SME0803 Visualização e Exploração de Dados

Medidas descritivas de dados quantitativos - Parte 2

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 Análise Exploratória de Dados. Mário de Castro, ICMC-USP, 2010.

Momentos amostrais

A **média**, a **variância**, a **assimetria** e a **curtose** são parâmetros que ajudam a caracterizar a distribuição de uma variável aleatória.

Para uma amostra observada, procuramos medidas para estimar esses parâmetros, algumas delas com base em **momentos amostrais**.

Momentos amostrais

O **momento de ordem** k (k inteiro e positivo) em relação à origem é dado por

$$m'_k = \frac{\sum_{i=1}^{n} x_i^k}{n}$$

Momentos amostrais

O **momento de ordem** k (k inteiro e positivo) em relação à origem é dado por

$$m'_k = \frac{\sum_{i=1}^{n} x_i^k}{n}$$

- A **média amostral** é o primeiro momento em relação à origem, com $k = 1$

$$m'_1 = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Momentos amostrais centrais

O **momento de ordem** k (k inteiro e positivo) em relação à média é dado por

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

1. A **variância amostral** é o segundo momento em relação a média, com $k = 2$

$$m_2 = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}.$$

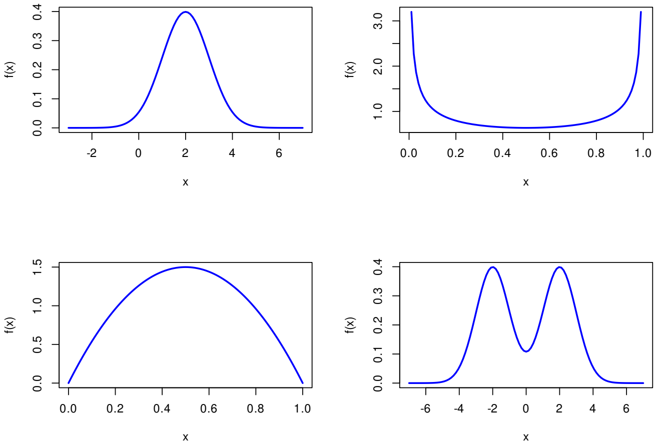
Medidas de assimetria

Distribuição simétrica

Uma variável aleatória X tem distribuição simétrica em relação a um valor x_0 se $f(x_0 - a) = f(x_0 + a)$ para todo a.

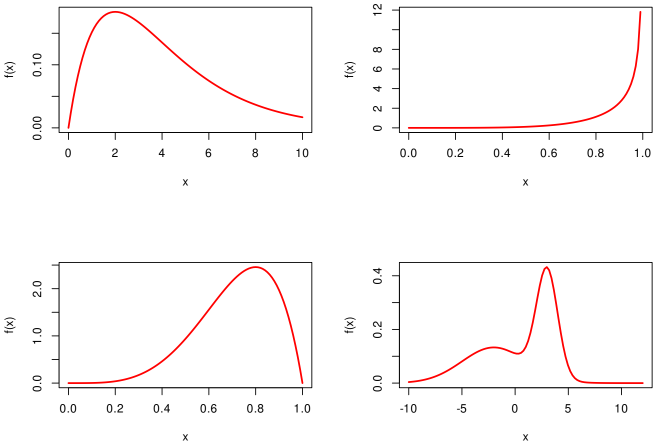
Observação: $f(x)$ é a função densidade de probabilidade de X, que será vista formalmente mais adiante. Por ora, veja `https://pt.wikipedia.org/wiki/Fun%C3%A7%C3%A3o_densidade`.

Distribuições simétricas: exemplos



Fonte: Elaborado pela autora, adaptado das notas de aula de M. de Castro.

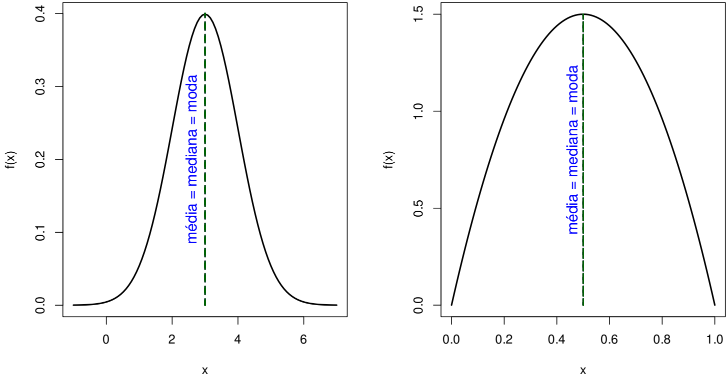
Distribuições assimétricas: exemplos



Fonte: Elaborado pela autora, adaptado das notas de aula de M. de Castro.

Relação entre moda, média e mediana

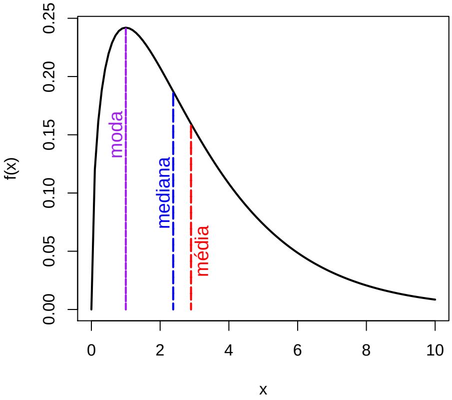
Suponha que a distribuição é unimodal e que a média existe. Se a distribuição é simétrica, temos média = mediana = moda.



Fonte: Elaborado pela autora, adaptado das notas de aula de M. de Castro.

Assimetria à direita

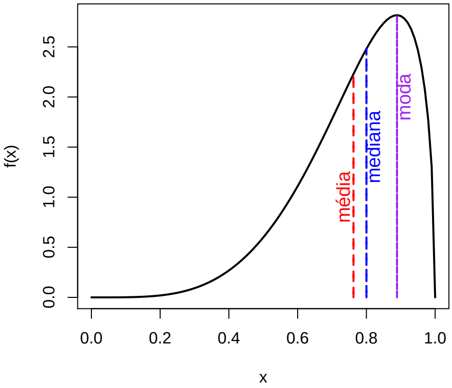
Para distribuições com **assimetria à direita** (assimetria positiva), temos **moda < mediana < média.**



Fonte: Elaborado pela autora, adaptado das notas de aula de M. de Castro.

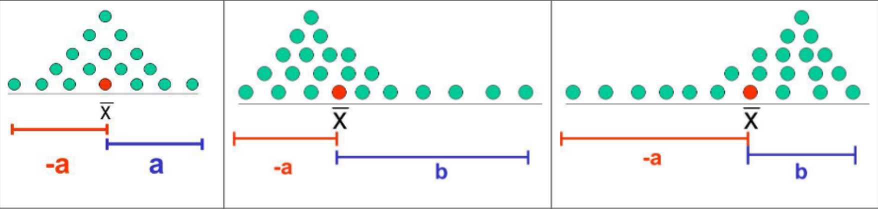
Assimetria à esquerda

Para distribuições com **assimetria à esquerda** (assimetria negativa), temos **média < mediana < moda**



Fonte: Elaborado pela autora, adaptado das notas de aula de M. de Castro.

Assimetria em conjuntos de dados



Simetria	Assimetria Positiva	Assimetria Negativa
Numa distribuição perfeitamente simétrica, se existir um ponto a uma distância a acima da média existirá um outro ponto, localizado à mesma distância abaixo da média.	Numa distribuição assimétrica positiva, a tendência é que hajam desvios positivos muito maiores do que os negativos	Numa distribuição assimétrica negativa, a tendência é que hajam desvios negativos muito maiores do que os positivos

Figura 5: Histogramas estilizados de distribuições com diferentes tipos de assimetria

Fonte: Coeficiente de Assimetria, por Rinaldo Artes (Insper). Disponível em `https://www.ime.usp.br/~mbranco/MedidasdeAssimetria_2014.pdf` Acesso em 04/05/2021.

Assimetria em conjuntos de dados

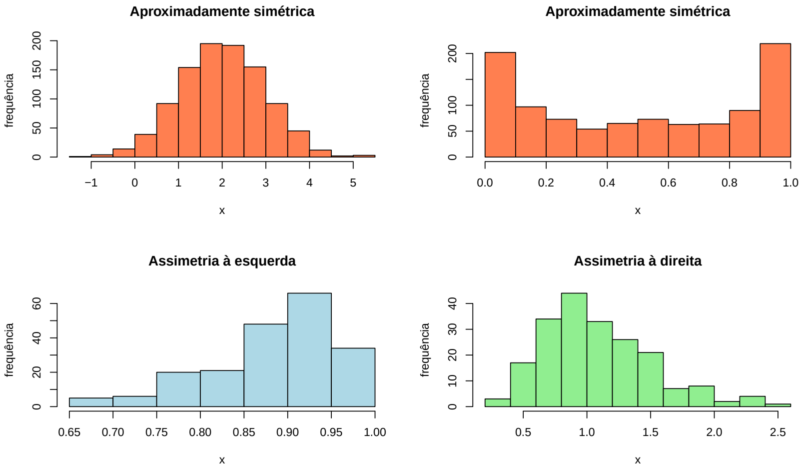
Amostras de uma variável aleatória X simétrica não necessariamente são perfeitamente simétricos.

Um conjunto de dados $x_1, \ldots, x_n$ é perfeitamente simétrico em relação a x_0 se, para todo x_j, existe x_k tal que $(x_j - x_0) = -(x_k - x_0)$, $j, k = 1, \ldots, n$ e $j \neq k$.

Se essa relação é observada, então $x_0 = \bar{x}$.

Como medir a assimetria? Precisamos de uma medida que seja próxima de 0 caso os dados sejam aproximadamente simétricos.

Assimetria em conjuntos de dados



Fonte: Elaborado pela autora, adaptado das notas de aula de M. de Castro.

Medida de Assimetria de Pearson

A **medida de assimetria de Pearson** é dada por

$$g = \frac{\bar{x} - Mo}{s},$$

com $Mo =$ moda (requer o cálculo da moda).

Propriedade: g é adimensional.

Medida de Assimetria de Pearson

A **medida de assimetria de Pearson** é dada por

$$g = \frac{\bar{x} - Mo}{s},$$

com $Mo =$ moda (requer o cálculo da moda).

Propriedade: g é adimensional.

Como o cálculo da moda nem sempre é possível, uma possível proposta seria

$$g^{\star} = \frac{\bar{x} - m_d}{s}, \text{ com } m_d = \text{mediana}.$$

Assimetria de uma distribuição

Propriedade:

Se uma variável aleatória X com média (populacional) μ tem distribuição simétrica, seus momentos centrais de ordem k ímpar, se existirem, são todos nulos:

$$E[(X-\mu)]^k = 0, \text{ sendo que } \mu = E(X).$$

Assimetria de uma distribuição

Propriedade:
Se uma variável aleatória X com média (populacional) μ tem distribuição simétrica, seus momentos centrais de ordem k ímpar, se existirem, são todos nulos:

$$E[(X-\mu)]^k = 0, \text{ sendo que } \mu = E(X).$$

Sendo assim, podemos utilizar o 3^o momento central para propor uma medida de assimetria.

Obs: $E[(X-\mu)]^3 = 0$ não implica que a distribuição de X é simétrica.

Medida de Assimetria de Fisher

A **medida de assimetria de Fisher** (baseado em momentos centrais amostrais) é dada por

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$$

Medida de Assimetria de Fisher

A **medida de assimetria de Fisher** (baseado em momentos centrais amostrais) é dada por

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$$

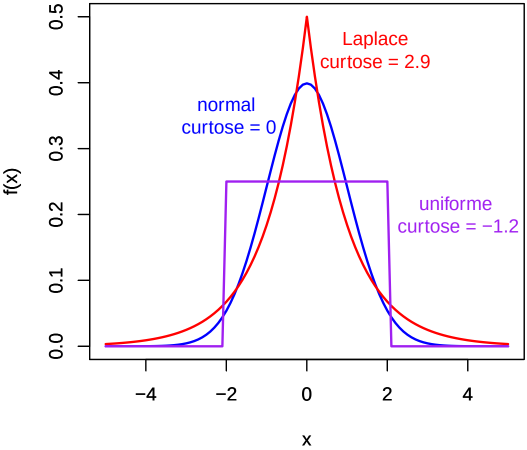
Propriedades:

1. g_1 é adimensional.
2. g_1 é um número real.

Medida de Assimetria de Fisher

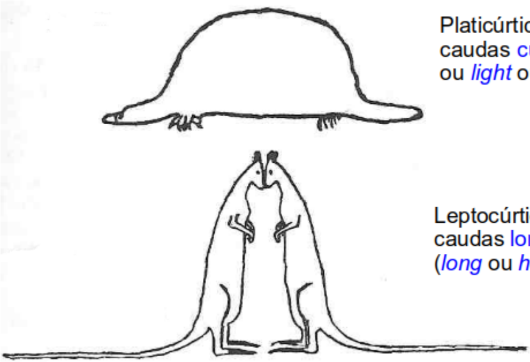
- $g_1 < 0$ indica distribuição assimétrica à esquerda.
- $g_1 = 0$ indica distribuição simétrica
- $g_1 > 0$ indica distribuição assimétrica à direita.

Curtose: Achatamento da distribuição



Fonte: Elaborado pela autora, adaptado das notas de aula de M. de Castro.

Curtose: Achatamento da distribuição



Platicúrtica (*platykurtic*): caudas curtas ou leves (*short* ou *light* ou *thin*).

Leptocúrtica (*leptokurtic*): caudas longas ou pesadas (*long* ou *heavy* ou *thick* ou *fat*).

Fonte. Bulmer, M. G. (1979), *Principles of Statistics*, Dover: New York.

Mesocúrtica (*mesokurtic*): caudas neutras (nem curtas e nem longas).

Fonte: Notas de aula de M. de Castro.

Medida de curtose

A **medida de curtose** (baseada em momentos centrais amostrais) é dada por

$$g_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

Medida de curtose

A **medida de curtose** (baseada em momentos centrais amostrais) é dada por

$$g_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

Propriedades:

1. g_2 é adimensional.
2. g_2 é um número real.
3. Para $X \sim N(\mu, \sigma^2)$, $\dfrac{E[(X-\mu)^4]}{E[(X-\mu)^2]^2} = \dfrac{E[(X-\mu)^4]}{\sigma^4} = 3.$

Medida de curtose

- Se $g_2 < 3$, a distribuição é platicúrtica.
- Se $g_2 = 3$, a distribuição é mesocúrtica.
- Se $g_2 > 3$, a distribuição é leptocúrtica.

Medida de curtose

É comum definir a curtose como

$$g_{2e} = g_2 - 3 \text{ (excess).}$$

Medida de curtose

É comum definir a curtose como

$$g_{2e} = g_2 - 3 \text{ (excess)}.$$

- Se $g_{2e} < 0$, a distribuição é platicúrtica.
- Se $g_{2e} = 0$, a distribuição é mesocúrtica.
- Se $g_{2e} > 0$, a distribuição é leptocúrtica.