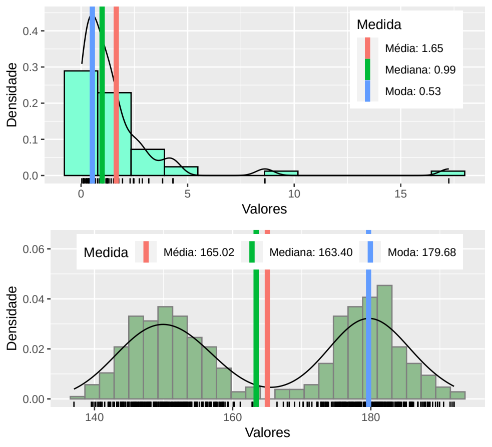


Figura 9. Média, mediana e moda para distribuição assimétrica (topo) e distribuição bimodal (base).

Medidas de posição relativa (separatrizes)

Medidas de posição relativa (separatrizes)

- Descrevem **posição relativa**, em termos de frequência, de um particular valor na amostra.
- Por isso, as separatrizes também são chamadas de medidas de posição relativa.
- São de importância prática 3 tipos de separatrizes.
  - Os **quartis**: dividem a amostra em 4 partes com frequência 1/4.
  - Os **decis**: dividem a amostra em 10 partes com frequência 1/10.
  - Os **percentis**: dividem a amostra em 100 partes com frequência 1/100.

Quartis

- Um quartil qº ($q \in \{1, 2, 3\}$) de um conjunto de n valores (distintos), ordenados em ordem crescente, é um número tal que $(100q/4)$% se localizam abaixo dele.
- Dessa forma, tem-se 1º, 2º e 3º quartis.
- O 2º quartil é a mediana.
- O gráfico de caixas e bigodes é uma representação gráfica baseada nos quartis.

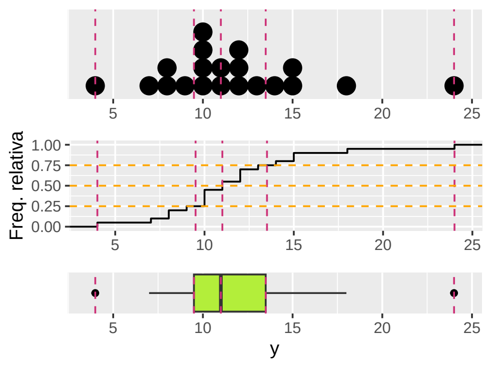


Figura 10. Gráfico de pontos empilhados (topo), gráfico de frequências relativas (meio) e gráfico de caixas e bigodes (base). Linhas verticais indicam os quartis e valores extremos.

Cálculo dos quartis

- Pode-se calcular os quartis 1 e 3 repetindo-se o procedimento de cálculo da mediana, mas aplicado a cada uma das metades da amostra.
- Cálculo do 1º quartil é a mediana da primeira porção 4, 7, 8, 8, 9, 10, 10, 10, 10, 11. Logo, $q_1 = 9.5$.
- Cálculo do 3º quartil é a mediana da segunda porção 11, 12, 12, 12, 13, 14, 15, 15, 18, 24. Logo, $q_3 = 13.5$.
- Apesar de simples, essa forma de calcular não é a única.
- Existem pelo menos 9 formas diferentes de calcular.
- O importante é que a diferença entre elas se torna irrelevante à medida que a amostra é maior.

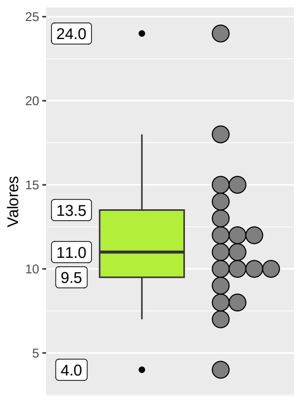


Figura 11. Gráfico de caixas com anotações dos valores das separações e diagrama de pontos incluído.

Ilustração do cálculo dos quartis

n múltiplo de 4

$q_1 \quad q_2 \quad q_3$

$$\underbrace{y_{(1)}\; y_{(2)}\; y_{(3)}\; y_{(4)}\; y_{(5)}\; y_{(6)}}_{1^{\text{a}}\text{ metade}}\; \underbrace{y_{(7)}\; y_{(8)}\; y_{(9)}\; y_{(10)}\; y_{(11)}\; y_{(12)}}_{2^{\text{a}}\text{ metade}}$$

$n + 1$ múltiplo de 4

$q_1 \quad q_2 \quad q_3$

$$\underbrace{y_{(1)}\; y_{(2)}\; y_{(3)}\; y_{(4)}\; y_{(5)}\; y_{(6)}}_{1^{\text{a}}\text{ metade}}\; \underbrace{y_{(6)}\; y_{(7)}\; y_{(8)}\; y_{(9)}\; y_{(10)}\; y_{(11)}}_{2^{\text{a}}\text{ metade}}$$

n múltiplo de 2

$q_1 \quad q_2 \quad q_3$

$$\underbrace{y_{(1)}\; y_{(2)}\; y_{(3)}\; y_{(4)}\; y_{(5)}}_{1^{\text{a}}\text{ metade}}\; \underbrace{y_{(6)}\; y_{(7)}\; y_{(8)}\; y_{(9)}\; y_{(10)}}_{2^{\text{a}}\text{ metade}}$$

$n + 1$ múltiplo de 2

$q_1 \quad q_2 \quad q_3$

$$\underbrace{y_{(1)}\; y_{(2)}\; y_{(3)}\; y_{(4)}\; y_{(5)}}_{1^{\text{a}}\text{ metade}}\; \underbrace{y_{(5)}\; y_{(6)}\; y_{(7)}\; y_{(8)}\; y_{(9)}}_{2^{\text{a}}\text{ metade}}$$

Figura 12. Cálculo dos quartis para as 4 situações possíveis conforme o tamanho da amostra pelo método dos 5 números de Tukey (*Tukey's hinge method*).

Amplitude interquartílica

- A amplitude interquartílica é a distância entre o q_1 e q_3, ou seja
$$\text{AIQ} = q_3 - q_1.$$
- A partir da AIQ e dos quartis q_1 e q_3 são delimitados valores limites, além dos quais as observações são representadas isoladamente. Esses valores são
$$q_1 - k \cdot \text{AIQ} \quad \text{e} \quad q_3 + k \cdot \text{AIQ},$$
em que k é uma constante amplamente utilizada com o valor 1,5.
- No gráfico ao lado foi usada outra forma de determinar os quartis q_1 e q_3.

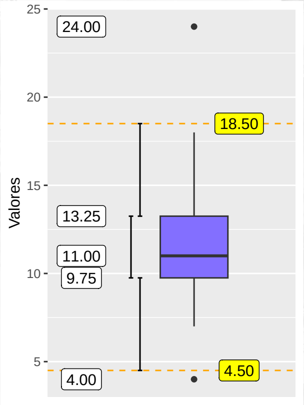


Figura 13. Linhas limítrofes para destaque de pontos individuais. Caixas feitas com outro método de determinação de quartis.

Percentis e decis

- Um conjunto de n valores, organizados de forma crescente, o P-ésimo percentil é um número tal que P% dos valores estejam à sua esquerda e (100 - P)% à sua direita.
- Os decis nada mais são que os **percentis múltiplos de 10**.
- Da mesma forma que os quartis são **percentis múltiplos de 25**.
- As separatrizes podem ser obtidas por meio do gráfico de frequências acumuladas.

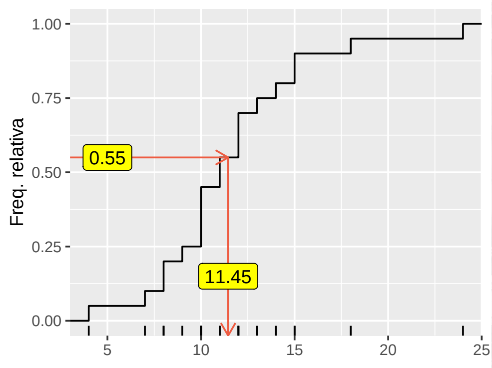


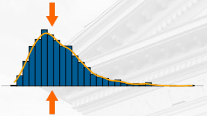
Figura 14. Gráfico de probabilidades acumulada indicando o uso para determinação de percentis.

Considerações finais

Considerações finais

Revisão

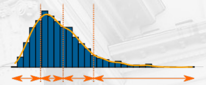
Medidas de posição



Expressam:
- → Posição central ocupada pela variável.
- → Valores que dividem a amostra em partes iguais.
- → Valores típicos onde há maior densidade.

São elas:
- → Média artimética.
- → Mediana.
- → Moda.
- → Média geométrica.
- → Média harmônica.
- → Média aparada.

Medidas de posição relativa



Expressam:
- → Pontos no domínio da variável que definem porções com frequências conhecidas.

São elas:
- → Quartis.
- → Decis.
- → Percentis.
- → Máximo e mínimo.

Figura 15. Medidas de posição usadas em análise descritiva de dados.