

Coefficiente de Assimetria

Rinaldo Artes
Insper

Nesta etapa do curso estudaremos medidas associadas à forma de uma distribuição de dados, em particular, os coeficientes de assimetria e curtose. Tais medidas são úteis não só para descrição dessas características, mas também para verificar se uma distribuição se aproxima de um modelo normal.

Neste texto, abordaremos os coeficientes de assimetria.

1. O Modelo Normal

O histograma apresentado na Figura 1 representa um padrão de comportamento conhecido como Modelo Normal. Trata-se de uma distribuição centrada na média (μ), unimodal e simétrica. O modelo é caracterizado¹ pela média da distribuição e pelo seu desvio-padrão, σ . A figura ilustra uma propriedade importante desse modelo: a proporção de observações num intervalo com centro em μ e largura σ é cerca de 68%; de largura 2σ é cerca de 95% e de largura 3σ , cerca de 99,7%. Isso indica que nem toda distribuição unimodal e simétrica segue esse modelo.

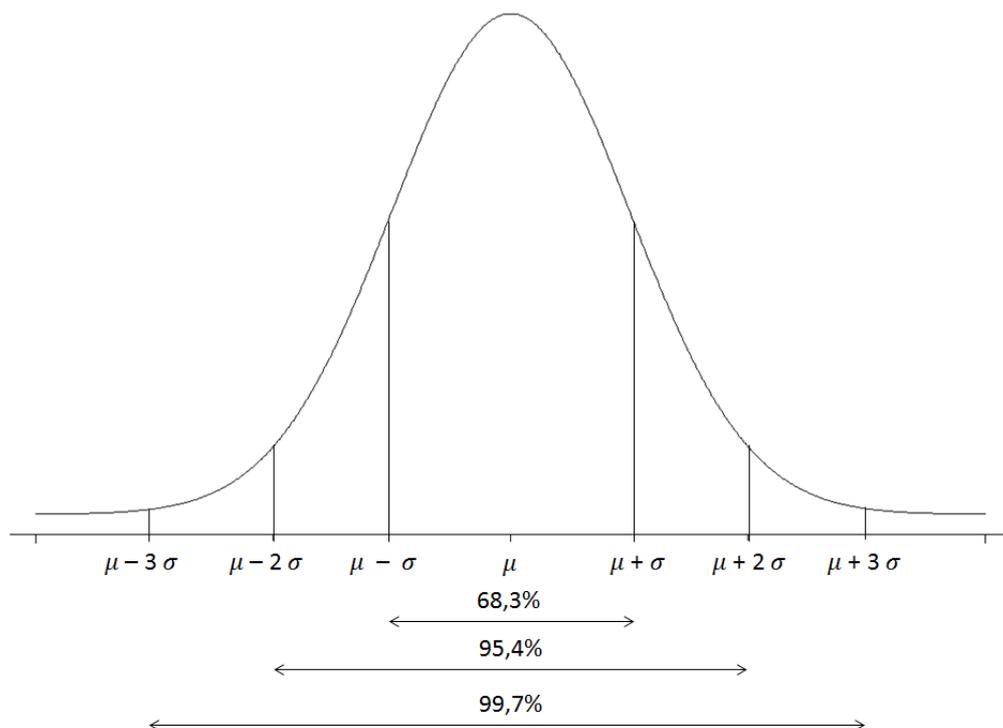


Figura 1: Modelo normal

¹ O contorno apresentado é dado por: $f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$, $-\infty < x < \infty$; $-\infty < \mu < \infty$; $\sigma > 0$. f é a função densidade de probabilidade de uma distribuição normal.

O modelo normal tem um papel central na inferência estatística. Inúmeras técnicas foram desenvolvidas partindo da suposição que os dados seguem esse modelo. Isso torna importante o desenvolvimento de critérios que permitam avaliar o quanto o comportamento de um conjunto se afasta do modelo normal.

Considere os histogramas apresentados na Figura 2. São representados os histogramas para três diferentes conjuntos de dados; em cada um foi sobreposto o contorno esperado do histograma se os dados seguissem um modelo normal. Tanto os dados do Gráfico A como os do B têm uma distribuição aproximadamente simétrica, no entanto, percebe-se que o modelo normal não se ajusta bem aos dados do Gráfico B: há muito mais observações na parte central do que se esperaria se o modelo normal fosse adequado. O Gráfico C, por sua vez, não tem um comportamento simétrico, o que torna o modelo normal inadequado. Apenas os dados do Gráfico A parecem ter um comportamento compatível com o modelo normal.

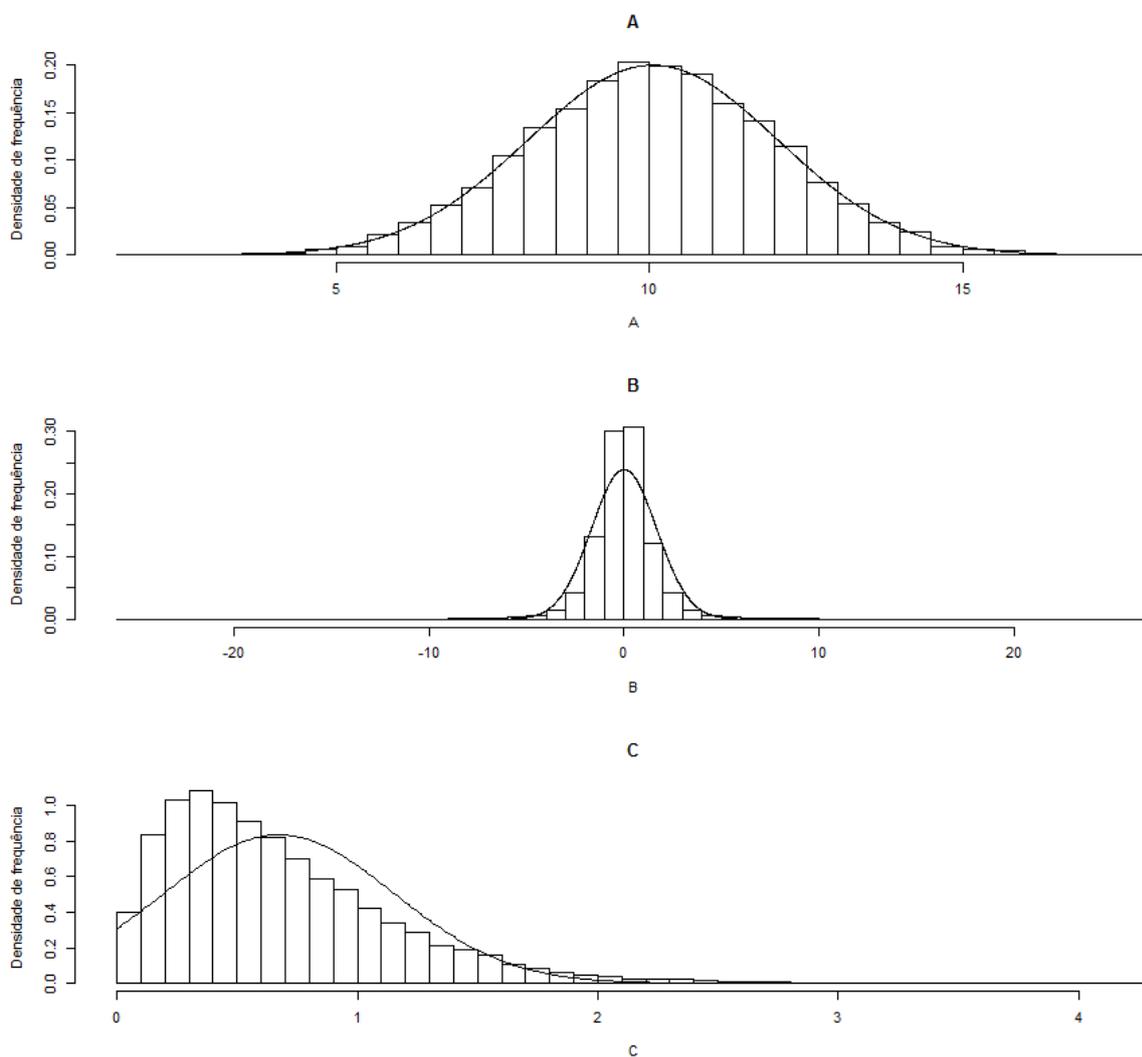


Figura 2: Histograma de três conjuntos de dados, com o ajuste do respectivo modelo normal.

Iremos desenvolver, neste texto, coeficientes que permitam avaliar o grau de assimetria e de achatamento de uma distribuição. A partir desses coeficientes será possível propor critérios para avaliar se a suposição de normalidade de um conjunto de dado é razoável.

2. Conceitos preliminares

Apresentamos os conceitos de padronização e momentos amostrais, importantes para o desenvolvimento da teoria a ser apresentada.

2.1. Padronização

Seja X_1, \dots, X_n uma amostra de uma variável com média \bar{X} e desvio-padrão S . Então a variável Z , definida como $Z_i = \frac{X_i - \bar{X}}{S}$ (a variável Z é uma variável padronizada, construída a partir de X), tem as seguintes propriedades:

- $\bar{Z} = 0$.
- $var(Z) = 1$.
- Z é uma variável adimensional.

2.2. Momentos Amostrais

Definição 1: Seja X_1, X_2, \dots, X_n um conjunto de dados, definem-se:

- momento amostral de ordem k dos dados ao valor: $M_k = \frac{\sum_{i=1}^n X_i^k}{n}$ e
- momento central amostral de ordem k dos dados ao valor: $m_k = \frac{\sum_{i=1}^n (X_i - \bar{X})^k}{n}$. Alguns autores preferem dividir o somatório por $(n - 1)$.

3. Assimetria

A Figura 3 traz histogramas e boxplots de três conjuntos de dados. Esses dados têm as seguintes características:

- Primeiro conjunto de dados – X – as observações distribuem-se de modo aproximadamente simétrico ao redor de 10. A distribuição é unimodal.
- Segundo conjunto de dados – Y – neste caso a distribuição é assimétrica. Há uma alta concentração de dados nos valores mais baixos. A cauda mais longa da distribuição fica à direita, indicando a ocorrência de valores altos com baixa frequência. Esse tipo de distribuição é denominada **assimétrica positiva** ou **à direita**, sendo bastante comum em administração e economia: variáveis como preços, PIB, salários, etc., possuem, em geral, este comportamento.
- Terceiro conjunto de dados – Z – a distribuição também é assimétrica, só que agora, a maior concentração de dados está nos valores mais altos. A cauda mais longa da distribuição fica à esquerda. Esse tipo de distribuição é denominada **assimétrica negativa** ou **à esquerda**.

Conhecer o tipo e intensidade da assimetria de um conjunto de dados pode trazer informações úteis ao analista. Por exemplo, caso a distribuição tenha uma forte assimetria positiva, sabe-se que apesar da alta concentração de dados em valores mais baixos, a média sofrerá influência da cauda à direita deslocando-se em sua direção. Nesse caso, haverá mais observações abaixo da média do que acima dela. O inverso acontece se a assimetria for negativa. Admita que o interesse seja analisar os retornos de duas aplicações, ambas com mesma média e mesma variância, no entanto, uma delas com assimetria positiva e a outra negativa. No caso de assimetria negativa, espera-se que a quantidade de dias com retornos inferiores a média seja maior do que acima, no entanto, a ocorrência de valores muito maiores do que a média é mais comum do que valores muito abaixo dela (o inverso acontece com os retornos

do ativo com assimetria negativa). Essa informação pode ser útil, caso o investidor tenha que se decidir por uma dessas aplicações.

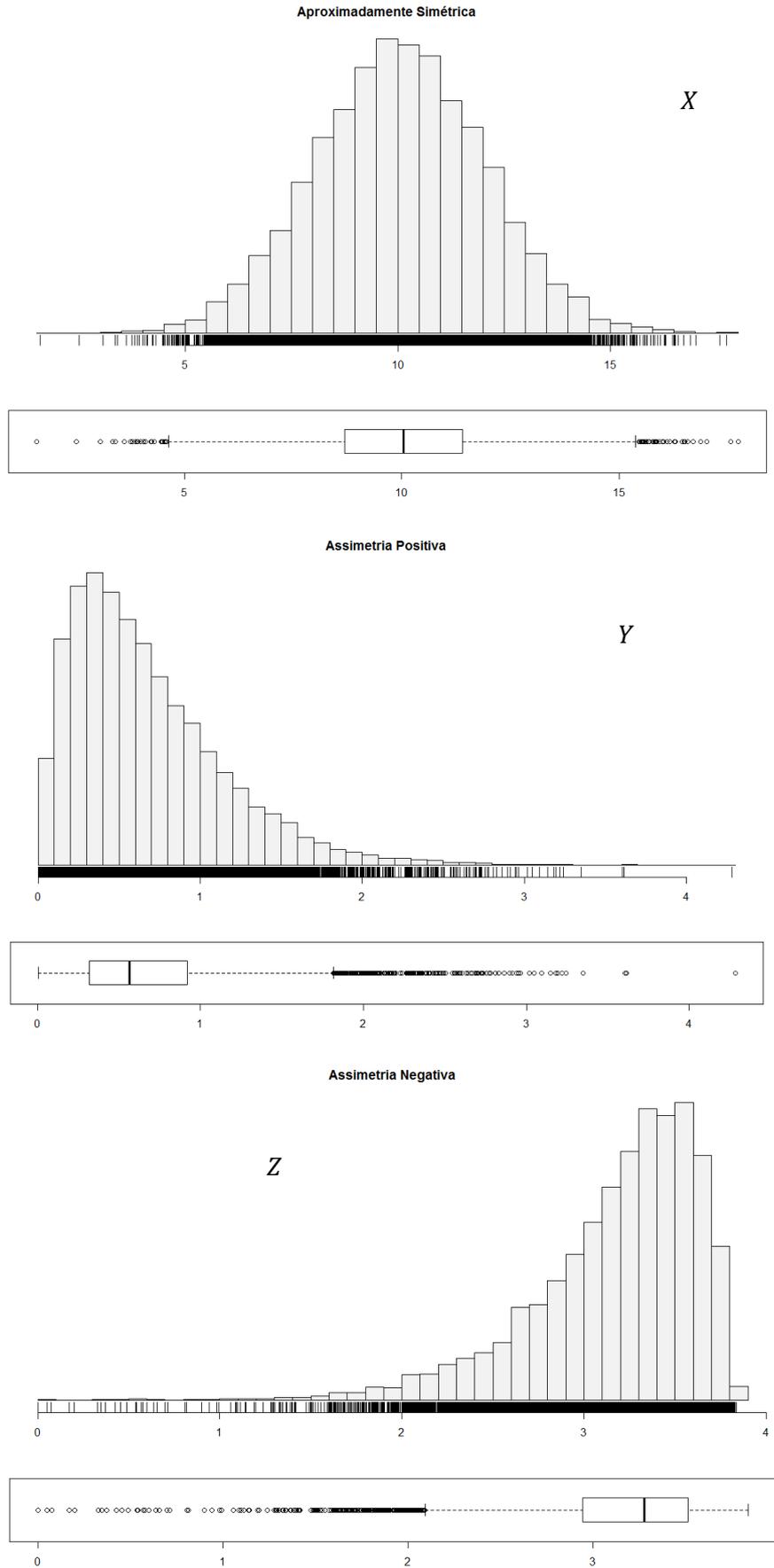


Figura 3: Histogramas e Boxplots de três conjuntos de dados

A Tabela 1 traz algumas medidas descritivas para as variáveis X , Y e Z .

Tabela 1: Estatísticas descritivas para os dados representados na Figura 3.

Estatística	X	Y	Z
Mínimo	1,592	0,004	0,000
Q_1 : primeiro quartil	8,691	0,319	2,946
m_d : mediana	10,050	0,567	3,278
Q_3 : terceiro quartil	11,400	0,918	3,514
Máximo	17,740	4,281	3,838
Média	10,050	0,675	3,172
S : desvio-padrão	1,993	0,477	0,469
m_o : moda ²	10,040	0,294	3,386

Tamanho das amostras = 10.000

Apresentaremos quatro diferentes coeficientes de assimetria: os dois primeiros foram desenvolvidos a partir do comportamento esperado de medidas de tendência central, o terceiro a partir de quartis e o último. A partir do estudo do comportamento dos desvios das observações em relação à sua média. Este último é um dos mais utilizados em modelagem estatística e parece com o nome *skewness* em pacotes estatísticos.

3.1. Coeficiente de assimetria baseados nas medidas de tendência central

O Coeficiente de Assimetria de Pearson, A_p , baseia-se na posição relativa das medidas de tendência central de acordo com o tipo de assimetria dos dados (ver Figura 4). Ele é definido como

$$A_p = \frac{\bar{x} - m_o}{S}.$$

Temos

- Distribuições simétricas unimodais: $\bar{x} = m_d = m_o$; nesse caso, $A_p = 0$
- Distribuições assimétricas positivas: $\bar{x} > m_d > m_o$; então $A_p > 0$
- Distribuições assimétricas negativas: $\bar{x} < m_d < m_o$, fazendo com que $A_p < 0$.

É importante chamar a atenção ao sentido das relações descritas nos itens acima. O tipo de assimetria implica nos diferentes valores de A_p e não o inverso. Na prática, podemos ter distribuições de dados que não se comportam como os histogramas da Figura 3 (por exemplo, distribuições bimodais). Assim, recomenda-se que a análise final sobre o tipo e assimetria seja feita após uma análise gráfica, por exemplo, a construção de um histograma. Os coeficientes de assimetria são úteis para comparar o grau de assimetria entre diferentes conjuntos de dados e o quanto o comportamento observado se afasta de uma distribuição simétrica. Este parágrafo se aplica aos demais coeficientes propostos neste texto.

O fato do denominador de A_p ser o desvio-padrão faz com que essa medida seja adimensional, o que permite sua comparação mesmo quando se trabalha com dados em diferentes escalas (por exemplo, preços em reais ou em dólares).

² Obtida pelo método de Lientz.

Trata-se de uma medida simples, mas com um sério inconveniente. A determinação da moda para dados contínuos não é trivial. Pode-se ter uma amostra de 1000 valores diferentes, por exemplo. Isso requer o uso de algoritmos que levam a diferentes estimativas dessa medida. Uma alternativa é utilizar o coeficiente

$$A_{P2} = \frac{\bar{x} - m_d}{S}.$$

Na Tabela 2, estão apresentados os coeficientes propostos para os dados da Tabela 1. Há indícios de assimetria fraca (quase simetria) para a variável X , assimetria positiva para Y e negativa para Z .

Tabela 2: Coeficientes de assimetria baseados em medidas de tendência central

Variável	A_p	A_{P2}
X	0,005	0,000
Y	0,799	0,226
Z	-0,456	-0,226

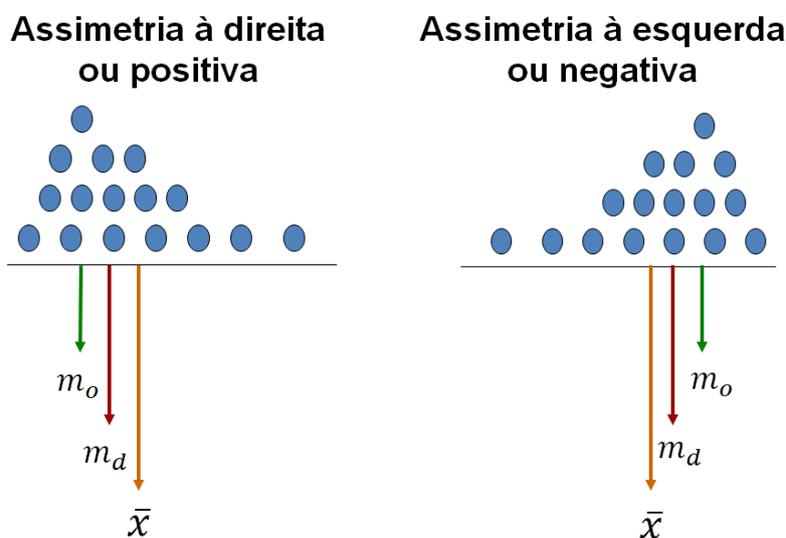


Figura 4: Posição relativa de medidas de tendência central sob assimetria dos dados

3.2. Coeficiente de assimetria baseado em quartis

Para distribuições simétricas, temos que $(Q_3 - m_d) = (m_d - Q_1)$. Por outro lado, é possível perceber ao analisar os boxplots apresentados na Figura 3 que:

- a) Para distribuições assimétricas positivas $(Q_3 - m_d) > (m_d - Q_1)$.
- b) Para distribuições assimétricas negativas $(Q_3 - m_d) < (m_d - Q_1)$.

Observando esses fatos, foi proposto o seguinte coeficiente:

$$A_Q = \frac{(Q_3 - m_d) - (m_d - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2m_d}{Q_3 - Q_1}.$$

A função do denominador, assim como em A_p é fazer com que este coeficiente seja adimensional, permitindo a comparação entre conjuntos de dados medidos em diferentes escalas.

A interpretação é feita da seguinte maneira

- a) Se a distribuição foi simétrica, então $A_Q = 0$.
- b) Se a distribuição foi assimétrica positiva, então $A_Q > 0$.
- c) Se a distribuição foi assimétrica negativa, então $A_Q < 0$.

Para os dados da Tabela 1, temos:

- Para X , $A_Q = -0,003$,
- Para Y , $A_Q = 0,172$ e
- Para Z , $A_Q = -0,169$.

Confirmando as conclusões tiradas na Seção 3.1.

3.3 Coeficiente de assimetria (b_1)

O coeficiente b_1 é um dos mais utilizados para avaliar a assimetria de um conjunto de dados. A lógica de seu desenvolvimento tem origem nos gráficos apresentados na Figura 5.

Simetria	Assimetria Positiva	Assimetria Negativa
<p>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.</p>	<p>Numa distribuição assimétrica positiva, a tendência é que hajam desvios positivos muito maiores do que os negativos</p>	<p>Numa distribuição assimétrica negativa, a tendência é que hajam desvios negativos muito maiores do que os positivos</p>

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

Na Tabela 3 estão dispostas sete observações de três variáveis hipotéticas. Todas têm a mesma média e mesmo desvio-padrão amostral (pelo menos até a segunda casa decimal), no entanto, elas claramente apresentam comportamentos diferentes. A distribuição dos dados da variável A apresenta assimetria positiva; de B negativa e a de C é simétrica. Iremos apresentar o desenvolvimento do Coeficiente de Assimetria b_1 utilizando esses dados.

Retome os histogramas da Figura 5. Os valores a e b indicam desvios em relação à média amostral. Na Tabela 4, apresentamos esses desvios para os dados da Tabela 3. Note que:

- a) para a variável A , há mais desvios negativos, no entanto, de magnitude menor do que os positivos;
- b) para a variável B , ocorre o oposto, há mais desvios positivos, no entanto, de magnitude menor do que os negativos;
- c) em C , para cada desvio negativo, existe um positivo com o mesmo módulo.
- d) Poderíamos, então, propor o cálculo da média dos desvios. Esperaríamos que os sinais dos desvios de maior magnitude predominassem e indicassem o tipo de assimetria presente nos dados. No entanto, pode-se provar que a soma dos desvios em relação à média amostral sempre será zero (ver Resultado 1). Para

eliminar esse problema, e ainda preservar os sinais dos desvios, poderíamos elevá-lo a qualquer potência ímpar e então calcular sua média. A Tabela 5 descreve essa operação utilizando-se a potência 3.

Tabela 3: Conjunto de dados hipotético

Observação	A_i	B_i	C_i
1	10	14	8,87
2	10,2	13,8	10
3	10,8	13,2	11
4	11	13	12
5	12	12	13
6	14	10	14
7	16	8	15,13
Média	12	12	12
Desvio-padrão	2,06	2,06	2,06

Obs: O desvio-padrão foi calculado como a raiz quadrada do segundo momento central amostral de ordem 2 dos dados.

Tabela 4: Desvios em relação a média dos dados da Tabela 1.

i	A_i	$A_i - \bar{A}$	B_i	$B_i - \bar{B}$	C_i	$C_i - \bar{C}$
1	10	-2	14	2	8,87	-3,13
2	10,2	-1,8	13,8	1,8	10	-2
3	10,8	-1,2	13,2	1,2	11	-1
4	11	-1	13	1	12	0
5	12	0	12	0	13	1
6	14	2	10	-2	14	2
7	16	4	8	-4	15,1	3,13
Média	12		12		12	
DP	2,06		2,06		2,06	

Resultado 1: Seja X_1, \dots, X_n , uma amostra de uma variável quantitativa. Então,

$$\sum_{i=1}^n (X_i - \bar{x}) = 0, \quad \bar{x} = \sum_{i=1}^n \frac{x_i}{n}.$$

Prova: $\sum_{i=1}^n (X_i - \bar{x}) = \sum_{i=1}^n X_i - n \bar{x} = \sum_{i=1}^n X_i - \sum_{i=1}^n X_i = 0$.

Os valores das médias dos desvios ao cubo para A , B e C são, respectivamente, 7,92; -7,92 e 0. Notem que o sinal indica o tipo de assimetria presente nos dados, e que esses valores correspondem ao momento central amostral de ordem 3. O momento amostral é uma média, e, no exemplo, os sinais das maiores distâncias acabam predominando no cálculo dessa média. Em geral, os momentos m_k , sendo $k > 1$ um número ímpar podem ser utilizados como indicadores do tipo de assimetria presente nos dados.

Os momentos m_k , no entanto, têm um inconveniente. Eles dependem da unidade de medida dos dados. Imagine uma amostra de preços em dólares convertida para reais. Obviamente nada mudou em termos da assimetria, todavia, os terceiros momentos amostrais não irão coincidir, já que $m_k(\text{reais}) = (\text{taxa de câmbio})^k m_k(\text{dólares})$.

Tabela 5: Desvios em relação a média dos dados da Tabela 1.

i	A_i	$A_i - \bar{A}$	$(A_i - \bar{A})^3$	B_i	$B_i - \bar{B}$	$(B_i - \bar{B})^3$	C_i	$C_i - \bar{C}$	$(C_i - \bar{C})^3$
1	10	-2	-8	14	2	8	8,87	-3,1	-30,66
2	10,2	-1,8	-5,83	13,8	1,8	5,832	10	-2	-8
3	10,8	-1,2	-1,73	13,2	1,2	1,728	11	-1	-1
4	11	-1	-1	13	1	1	12	0	0
5	12	0	0	12	0	0	13	1	1
6	14	2	8	10	-2	-8	14	2	8
7	16	4	64	8	-4	-64	15,1	3,13	30,66
Média	12		7,92	12		-7,92	12		0
DP	2,06			2,06			2,06		

Um modo de contornar esse problema é refazer os cálculos utilizando-se os dados padronizados. A Tabela 6 apresenta essas contas. Agora, mesmo que mudemos a escala de uma coluna o terceiro momento amostral da variável padronizada não sofrerá alterações.

Tabela 6: Desvios em relação a média dos dados da Tabela 3, dados padronizados.

i	A_i	Z_{A_i}	$Z_{A_i}^3$	B_i	Z_{B_i}	$Z_{B_i}^3$	C_i	Z_{C_i}	$Z_{C_i}^3$
1	10	-0,97	-0,92	14,00	0,97	0,92	8,87	-1,52	-3,53
2	10,2	-0,87	-0,67	13,80	0,87	0,67	10,00	-0,97	-0,92
3	10,8	-0,58	-0,20	13,20	0,58	0,20	11,00	-0,49	-0,12
4	11	-0,49	-0,11	13,00	0,49	0,11	12,00	0,00	0,00
5	12	0,00	0,00	12,00	0,00	0,00	13,00	0,49	0,12
6	14	0,97	0,92	10,00	-0,97	-0,92	14,00	0,97	0,92
7	16	1,94	7,33	8	-1,9	-7,3	15,1	1,52	3,53
Média	12		0,91	12		-0,91	12		0
DP	2,06			2,06			2,06		

Definição 2: Seja X_1, X_2, \dots, X_n um conjunto de dados e $Z_i = \frac{X_i - \bar{X}}{S}$, $i = 1, 2, \dots, n$. Define-se o Coeficiente de Assimetria (Amostral) dos dados por

$$b_1 = \sum_{i=1}^n \frac{Z_i^3}{n}.$$

Alternativamente, b_1 pode ser reescrito como

$$b_1 = \sum_{i=1}^n \frac{Z_i^3}{n} = \frac{m_3}{S^3}.$$

Em resumo temos:

- se a distribuição é assimétrica positiva $\Rightarrow b_1 > 0$;
- se a distribuição é assimétrica negativa $\Rightarrow b_1 < 0$;
- se a distribuição é (perfeitamente) simétrica $\Rightarrow b_1 = 0$.

Retomando o exemplo tratado na Tabela 1, $b_1(X) = 0,010$, $b_1(Y) = 1,361$ e $b_1(Z) = -1,486$.

3.3.1. Determinação de b_1 para dados agrupados

Os dados da Tabela 7 resumem o grau de endividamento de clientes de uma carteira de empréstimos (Carteira Alfa). A partir dessa tabela chegou-se a uma média de 18,17 e desvio-padrão 14,09. A Figura 6 é o histograma construído a partir dos dados apresentados. Há claros indícios de existência de uma distribuição assimétrica positiva. O coeficiente b_1 trará indicações sobre a intensidade dessa assimetria.

Tabela 7: Distribuição de frequências e densidades de frequência do grau de endividamento de clientes da carteira Alfa

Grau de Endividamento	n_i	$f_i \times 100$	$F_i \times 100$
0 -- 5	61	12,2	12,2
5 -- 10	107	21,4	33,6
10 -- 15	97	19,4	53,0
15 -- 20	77	15,4	68,4
20 -- 30	77	15,4	83,8
30 -- 50	63	12,6	96,4
50 -- 75	18	3,6	100
Total	500	1,000	

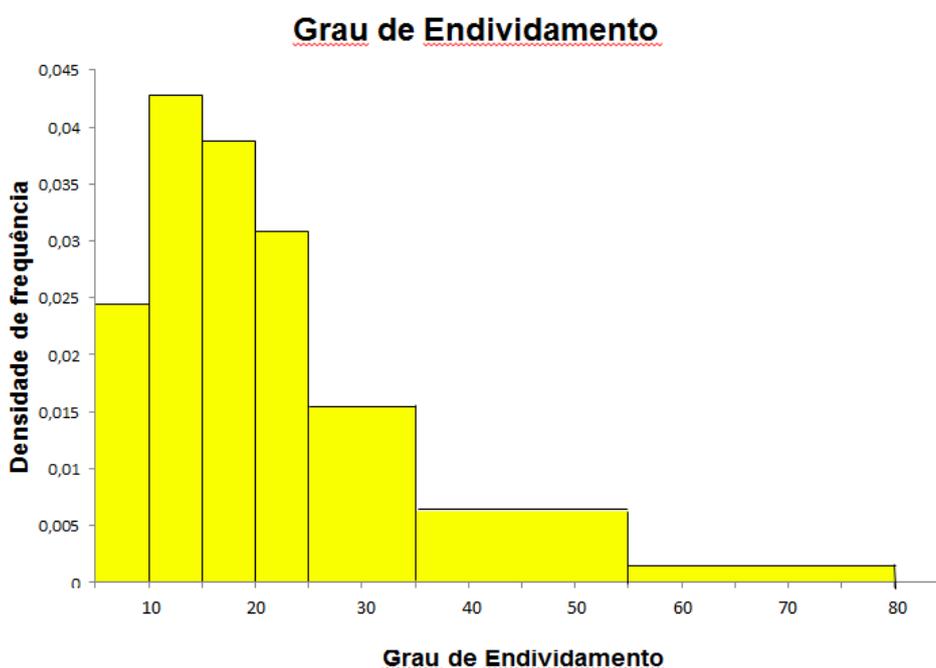


Figura 6: Histograma do Grau de Endividamento dos clientes da Carteira Alfa.

A lógica do cálculo desse indicador é a mesma utilizada quando determinamos a média e a variância a partir de dados agrupados. Assumimos que os dados se distribuem uniformemente em cada faixa de grau de endividamento,

de modo que $\sum_{j=1}^n (x_j - \bar{x})^3$ pode ser aproximada por $\sum_{i=1}^k n_i (c_i - \bar{x})^3$, sendo c_i o ponto médio da faixa i e k , o número de faixas utilizadas na tabela. Desse modo,

$$m_3 = \sum_{j=1}^k \frac{n_i (c_i - \bar{x})^3}{n} = \sum_{j=1}^k \frac{n_i}{n} (c_i - \bar{x})^3 = \sum_{j=1}^k f_i (c_i - \bar{x})^3.$$

A Tabela 8, resume o cálculo de m_3 para os dados da Tabela 7. Utilizando o valor obtido, temos que

$$b_1 = \frac{m_3}{S^3} = \frac{3731,20}{14,09^3} = 1,33.$$

Tabela 8: Determinação de m_3 para os dados da Carteira Alfa.

Grau de Endividamento	n_i	c_i	$(c_i - \bar{x})^3$	$n_i (c_i - \bar{x})^3$	f_i	$f_i (c_i - \bar{x})^3$
0 -- 5	61	2,5	-3847,75	-234712,83	0,122	-469,43
5 -- 10	107	7,5	-1214,77	-129980,15	0,214	-259,96
10 -- 15	97	12,5	-182,28	-17681,574	0,194	-35,36
15 -- 20	77	17,5	-0,30	-23,158751	0,154	-0,05
20 -- 30	77	25,0	318,61	24533,123	0,154	49,07
30 -- 50	63	40,0	10403,06	655392,937	0,126	1310,79
50 -- 75	18	62,5	87115,05	1568070,91	0,036	3136,14
Total	500			1865599,26		
				1865599,26/500=	$m_3 =$	3731,20
			$m_3 =$	3731,20		