

Estatística e Probabilidade

Medidas de Assimetria e Curtose





Desenvolvimento do material

Gregório Dalle Vedove Nosaki

1^a Edição

Copyright © 2021, Unigranrio

Nenhuma parte deste material poderá ser reproduzida, transmitida e gravada, por qualquer meio eletrônico, mecânico, por fotocópia e outros, sem a prévia autorização, por escrito, da Unigranrio.

Sumário

Medidas de Assimetria e Curtose

Para início de conversa...	3
Objetivo	3
1. Assimetria	4
2. Curtose	11
Referências	14

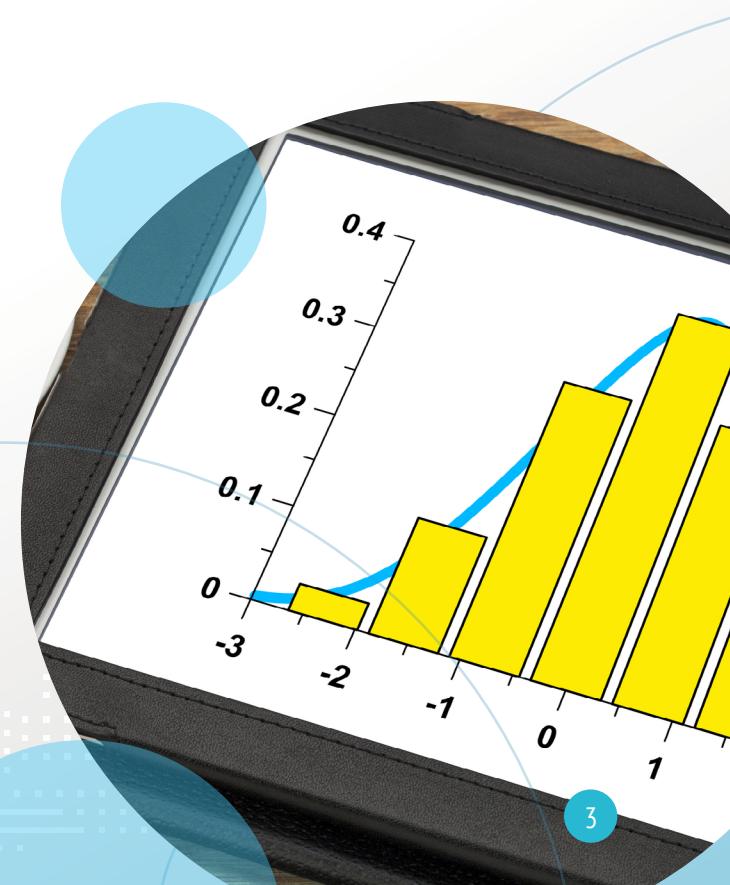


Para início de conversa...

Ao analisar uma distribuição de frequências de uma determinada variável, podemos compreender um pouco mais o seu comportamento por meio das medidas de posição e medidas de dispersão apresentadas anteriormente. Neste capítulo, iremos trabalhar com outros dois conceitos importantes com relação às distribuições de frequências e os coeficientes que podem nos auxiliar a realizar esse estudo. O primeiro conceito que iremos abordar é a assimetria que uma distribuição pode ter quando o valor da moda se distancia do valor da média aritmética. Esse comportamento recebe esse nome porque, ao traçar uma curva da distribuição de frequência centralizada na média, a curva tem seu ponto mais alto desviado ora para direita e ora para esquerda. Avaliaremos esses dois casos e como calcular alguns coeficientes de assimetria. Outro elemento importante na análise da distribuição de uma frequência é a curtose, que aborda o grau de achatamento de uma distribuição. Também apresentaremos alguns coeficientes para caracterizar a curtose e como avaliá-los para obter informações sobre a amplitude dos valores da nossa variável.

Objetivo

Calcular e interpretar as medidas de assimetria e curtose.



1. Assimetria

Uma das distribuições de frequência mais comuns e recorrentes para diversas variáveis e em diferentes exemplos é a **distribuição normal**, também chamada de **distribuição gaussiana**, que tem um formato de sino como apresentada no gráfico a seguir.

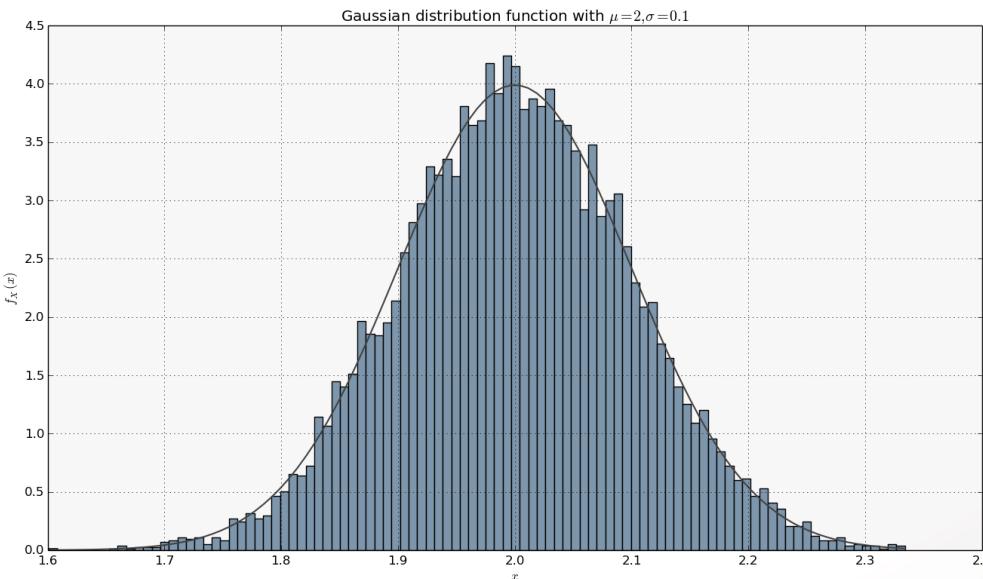


Figura 1: Distribuição normal com média 2 e desvio padrão de 0,1. Fonte: Wikimedia.

Esse tipo de distribuição tem a característica de ser simétrica com relação à média aritmética, mas nem todas as distribuições possuem esse aspecto. Iremos tratar justamente de parâmetros que medem esse tipo de comportamento neste capítulo. Vamos começar nosso estudo com o conceito de momento definido a seguir.

Definição: seja X uma coleção de observáveis x_1, x_2, \dots, x_n com n entradas e $r \in \mathbb{Z}$ um inteiro positivo. O **momento de ordem r** é definido como sendo

$$m_X(r) = \frac{x_1^r + x_2^r + \dots + x_n^r}{n} = \frac{\sum_{i=1}^n x_i^r}{n}.$$

Além disso, seja $a \in \mathbb{R}$, definimos o **momento de ordem r centrado na origem a** como sendo

$$m_X(r, a) = \frac{(x_1 - a)^r + (x_2 - a)^r + \dots + (x_n - a)^r}{n} = \frac{\sum_{i=1}^n (x_i - a)^r}{n}.$$

Antes de apresentarmos alguns exemplos, cabe algumas observações sobre as definições acima: se o momento de ordem 1 é igual à média aritmética, ou seja, $m_X(1) = \bar{X}$ onde \bar{X} é a média aritmética da coleção X ; se tomarmos $a = \bar{X}$, então o momento de ordem 1 centrado em \bar{X} é igual a zero para qualquer que seja a coleção X , ou seja, $m_X(1, \bar{X}) = 0$; ainda tomando $a = \bar{X}$, vale que $m_X(2, \bar{X}) = \text{var}(X)$ onde $\text{var}(X)$ é variação de X .

Considere a seguinte coleção de dados:

$$X = 2 \ 3 \ 5 \ 6 \ 9.$$

O momento de ordem 1 da coleção X é tal que

$$m_X(1) = \frac{2 + 3 + 5 + 6 + 9}{5} = \frac{25}{5} = 5$$

O cálculo anterior é exatamente a definição de média aritmética e, portanto, $\bar{X} = 5$. Calculando o momento de ordem 2, temos que

$$m_X(2) = \frac{2^2 + 3^2 + 5^2 + 6^2 + 9^2}{5} = \frac{155}{5} = 31$$

Tomando $a = \bar{X}$, podemos calcular os momentos centrados na origem \bar{X} . Note que

$$m_X(1, \bar{X}) = \frac{(2 - 5) + (3 - 5) + (5 - 5) + (6 - 5) + (9 - 5)}{5} = \frac{0}{5} = 0$$

e

$$m_X(1, \bar{X}) = \frac{(2 - 5)^2 + (3 - 5)^2 + (5 - 5)^2 + (6 - 5)^2 + (9 - 5)^2}{5} = \frac{30}{5} = 6$$

O cálculo anterior é exatamente o cálculo da variação da coleção X e, portanto, $\text{var}(X) = 6$.

Os momentos também podem ser calculados para dados agrupados. Apresentamos a definição formal do cálculo de momentos para este caso a seguir.

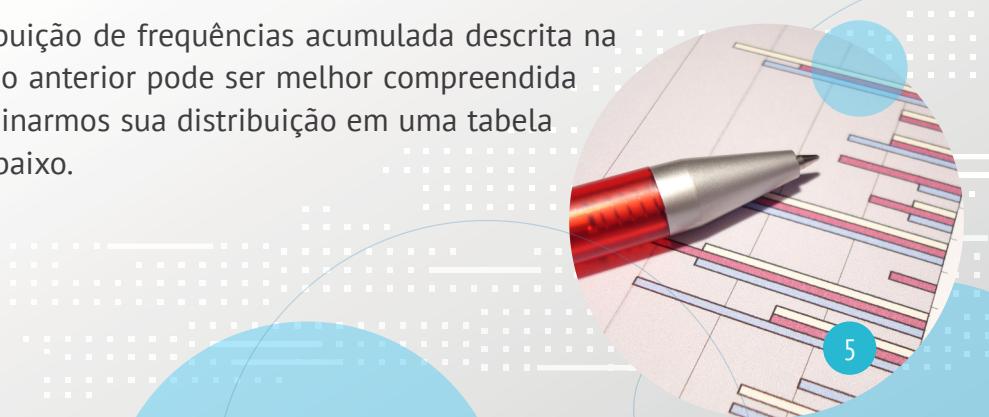
Definição: considere uma distribuição de frequência com dados agrupados em n classes tais que os intervalos de classe tem pontos médios x_1, x_2, \dots, x_n e a frequência de cada classe é f_1, f_2, \dots, f_n , respectivamente. O **momento de ordem r** , onde $r \in \mathbb{Z}$ é um inteiro positivo é definido como sendo

$$m_X(r) = \frac{f_1 \cdot x_1^r + f_2 \cdot x_2^r + \dots + f_n \cdot x_n^r}{n} = \frac{\sum_{i=1}^n f_i \cdot x_i^r}{n}$$

Além disso, seja $a \in \mathbb{R}$, definimos o **momento de ordem r centrado na origem a** como sendo

$$m_X(r, a) = \frac{f_1(x_1 - a)^r + f_2(x_2 - a)^r + \dots + f_n(x_n - a)^r}{n} = \frac{\sum_{i=1}^n f_i(x_i - a)^r}{n}$$

A distribuição de frequências acumulada descrita na definição anterior pode ser melhor compreendida se imaginarmos sua distribuição em uma tabela como abaixo.



Intervalo de classe	Frequência	Frequência acumulada
$\ell_0 \cup \ell_1$	f_1	f_1
$\ell_1 \cup \ell_2$	f_2	$f_1 + f_2$
\vdots	\vdots	\vdots
$\ell_{i-1} \cup \ell_i$	f_i	$f_1 + f_2 + \dots + f_i$
\vdots	\vdots	\vdots
$\ell_{n-1} \cup \ell_n$	f_n	$f_1 + \dots + f_n$

Na tabela, cada linha refere-se a um intervalo de classe cujos extremos são apresentados na primeira coluna. Na segunda coluna, temos a frequência de cada uma das classes e na terceira a coluna a frequência acumulada.

Como dito anteriormente, uma assimetria é um desvio da distribuição das frequências em relação à média das observáveis. Existem diferentes coeficientes que são utilizados para indicar o grau de assimetria de uma distribuição. Podemos dividir as assimetrias em dois tipos: assimetrias positivas e assimetria negativas. Observe a representação abaixo desses dois tipos de assimetria.

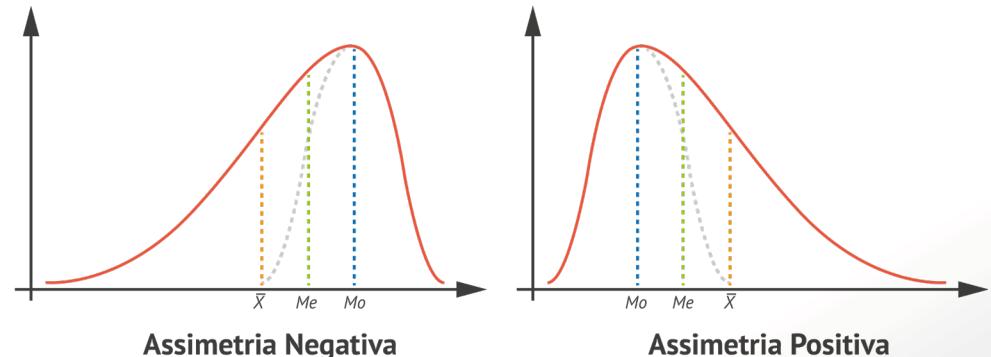


Figura 2: As assimetrias podem ser divididas em positivas e negativas. Fonte: Wikimedia.

A assimetria positiva recebe esse nome porque a diferença entre a média e a moda da distribuição é um valor positivo, enquanto que a assimetria negativa recebe esse nome porque a diferença entre a média e a moda da distribuição é negativa. Perceba que o ponto mais alto da curva que descreve a distribuição nas representações acima é o valor da moda, pois é o valor cuja frequência é a maior entre todas as outras.

Apresentaremos alguns desses coeficientes que nos ajudam a caracterizar e descrever os comportamentos assimétricos de distribuições. Iremos considerar uma coleção X com n entradas, sendo x_1, x_2, \dots, x_n . Além disso, podemos considerar que a coleção X está ordenada em ordem crescente tal que

$$x_1 \leq x_2 \leq \dots \leq x_n$$

sem perda de generalidade. Denotaremos por \bar{X} a média aritmética da coleção X e por X_{moda} e $X_{mediana}$ a moda e mediana, respectivamente. Além disso, $var(X)$ denota a variação e $dp(X)$ o desvio padrão.

Definição: o **primeiro coeficiente de assimetria de Pearson** será denotado como a_P e é definido como sendo

$$a_P = \frac{\bar{X} - X_{moda}}{dp(X)}.$$

O **segundo coeficiente de assimetria de Pearson** será denotado como a'_P e é definido como sendo

$$a'_P = \frac{3 \cdot (\bar{X} - X_{mediana})}{dp(X)}.$$



Curiosidade

Quando uma distribuição é moderadamente assimétrica, é possível mostrar a seguinte relação entre as medidas de tendência central

$$\bar{X} - X_{moda} = 3 \cdot (\bar{X} - X_{mediana})$$

e, portanto, o segundo coeficiente de assimetria de Pearson apresentado na definição anterior é obtido a partir do primeiro. (MATTOS; AZAMBUJA; KONRATH, 2017).

Existem diferentes interpretações para os coeficientes de Pearson. Segundo Mattos, Azambuja e Konrath (2017), o segundo coeficiente de Pearson apresentado aqui pode classificar as assimetrias segundo os seguintes critérios:

- se $|a'_P| \leq 0,15$, então a distribuição é praticamente simétrica;
- se $0,15 \leq |a'_P| \leq 1$, então a distribuição é moderadamente assimétrica;
- se $|a'_P| > 1$, então a distribuição é acentuadamente assimétrica.

Cabe destacar aqui que o segundo coeficiente de assimetria de Pearson varia entre -3 e 3 e que para determinar se a assimetria é negativa ou positiva, podemos avaliar o sinal da diferença entre a média e a moda. Vejamos agora o cálculo desses coeficientes em um exemplo numérico.

Exemplo 1: considere a seguinte tabela, que apresenta as notas de uma turma de 120 alunos em uma prova.

Nota	Frequência	Frequência acumulada
6	10	10
7	50	60
8	30	90
9	20	110
10	10	120

Calculando as medidas de posição e de dispersão apresentadas nos capítulos anteriores para essa coleção de dados, obtemos os seguintes valores:

Média aritmética:

$$\bar{X} = \frac{6 \cdot 10 + 7 \cdot 50 + 8 \cdot 30 + 9 \cdot 20 + 10 \cdot 10}{120} = \frac{930}{120} = 7,75 ;$$

Moda: trata-se do maior com maior frequência, ou seja,

$$X_{\text{moda}} = 7 ;$$

Mediana: como temos 120 observáveis e elas estão organizadas em ordem crescente, tomamos a média entre o 60º e 61º valor para determinar a mediana, ou seja,

$$X_{\text{mediana}} = \frac{8 + 7}{2} = 7,5 ;$$

Variação:

$$\begin{aligned} \text{var}(X) &= \frac{(6 - 7,75)^2 \cdot 10 + (7 - 7,75)^2 \cdot 50 + (8 - 7,75)^2 \cdot 30 + (9 - 7,75)^2 \cdot 20 + (10 - 7,75)^2 \cdot 10}{120} \\ &= \frac{142,5}{120} = 1,1875 ; \end{aligned}$$

Desvio padrão:

$$dp(X) = \sqrt{1,1875} \approx 1,0897$$

A partir dessas informações, somos capazes de calcular os coeficientes de assimetria de Pearson. Utilizando as fórmulas apresentadas anteriormente e as medidas de posição e de dispersão calculadas, obtemos que

$$a_P = \frac{7,75 - 7}{1,0897} \approx 0,6882$$

e

$$a'_P = \frac{3 \cdot (7,75 - 7,5)}{1,0897} \approx 0,6882$$

Ambos os coeficientes de assimetria de Pearson têm o mesmo valor para os dados apresentados na tabela. Além disso, podemos considerar que essa é uma distribuição moderadamente assimétrica e que é do tipo positiva, isso é, o valor da moda é inferior ao valor da média aritmética.

Existem outros coeficientes de assimetria que podem ser usados para determinar o grau do deslocamento da distribuição de uma determinada variável. Apresentaremos outros dois coeficientes que se baseiam nas

separatrizes de uma coleção de dados. Considerando ainda as notações apresentadas anteriormente, iremos considerar os quartis que dividem a coleção X como sendo Q_1 , Q_2 e Q_3 . Como $X_{mediana} = Q_2$, usaremos a notação $X_m = X_{mediana}$ da mediana para o segundo quartil. Além disso, para a mesma coleção de dados, denotaremos os percentis, isso é, as separatrizes que dividem as observáveis em 100 subconjuntos de mesmo tamanho seguindo uma ordem crescente, pela notação P_i onde $1 \leq i \leq 99$ indicando cada um dos percentis.

Definição: o **coeficiente de assimetria de Yule**, denotado por a_Y , para a coleção X e calculado baseado nas diferenças $Q_3 - X_m$ e $X_m - Q_1$. Ele é definido como sendo

$$a_Y = \frac{(Q_3 - X_m) - (X_m - Q_1)}{(Q_3 - X_m) + (X_m - Q_1)} = \frac{Q_3 + Q_1 - 2X_m}{Q_3 - Q_1}.$$

Seguindo a mesma lógica aplicada no coeficiente de assimetria de Yule, definimos o coeficiente de assimetria de Kelley que se baseia na divisão da amostra em percentis.

Definição: o **coeficiente de assimetria de Kelley**, denotado por a_K , para a coleção X como sendo

$$a_K = \frac{P_{90} + P_{10} - 2X_m}{P_{90} - P_{10}}.$$

! Importante

Lembre-se de que, pela definição de percentis, o percentil P_{10} marca o valor tal que 10% dos valores da coleção X são menores que P_{10} , enquanto que P_{90} marca o valor tal que 10% dos valores da coleção X são maiores que P_{90} .

Exemplo 2: Considere a coleção de observáveis apresentada a seguir referente a contagem de carros que estavam estacionados em uma rua no centro de uma cidade em diferentes dias da semana: (MATTOS; AZAMBUJA; KONRATH, 2017. Adaptado).

4 4 5 7 7 7 8 8 9 9 9 10 10 11



Figura 3: Coleção de observáveis. Fonte: Dreamstime.

Os dados já foram ordenados em ordem crescente e não seguem mais a ordem da observação no decorrer do experimento. Denominando a coleção apresentada acima como X , vamos determinar os quartis dela para que possamos calcular o coeficiente de assimetria de Yule. Com os em ordem crescente, determinamos a mediana, encontrando o valor que está exatamente no meio das observáveis acima. Como temos 14 observáveis, a mediana é igual a média aritmética entre a 7^a e 8^a observáveis:

$$X_m = \frac{8 + 8}{2} = 8$$

Após determinarmos a mediana da coleção X , podemos determinar a mediana dos dois subconjuntos gerados por essa primeira divisão para encontrarmos os dois quartis restantes. Para o subconjunto

4 4 5 7 7 7 8

concluímos que o primeiro quartil é $Q_1 = 7$; para o subconjunto

8 9 9 9 10 10 11

concluímos que o terceiro quartil é $Q_3 = 9$.

Dessa forma, o coeficiente de assimetria de Yule é

$$ay = \frac{7 + 9 - 2 \cdot 8}{9 - 7} = 0$$

Observe o histograma da coleção X apresentado abaixo.

Distribuição de frequências

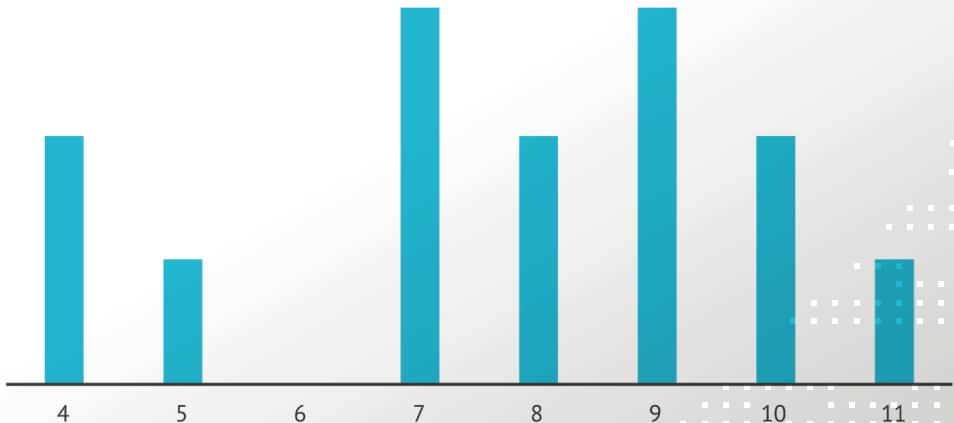


Figura 3: Histograma da coleção X. Fonte: Elaborada pelo autor.

2. Curtose

Seguindo na investigação sobre o comportamento de uma distribuição de frequência de uma variável, a curtose nos dá informações sobre o grau de achatamento da distribuição. Observe as seguintes representações de distribuição de frequências em gráficos de barras.

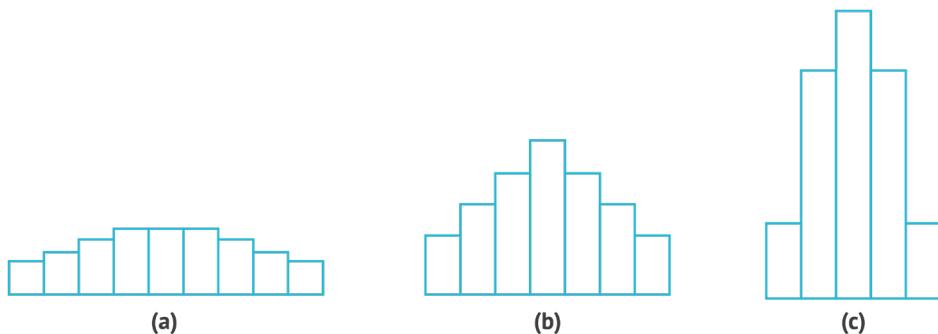


Figura 4: Representações de distribuição de frequências em gráficos de barras.
Fonte: Adaptado de Spiegel e Stephens (2009).

Com base no achatamento de uma distribuição, podemos classificá-la em três tipos:

- Platicúrtica (gráfico a): caracteriza-se por não apresentar grandes discrepâncias entre as frequências dos diferentes valores que a variável assume; tem uma distribuição mais achatada de todas.

- Mesocúrtica (gráfico b): trata-se de um grau de achatamento moderado entre os outros dois tipos de distribuições; tem um pico de frequência, mas não é muito pronunciado.
- Leptocúrtica (gráfico c): caracteriza-se por apresentar grandes discrepâncias entre as frequências dos valores que a variável assume; trata-se da distribuição menos achatada de todas.

Assim como na seção anterior, vamos definir alguns coeficientes que podem auxiliar na interpretação dos dados de uma distribuição de frequência quanto ao seu achatamento. O primeiro deles é baseado na definição de momento apresentada no início deste capítulo e o segundo baseado em separatrizes da coleção.

Definição: o coeficiente de curtose calculado a partir de momentos da coleção X será denotado por $b(X)$ e é definido como sendo

$$b(X) = \frac{m_X(4, \bar{X})}{(m_X(2, \bar{X}))^2}$$

Seguindo a interpretação proposta em Mattos, Azambuja e Konrath (2017), o coeficiente de curtose $b(X)$ pode nos auxiliar a classificar a distribuição de frequências sendo

- leptocúrtica, se $b(X) < 3$;

- mesocúrtica, se $b(X) = 3$; e
- platicúrtica, se $b(X) > 3$.

Definição: o coeficiente de curtose calculado a partir de separatrizes da coleção X será denotado por $c(X)$ e é definido como sendo

$$c(X) = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}.$$

Ainda em Mattos, Azambuja e Konrath (2017), o coeficiente de curtose baseado nas separatrizes da coleção denotado aqui por $c(X)$, classifica a distribuição de frequências sendo

- leptocúrtica, se $c(X) < 0,263$;
- mesocúrtica, se $c(X) = 0,263$; e
- platicúrtica, se $c(X) > 0,263$.

Exemplo 3: utilizando os dados apresentados no Exemplo 2, iremos calcular o coeficiente de curtose a partir dos momentos. O primeiro passo para calcular os momentos envolvidos no cálculo desse coeficiente é determinar o valor da média aritmética. Obtemos que

$$\bar{X} = \frac{4 + 4 + 5 + 7 + 7 + 7 + 8 + 8 + 9 + 9 + 9 + 10 + 10 + 11}{14} = \frac{108}{14} = 7,714.$$

Para facilitar o cálculo dos momentos $m_X(4, \bar{X})$ e $m_X(2, \bar{X})$, organizamos alguns dados em uma tabela apresentada a seguir.

Valor da variável (x)	Frequência	$x - \bar{X}$	$(x - \bar{X})^2$	$(x - \bar{X})^4$
4	2	-3,714	13,794	190,27
5	1	-2,714	7,365	54,255
7	3	-0,714	0,51	0,26
8	2	0,286	0,082	0,007
9	3	1,286	1,653	2,735
10	2	2,286	5,225	27,31
11	1	3,286	10,797	116,592

A partir dos dados da tabela, obtemos que

$$\begin{aligned}
 m_X(4, \bar{X}) &= \frac{2 \cdot 190,27 + 1 \cdot 54,255 + 3 \cdot 0,26 + 2 \cdot 0,007 + 3 \cdot 2,735 + 2 \cdot 27,31 + 1 \cdot 116,592}{14} \\
 &= \frac{380,54 + 54,255 + 0,78 + 0,014 + 8,205 + 54,62 + 116,592}{14} \\
 &= \frac{615,006}{14} = 43,929
 \end{aligned}$$

e

$$\begin{aligned}m_X(2, \bar{X}) &= \frac{2 \cdot 13,794 + 1 \cdot 7,365 + 3 \cdot 0,51 + 2 \cdot 0,082 + 3 \cdot 1,635 + 2 \cdot 5,225 + 1 \cdot 10,797}{14} \\&= \frac{27,588 + 7,365 + 1,53 + 0,164 + 4,905 + 10,45 + 10,797}{14} \\&= \frac{62,799}{14} = 4,485\end{aligned}$$

Portanto

$$b(X) = \frac{m_X(4, \bar{X})}{(m_X(2, \bar{X}))^2} = \frac{43,929}{(4,485)^2} = 2,1838$$

Como $b(X) < 3$, a distribuição de frequência da variável X é leptocúrtica

! Importante

O cálculo dos coeficientes de assimetria em coleções de dados pequenos podem conduzir a interpretações que não correspondem exatamente à realidade dos dados. O Exemplo 2 apresentado anteriormente é um caso desses, pois estamos trabalhando apenas com 14 observáveis e uma variação muito pequena dos valores que a variável assume. Em Spiegel e Stephens (2009), você poderá encontrar mais exemplos do cálculo dos coeficientes apresentados aqui utilizando softwares matemáticos.

Gráfico de dispersão de Alturas, Idade de casamento, Idade de óbito e Quantidade de refrigerante

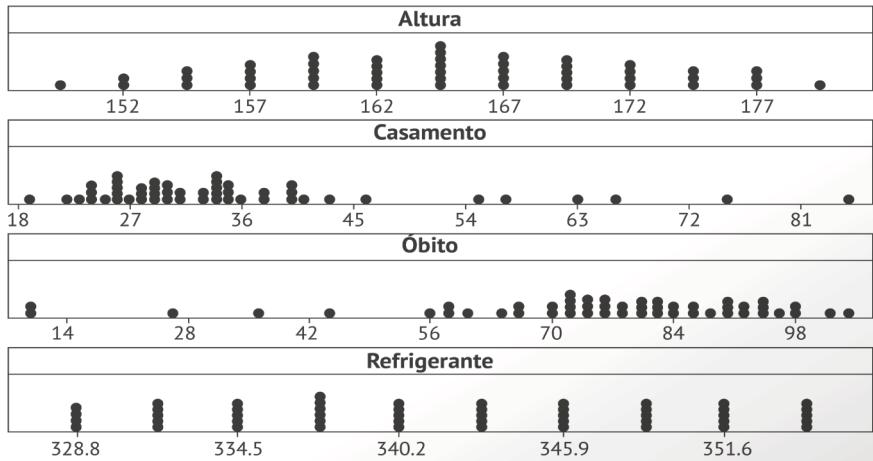


Figura 5: Gráfico Minitab para as quatro distribuições: normal, assimétrica à direita, assimétrica à esquerda e uniforme. Fonte: Adaptado de Spiegel e Stephens (2009).

Nesta unidade, trabalhamos inicialmente com o conceito de momento dentro de estatística e como ele pode ser calculado. Vimos também a relação dos momentos com a média aritmética e a variância de um determinado conjunto de observáveis. Tratamos dos tipos de assimetria que podem ser observados em uma distribuição de frequências e como classificá-los a partir das medidas de posição. Estudamos alguns dos principais coeficientes de assimetria que podem ser calculados

para determinar o grau de assimetria de uma distribuição. Por fim, trabalhamos com o conceito de curtose, que analisa o grau de achatamento da curva de distribuição de frequências e também está relacionada com o espectro que uma determinada variável tem dentro dos seus possíveis valores.

Referências

FONSECA, J. S.; MARTINS, G. A. **Curso de estatística**. 6. ed. São Paulo: Atlas, 2012.

MATTOS, V. L. D.; AZAMBUJA, A. M. V.; KONRATH, A.C. **Introdução à estatística**: aplicações em ciências exatas. Rio de Janeiro: LTC, 2017.

MOORE, D.; NOTZ, W.; FLIGNER, M. **A estatística básica e sua prática**. 7. ed. Rio de Janeiro: LTC, 2017.

MORETTIN, P.; BUSSAB, W. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017.

SPIEGEL, M.; SCHILLIER, J.; SRINIVASAN, A. **Probabilidade e estatística**. 3. ed. Porto Alegre: Bookman, 2013.

SPIEGEL, M.; STEPHENS, L. **Estatística**. Porto Alegre: Bookman, 2009.

TRIOLA, M. **Introdução à estatística**. 12. ed. Rio de Janeiro: LTC, 2017.

VIEIRA, S. **Fundamentos de estatística**. 6 ed. São Paulo: Atlas, 2019.