SME0803 Visualização e Exploração de Dados

Medidas descritivas de dados quantitativos

Prof. Cibele Russo

cibele@icmc.usp.br

Baseado em

Murteira, B. J. F., Análise Exploratória de Dados. McGraw-Hill, Lisboa, 1993.

Notas de aula de Mário de Castro, 2010.

Medidas-resumo

Objetivo

Reduzir *n* observações a **medidas** que ajudem a representar características importantes dos dados.

Medidas-resumo

Medidas de posição ou localização

- **Média**: boas propriedades estatísticas.
  Média aritmética, média aparada, média ponderada, média geométrica, média harmônica.
- **Mediana**: medida resistente a dados atípicos.
- **Moda**: valor mais frequente.
- **Quantis**: caracterização da distribuição dos dados.

Medidas-resumo

Medidas de posição ou localização

- **Média**: boas propriedades estatísticas.
  Média aritmética, média aparada, média ponderada, média geométrica, média harmônica.
- **Mediana**: medida resistente a dados atípicos.
- **Moda**: valor mais frequente.
- **Quantis**: caracterização da distribuição dos dados.

Medidas de dispersão

- **Desvio-padrão** .
- **Variância**.
- **Amplitude**.
- **Coeficiente de variação:** medida de dispersão relativa.

Medidas-resumo

Medidas de Assimetria:

- Assimetria da distribuição dos dados.

Medidas-resumo

Medidas de Assimetria:

- Assimetria da distribuição dos dados.

Medidas de Curtose:

- Achatamento da distribuição.

Medidas-resumo

Medidas de Assimetria:

- Assimetria da distribuição dos dados.

Medidas de Curtose:

- Achatamento da distribuição.

Medidas de associação:

- Covariância.
- Coeficiente de correlação de Pearson.
- Coeficiente de correlação de Spearman.

Medidas de posição: média

Em geral, não é possível calcular a média populacional de uma variável, μ. Usa-se então um **estimador**, por exemplo a média amostral, ou seja, a média que será obtida de uma amostra (representativa) da população (**estimativa**).

Vamos estabelecer que $X_1, \ldots, X_n$ é uma amostra aleatória e $x_1, \ldots, x_n$ os dados observados dessa amostra. As medidas aqui apresentadas são **amostrais** e são obtidas a partir de $x_1, \ldots, x_n$.

Medidas de posição: média

Em geral, não é possível calcular a média populacional de uma variável, μ. Usa-se então um **estimador**, por exemplo a média amostral, ou seja, a média que será obtida de uma amostra (representativa) da população (**estimativa**).

Vamos estabelecer que $X_1, \ldots, X_n$ é uma amostra aleatória e $x_1, \ldots, x_n$ os dados observados dessa amostra. As medidas aqui apresentadas são **amostrais** e são obtidas a partir de $x_1, \ldots, x_n$.

A **média** (amostral observada, *mean*) é definida como

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Propriedades da média (aritmética)

P1 Se $y_1, \ldots, y_n$ são tais que $y_i = a + bx_i$, para $i = 1, \ldots, n$, com a e b constantes, então $\bar{y} = a + b\bar{x}$.

Propriedades da média (aritmética)

P1 Se $y_1, \ldots, y_n$ são tais que $y_i = a + bx_i$, para $i = 1, \ldots, n$, com a e b constantes, então $\bar{y} = a + b\bar{x}$.

P2 A média é o centro de massa (ou centro de gravidade) dos dados: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Propriedades da média (aritmética)

P1 Se $y_1, \ldots, y_n$ são tais que $y_i = a + bx_i$, para $i = 1, \ldots, n$, com a e b constantes, então $\bar{y} = a + b\bar{x}$.

P2 A média é o centro de massa (ou centro de gravidade) dos dados:
$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

P3 Se $x_1, \ldots, x_k$ são os valores únicos que X assume na amostra, com frequências $f_1, \ldots, f_k$, respectivamente, então
$$\sum_{j=1}^{k} f_j(x_j - \bar{x}) = 0.$$

Propriedades da média (aritmética)

P1 Se $y_1, \ldots, y_n$ são tais que $y_i = a + bx_i$, para $i = 1, \ldots, n$, com a e b constantes, então $\bar{y} = a + b\bar{x}$.

P2 A média é o centro de massa (ou centro de gravidade) dos dados:
$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

P3 Se $x_1, \ldots, x_k$ são os valores únicos que X assume na amostra, com frequências $f_1, \ldots, f_k$, respectivamente, então
$$\sum_{j=1}^{k} f_j(x_j - \bar{x}) = 0.$$

P4 A quantidade $\sum_{i=1}^{n}(x_i - \nu)^2$ é minimizada se, e somente se $\nu = \bar{x}$.

Propriedades da média (aritmética)

Exercícios:

1. Demonstre as propriedades P1-P4.

Propriedades da média (aritmética)

Exercícios:

1. Demonstre as propriedades P1-P4.
2. O que acontece com a média se os dados forem acrescidos de duas unidades?

Propriedades da média (aritmética)

Exercícios:

1. Demonstre as propriedades P1-P4.
2. O que acontece com a média se os dados forem acrescidos de duas unidades?
3. O que acontece com a média se os dados forem multiplicados por 2?

Propriedades da média (aritmética)

Exercícios:

1. Demonstre as propriedades P1-P4.
2. O que acontece com a média se os dados forem acrescidos de duas unidades?
3. O que acontece com a média se os dados forem multiplicados por 2?
4. Que transformação devemos fazer para que os dados transformados tenham média zero?

Propriedades da média (aritmética)

Vantagens da média:

- É uma medida conhecida.
- Tem boas propriedades estatísticas.
- Facilidade de cálculo.

Propriedades da média (aritmética)

Vantagens da média:

- É uma medida conhecida.
- Tem boas propriedades estatísticas.
- Facilidade de cálculo.

Desvantagens da média:

- É muito influenciada por valores atípicos.
- Bastante afetada por distribuições assimétricas.
- Só pode ser calculada para dados quantitativos.
- Nem sempre pode ser calculada.

Outras propostas para a média

Média aparada: (média truncada, tri-média ou *Winsorized mean*)

Média aparada de $100\alpha\%$: média aritmética dos dados após a eliminação das $100\alpha\%$ menores e das $100\alpha\%$ maiores observações, com $0 < \alpha < 1/2$.

É uma medida mais resistente a valores atípicos do que a média aritmética. Considere agora os dados ordenados $x_{(1)}, \ldots, x_{(n)}$.

Outras propostas para a média

Média aparada: (média truncada, tri-média ou *Winsorized mean*)

Média aparada de $100\alpha\%$: média aritmética dos dados após a eliminação das $100\alpha\%$ menores e das $100\alpha\%$ maiores observações, com $0 < \alpha < 1/2$.

É uma medida mais resistente a valores atípicos do que a média aritmética. Considere agora os dados ordenados $x_{(1)}, \ldots, x_{(n)}$.

Se $n = 20$ e $\alpha = 0,1$, então eliminamos as duas menores e as duas maiores medidas para calcular a média aparada:

$$\bar{x}_\alpha = \frac{x_{(3)} + x_{(4)} + \ldots + x_{(17)} + x_{(18)}}{16}.$$

Outras propostas para a média

Média ponderada: (weighted mean)

$$\bar{x}_p = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$, em que w_i é o peso para a i-ésima observação.

Outras propostas para a média

Média ponderada: (weighted mean)

$$\bar{x}_p = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$, em que w_i é o peso para a i-ésima observação.

Em particular, se existem somente k valores únicos na amostra e f_i é a frequência absoluta da observação x_i, para $i = 1, \ldots, k$, então a média ponderada é obtida por

$$\bar{x}_p = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{n} f_i} =$$

Outras propostas para a média

Média ponderada: (weighted mean)

$$\bar{x}_p = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$, em que w_i é o peso para a i-ésima observação.

Em particular, se existem somente k valores únicos na amostra e f_i é a frequência absoluta da observação x_i, para $i = 1, \ldots, k$, então a média ponderada é obtida por

$$\bar{x}_p = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{k} f_i x_i}{n}.$$

Outras propostas para a média

Média geométrica: (para valores positivos)

$$\bar{x}_g = \left\{ \prod_{i=1}^{n} x_i \right\}^{1/n} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \times \ldots \times x_n}.$$

Média harmônica: (para grandezas inversamente proporcionais)

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

Exemplo - Crescimentos proporcionais

Suponha que uma laranjeira produz 100 laranjas em um ano, e então 180, 210 e 300 nos anos seguintes, então os crescimentos anuais são 80%, 16,66% e 42,86%.

Exemplo - Crescimentos proporcionais

Suponha que uma laranjeira produz 100 laranjas em um ano, e então 180, 210 e 300 nos anos seguintes, então os crescimentos anuais são 80%, 16,66% e 42,86%.

A **média aritmética** dos crescimentos é 46,51%. Todavia, crescimentos acumulados de 46,51% ao ano ano resultaria em 314 laranjas, não 300.

Exemplo - Crescimentos proporcionais

Suponha que uma laranjeira produz 100 laranjas em um ano, e então 180, 210 e 300 nos anos seguintes, então os crescimentos anuais são 80%, 16,66% e 42,86%.

A **média aritmética** dos crescimentos é 46,51%. Todavia, crescimentos acumulados de 46,51% ao ano ano resultaria em 314 laranjas, não 300.

Nesse caso, o correto é usar a **média geométrica**. A média geométrica de 1,80, 1,17 e 1,43 é $\sqrt[3]{1,80 \times 1,17 \times 1,43} = 1,443$; logo o crescimento médio por ano é 44,3%. Começando com 100 laranjas e crescimento de 44,3% cada ano, o resultado é 300 laranjas.

Fonte: `https://pt.wikipedia.org/wiki/M%C3%A9dia_geom%C3%A9trica`

Exemplo - média aritmética e média harmônica

Um veículo percorre 100km a uma velocidade de 60 km/h e, em seguida, a mesma distância a uma velocidade 40 km/h. A sua **velocidade média** é a **média harmônica** entre 60 e 40 (48 km/h), e seu tempo total de viagem é o mesmo como se tivesse viajado toda a distância com essa velocidade média.

No entanto, se o veículo se desloca por uma hora a uma velocidade 60km/h e em seguida uma hora, a uma velocidade de 40 km/h, a sua **velocidade média** é a **média aritmética** das velocidades, o que no exemplo acima é de 50 km/h.

Fonte: `https://pt.wikipedia.org/wiki/M%C3%A9dia_harm%C3%B4nica`

Desigualdade das médias

Se $x_1, \ldots x_n$ são todos positivos, então a média harmônica é menor ou igual que a média geométrica, que por sua vez é menor ou igual do que a média aritmética, ou seja,

$$\bar{x}_h \leq \bar{x}_g \leq \bar{x}.$$

Fórmula geral das médias

As médias aritmética, geométrica e harmônica são casos particulares da fórmula geral das médias (para dados positivos):

$$\bar{x}_q = \left\{ \frac{\sum_{i=1}^{n} x_i^q}{n} \right\}^{1/q}.$$

Exercício: Para que valores de q obtemos a média aritmética, geométrica e harmônica?

Medidas de posição: mediana

Considere agora os dados ordenados $x_{(1)}, \ldots, x_{(n)}$, isto é,

$$x_{(1)} = min(x_1, \ldots, x_n) \text{ e } x_{(n)} = max(x_1, \ldots, x_n).$$

Qual é a posição central dos dados?

Medidas de posição: mediana

Considere agora os dados ordenados $x_{(1)}, \ldots, x_{(n)}$, isto é,

$$x_{(1)} = min(x_1, \ldots, x_n) \text{ e } x_{(n)} = max(x_1, \ldots, x_n).$$

Qual é a posição central dos dados?

Se n é ímpar, a posição central é $c = (n+1)/2$.
Se n é par, as posições centrais são $c = n/2$ e $c + 1 = n/2 + 1$.

Medidas de posição: mediana

Considere agora os dados ordenados $x_{(1)}, \ldots, x_{(n)}$, isto é,

$$x_{(1)} = min(x_1, \ldots, x_n) \text{ e } x_{(n)} = max(x_1, \ldots, x_n).$$

Qual é a posição central dos dados?

Se n é ímpar, a posição central é $c = (n+1)/2$.
Se n é par, as posições centrais são $c = n/2$ e $c + 1 = n/2 + 1$.

A **mediana** é definida como

$$Md = \begin{cases} x_{(c)}, \text{se n é ímpar} \\ \dfrac{x_{(c)} + x_{(c+1)}}{2}, \text{se n é par} \end{cases}$$

Propriedades da mediana

Vantagens da mediana:

- Mais resistente a valores atípicos.
- Pouco afetada por distribuições assimétricas.
- Pode ser obtida para variáveis qualitativas ordinais. Exemplo: ruim, ruim, ruim, ruim, bom, bom. Mediana: ruim.

Desvantagens da mediana:

- Menos conhecida que a média.
- Não tem boas propriedades estatísticas.

Medidas de posição: moda e quantis

A **moda** é o valor mais frequente da amostra.

Uma amostra pode ter uma moda, mais de uma moda, ou a moda pode não existir.

Medidas de posição: moda e quantis

A **moda** é o valor mais frequente da amostra.
Uma amostra pode ter uma moda, mais de uma moda, ou a moda pode não existir.

Exemplo:
Sejam os dados observados (ordenados): 1; 2; 2; 3; 3; 4; 4; 4; 5; 40.

Medidas de posição: moda e quantis

A **moda** é o valor mais frequente da amostra.
Uma amostra pode ter uma moda, mais de uma moda, ou a moda pode não existir.

Exemplo:
Sejam os dados observados (ordenados): 1; 2; 2; 3; 3; 4; 4; 4; 5; 40.

média: 6,8
mediana: 3,5
moda: 4

Medidas de posição: quantil ou separatriz

Um **quantil** é o valor que provoca uma divisão conveniente nos valores **ordenados**.

Medidas de posição: quantil ou separatriz

Um **quantil** é o valor que provoca uma divisão conveniente nos valores **ordenados**.

O **quantil** de 10%, q_{10} divide os dados de tal forma que 10% dos menores valores sejam menores que ele.

Medidas de posição: quantil ou separatriz

Um **quantil** é o valor que provoca uma divisão conveniente nos valores **ordenados**.

O **quantil** de 10%, q_{10} divide os dados de tal forma que 10% dos menores valores sejam menores que ele.

O **quantil** de 50%, q_{50} é a **mediana**.

Medidas de posição: quantil ou separatriz

Um **quantil** é o valor que provoca uma divisão conveniente nos valores **ordenados**.

O **quantil** de 10%, q_{10} divide os dados de tal forma que 10% dos menores valores sejam menores que ele.

O **quantil** de 50%, q_{50} é a **mediana**.

Os **quartis** Q_1, Q_2 e Q_3 dividem os dados em porções de 25%.

Medidas de posição: quantil ou separatriz

Um **quantil** é o valor que provoca uma divisão conveniente nos valores **ordenados**.

O **quantil** de 10%, q_{10} divide os dados de tal forma que 10% dos menores valores sejam menores que ele.

O **quantil** de 50%, q_{50} é a **mediana**.

Os **quartis** Q_1, Q_2 e Q_3 dividem os dados em porções de 25%.

Os **decis** dividem os dados em porções de 10%, d_α.

Medidas de posição: quantil ou separatriz

Um **quantil** é o valor que provoca uma divisão conveniente nos valores **ordenados**.

O **quantil** de 10%, q_{10} divide os dados de tal forma que 10% dos menores valores sejam menores que ele.

O **quantil** de 50%, q_{50} é a **mediana**.

Os **quartis** Q_1, Q_2 e Q_3 dividem os dados em porções de 25%.

Os **decis** dividem os dados em porções de 10%, d_α.

Os **percentis** dividem os dados em porções de 1%, p_α.

Medidas de posição: quantil

Exemplo: dados observados 1; 2; 2; 3; 3; 4; 4; 4; 5; 40.

Medidas de posição: quantil

Exemplo: dados observados 1; 2; 2; 3; 3; 4; 4; 4; 5; 40.

quantil de 10%: q_{10}=1

Medidas de posição: quantil

Exemplo: dados observados 1; 2; 2; 3; 3; 4; 4; 4; 5; 40.

quantil de 10%: q_{10}=1

quantil de 20%: q_{20}=2

Medidas de posição: quantil

Exemplo: dados observados 1; 2; 2; 3; 3; 4; 4; 4; 5; 40.

quantil de 10%: q_{10}=1

quantil de 20%: q_{20}=2

primeiro quartil: Q_1=2

Medidas de posição: quantil

Exemplo: dados observados 1; 2; 2; 3; 3; 4; 4; 4; 5; 40.

quantil de 10%: q_{10}=1

quantil de 20%: q_{20}=2

primeiro quartil: Q_1=2

segundo quartil: Q_2 = 3,5

Medidas de posição: quantil

Exemplo: dados observados 1; 2; 2; 3; 3; 4; 4; 4; 5; 40.

quantil de 10%: q_{10}=1

quantil de 20%: q_{20}=2

primeiro quartil: Q_1=2

segundo quartil: $Q_2 = 3{,}5$

terceiro quartil: $Q_3 = 4$

Medidas de posição: quantil

Exemplo: dados observados 1; 2; 2; 3; 3; 4; 4; 4; 5; 40.

quantil de 10%: q_{10}=1

quantil de 20%: q_{20}=2

primeiro quartil: Q_1=2

segundo quartil: $Q_2 = 3{,}5$

terceiro quartil: $Q_3 = 4$

Medidas de dispersão

A **variância** (amostral) é dada por

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

Medidas de dispersão

A **variância** (amostral) é dada por

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

O **desvio padrão** (amostral) é dado por

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

Medidas de dispersão

Observação:

Se a variável original X está em 'unidades', a variância está em 'unidades^2' e o desvio-padrão está em 'unidades'.

Medidas de dispersão

Observação:

Se a variável original X está em 'unidades', a variância está em 'unidades^2' e o desvio-padrão está em 'unidades'.

E a média? E a mediana?

Medidas de dispersão

É comum, entretanto, utilizar as medidas corrigidas:

Variância amostral corrigida:

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

Desvio padrão corrigido:

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

Medidas de dispersão

Sejam $x_{(1)}, \ldots, x_{(n)}$ os dados ordenados, ou seja, $x_{(1)} = min\{x_1, \ldots, x_n\}$ e $x_{(n)} = max\{x_1, \ldots, x_n\}$.

Medidas de dispersão

Sejam $x_{(1)}, \ldots, x_{(n)}$ os dados ordenados, ou seja, $x_{(1)} = min\{x_1, \ldots, x_n\}$ e $x_{(n)} = max\{x_1, \ldots, x_n\}$.

A **amplitude** é dada por

$$A = x_{(n)} - x_{(1)}.$$

Medidas de dispersão

Sejam $x_{(1)}, \ldots, x_{(n)}$ os dados ordenados, ou seja,
$x_{(1)} = min\{x_1, \ldots, x_n\}$ e $x_{(n)} = max\{x_1, \ldots, x_n\}$.

A **amplitude** é dada por

$$A = x_{(n)} - x_{(1)}.$$

A **amplitude interquartil** é dada por

$$AIQ = Q_3 - Q_1,$$

em que Q_1 é o primeiro quartil e Q_3 é o terceiro quartil da amostra.

Medidas de dispersão

O **coeficiente de variação** (amostral) é dado pela razão entre o desvio-padrão e a média

$$CV = \frac{s}{\bar{x}}$$

Medidas de dispersão

O **coeficiente de variação** (amostral) é dado pela razão entre o desvio-padrão e a média

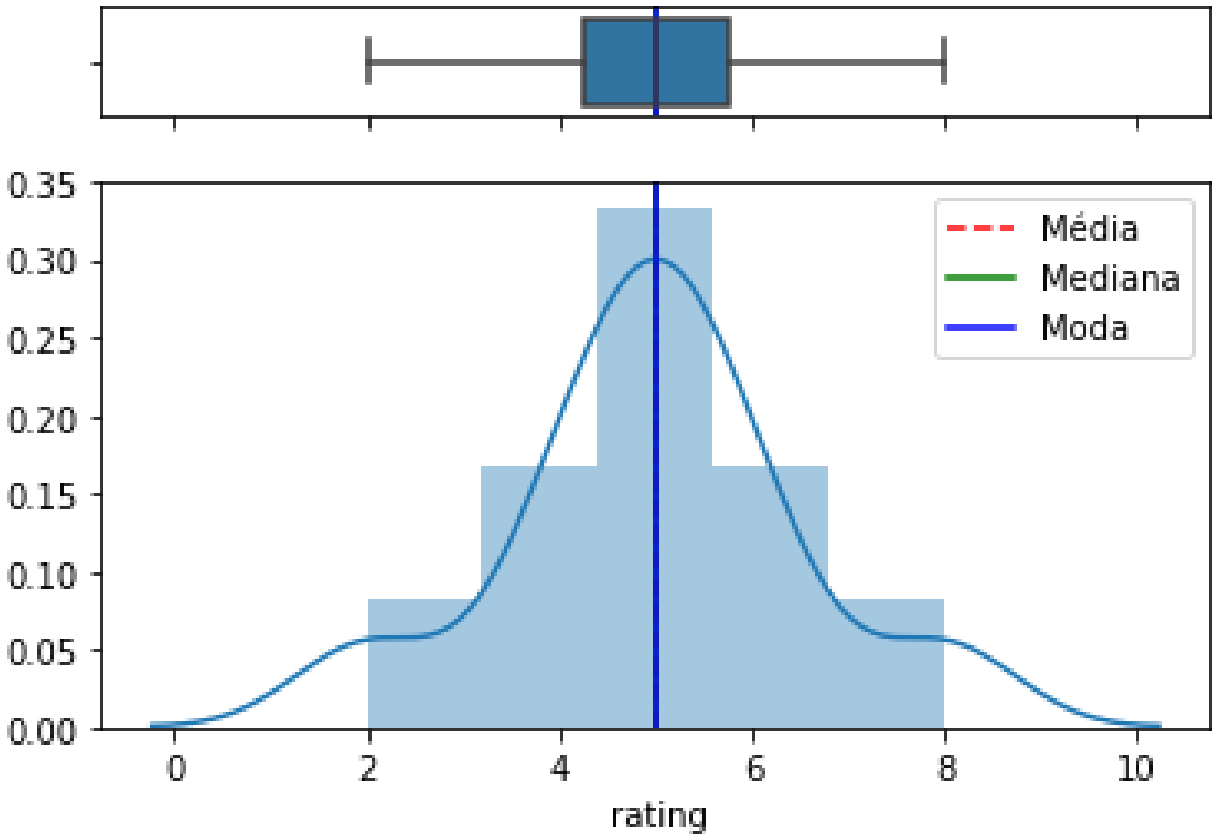
$$CV = \frac{s}{\bar{x}}$$

Em que unidade está o coeficiente de variação?

Medidas de assimetria

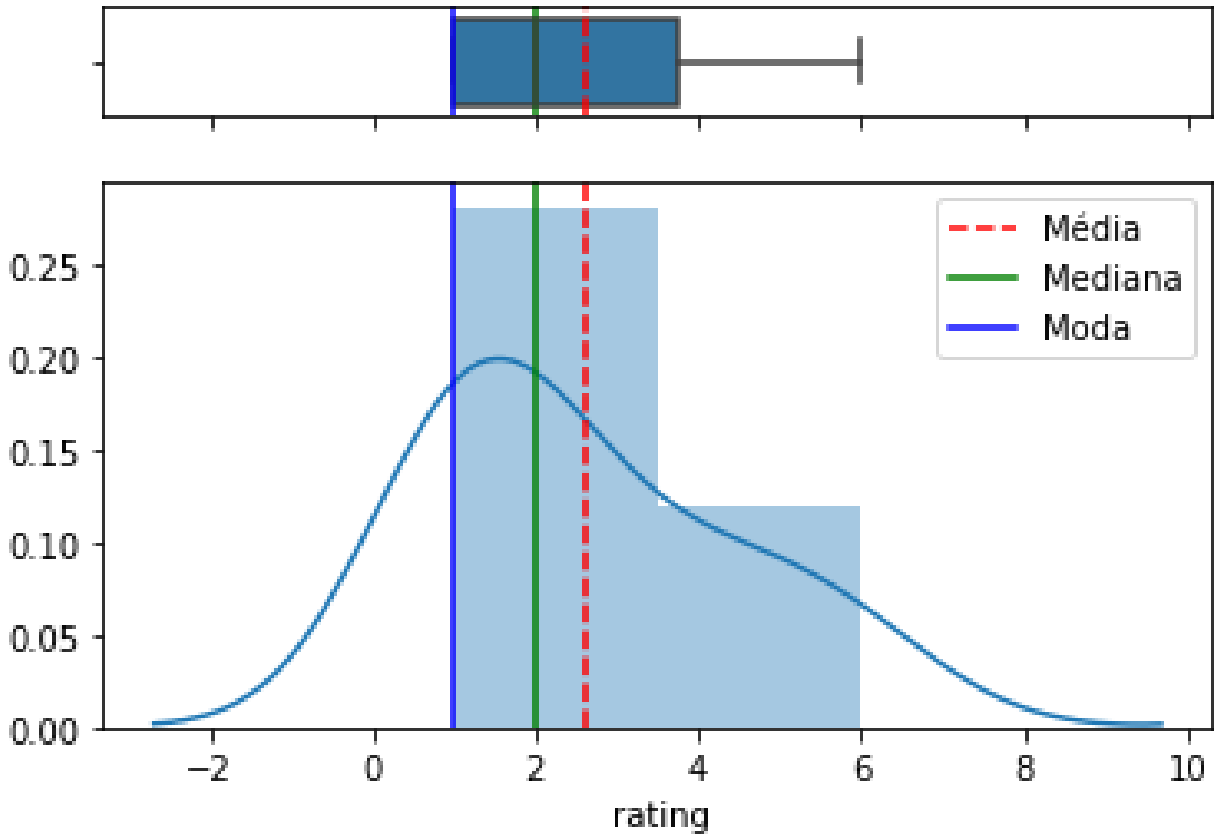
- **Distribuição simétrica:** média = mediana = moda
- **Distribuição assimétrica à direita:** moda < mediana < média
- **Distribuição assimétrica à esquerda:** média < mediana < moda

Medidas de assimetria



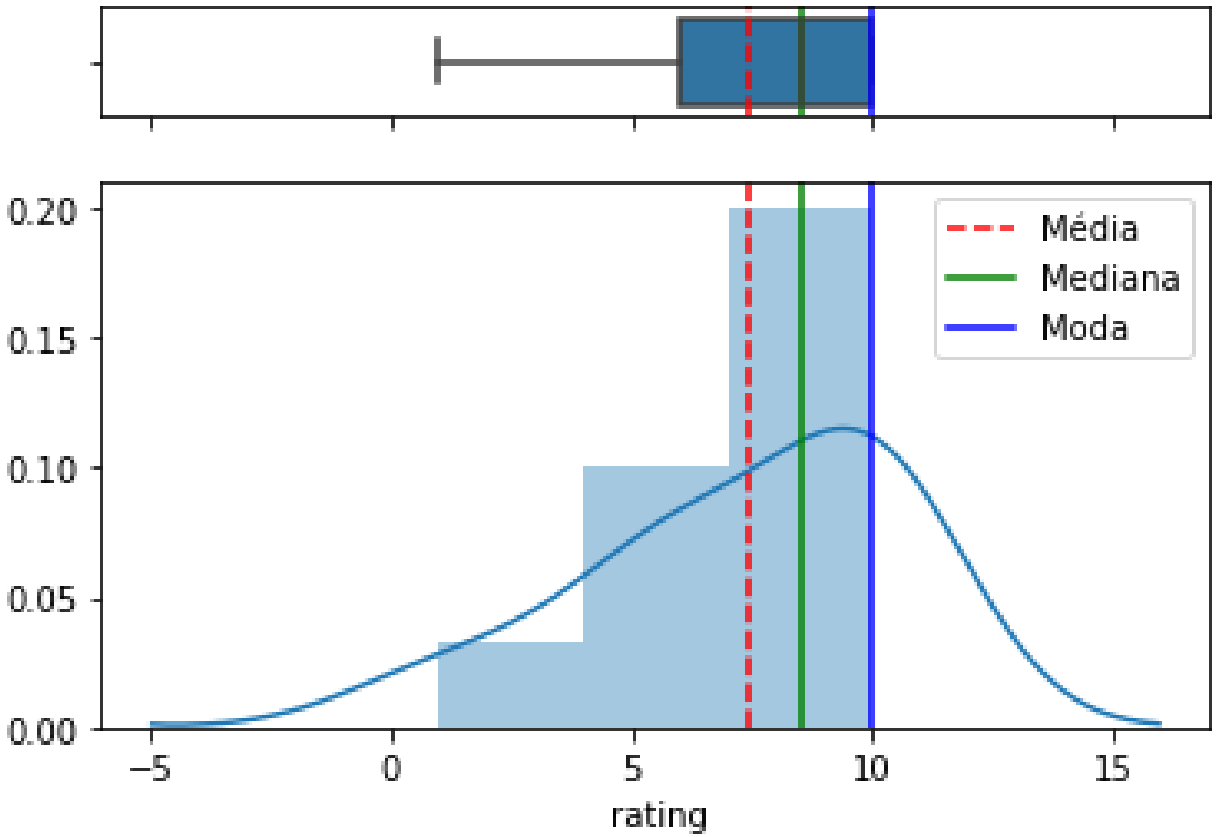
Exemplo de distribuição simétrica.

Medidas de assimetria



Exemplo de distribuição assimétrica à direita.

Medidas de assimetria



Exemplo de distribuição assimétrica à esquerda.

Medidas de curtose

- **Distribuições mesocúrticas:** achatamento da distribuição normal
- **Distribuições leptocúrticas:** distribuição mais concentrada
- **Distribuições platicúrticas:** distribuição mais achatada

Leitura complementar:

- `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kurtosis.html`
- `https://pt.wikipedia.org/wiki/Curtose`

Medidas de curtose

Medida que caracteriza o achatamento da curva.

- Curtose ≈ 0: achatamento da curva normal
- Curtose > 0: leptocúrtica, distribuição mais afunilada
- Curtose < 0: platicúrtica, distribuição mais achatada

Obs: Distribuição normal

`https://www.spss-tutorials.com/normal-distribution/`

Medidas de curtose