Medidas descritivas

Medidas resumo numéricas

- Tendência central dos dados
  - Média
  - Mediana
  - Moda
- Dispersão ou variação em relação ao centro
  - Amplitude
  - Intervalo interquartil
  - Variância
  - Desvio Padrão
  - Coeficiente de variação
- Medidas de simetria

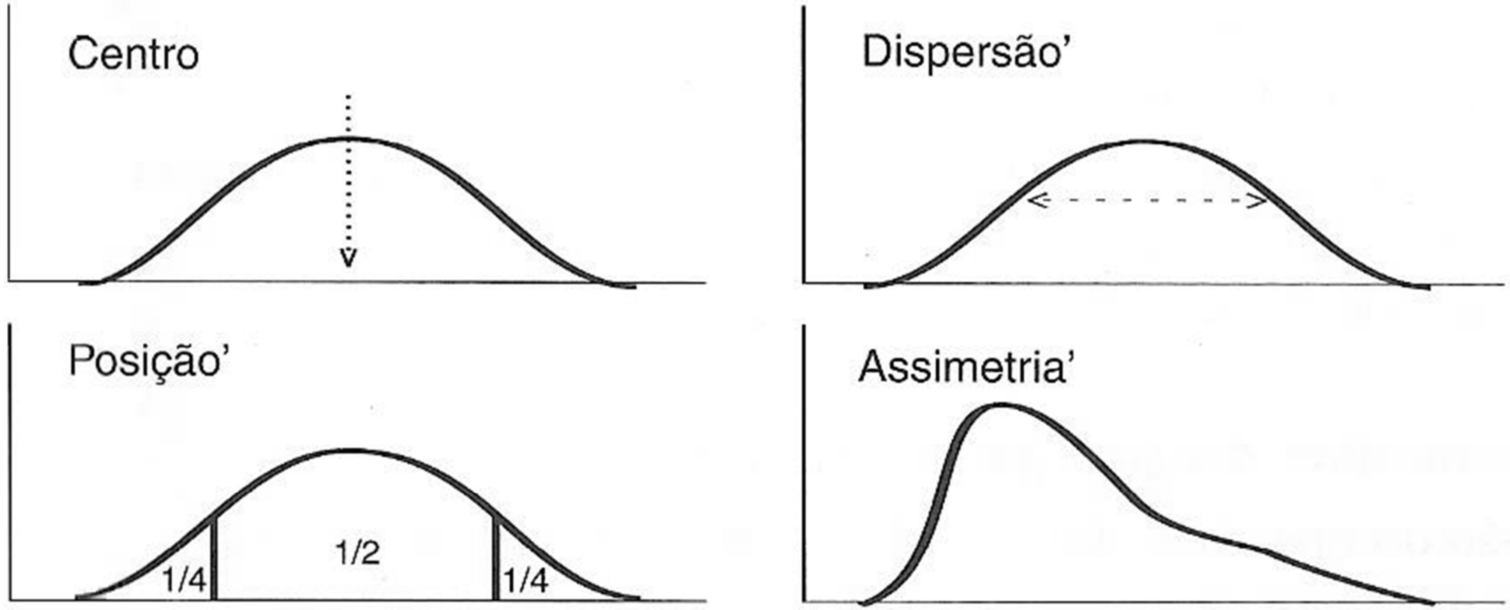


Figura 2.1 Medidas representativas de um conjunto de dados estatísticos.

Extraído de Diaz FR, López FJB. Bioestatística. São Paulo: Thomson Lerarning; 2007

Modalidade	Freqüências Absolutas	Freqüências Relativas	Freqüências Absolutas Acumuladas	Freqüências Relativas Acumuladas
C	n_i	f_i	N_i	F_i
c_1	n_1	$f_1 = \dfrac{n_1}{n}$	$N_1 = n_1$	$F_1 = \dfrac{N_1}{n} = f_1$
...	...	...	...	...
c_j	n_j	$f_j = \dfrac{n_j}{n}$	$N_j = n_1 + ... + n_j$	$F_j = \dfrac{N_j}{n} = f_1 + ... + f_j$
...	...	...	...	...
c_k	n_k	$f_k = \dfrac{n_k}{n}$	$N_k = n$	$F_k = 1$
	n	1		

Média Aritmética

É a soma de todas as observações de um conjunto de dados, e divisão do resultado pelo número total de medidas.

X	n_i	f_i
x_1	n_1	f_1
...	...	...
x_k	n_k	f_k

a média é o valor que podemos escrever das seguintes formas equivalentes:

$$\begin{aligned} \overline{x} &= x_1 f_1 + \ldots + x_k f_k \\ &= \frac{1}{n}\left(x_1 n_1 + \ldots + x_k n_k\right) \\ &= \frac{1}{n}\sum_{i=1}^{k} x_i n_i \end{aligned}$$

Se os dados não estão ordenados em uma tabela, então:

$$\boxed{\overline{x} = \frac{x_1 + \ldots + x_n}{n}}$$

Tab 3.1 Volumes expiratórios forçados em 1 segundo para 13 adolescentes que sofrem de asma. Local X, Ano Y.

Indivíduos	FEV (litros)
1	2,30
2	2,15
3	3,50
4	2,60
5	2,75
6	2,82
7	4,05
8	2,25
9	2,68
10	3,00
11	4,02
12	2,85
13	3,38

$$\begin{aligned}\bar{x} &= \frac{1}{13}\sum_{i=1}^{13} x_i \\ &= \left(\frac{1}{13}\right)(2{,}30 + 2{,}15 + 3{,}50 + 2{,}60 + 2{,}75 + 2{,}82 + 4{,}05 \\ &\qquad + 2{,}25 + 2{,}68 + 3{,}00 + 4{,}02 + 2{,}85 + 3{,}38) \\ &= \frac{38{,}35}{13} \\ &= 2{,}95 \text{ litros.}\end{aligned}$$

l_{i-1} l_i	n_i	x_i	$x_i n_i$
0 |- 10	60	5	300
10 |- 20	80	15	1200
20 |- 30	30	25	750
30 |- 100	20	65	1300
100 |- 500	10	300	3000
Total	200		$\Sigma x_i n_i = 6550$

$$\overline{X} = \frac{\Sigma x_i n_i}{n} = \frac{6550}{200} = 32{,}75$$

Alguns inconvenientes da média:

- **é sensível aos valores extremos**
- **na variável discreta, a média pode não pertencer ao conjunto de valores da variável**

Mediana

- Não é afetada pelas observações extremas
- É uma medida resumo para observações ordinais, dados discretos e contínuos
- Na variável discreta é sempre um valor observado
- É o primeiro valor que deixa abaixo de si 50% das observações
- Se n for ímpar, a mediana = [(n+1)/2]
- Se n for par, mediana= $\dfrac{(n/2) + [(n/2) + 1]}{2}$

Tab 3.1 Volumes expiratórios forçados em 1 segundo para 13 adolescentes que sofrem de asma. Local X, Ano Y.

Indivíduos	FEV (litros)
1	2,30
2	2,15
3	3,50
4	2,60
5	2,75
6	2,82
7	4,05
8	2,25
9	2,68
10	3,00
11	4,02
12	2,85
13	3,38

2,15
2,25
2,30
2,60
2,68
2,75
2,82 ←
2,85
3,00
3,38
3,50
4,02
4,05

$$\text{Med} = \frac{(n+1)}{2} = \frac{13+1}{2} = 7^{\text{a}} \text{ observ}$$

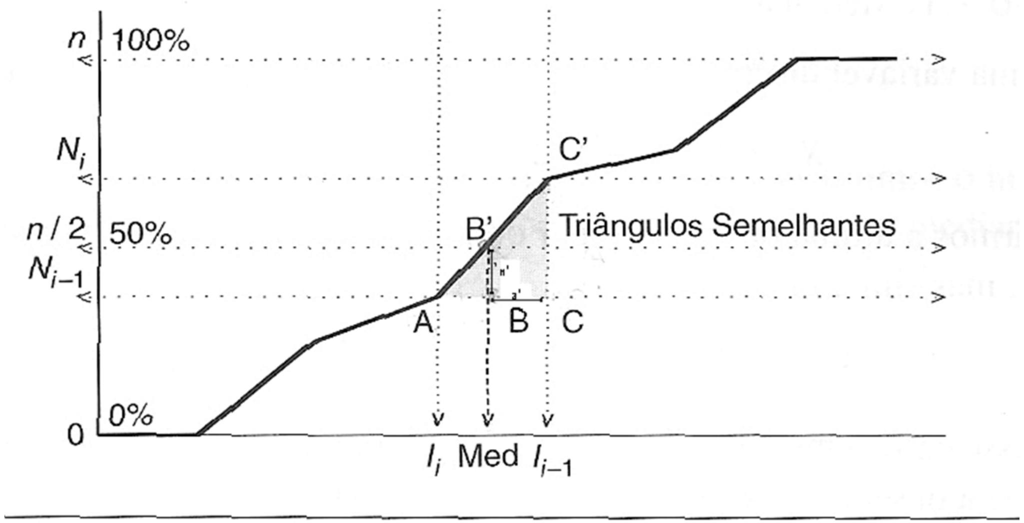


Figura 2.2 Cálculo geométrico da mediana.

Teorema de Tales

$$\frac{CC'}{AC} = \frac{BB'}{AB} \Longrightarrow \frac{n_i}{a_i} = \frac{\frac{n}{2} - N_{i-1}}{M_{ed} - l_{i-1}}$$

$$\Longrightarrow \boxed{M_{ed} = l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i}$$

l_{i-1} — l_i	n_i	x_i	$x_i n_i$	a_i	N_i
0 |- 10	60	5	300	10	60
10 |- 20	80	15	1200	10	140
20 |- 30	30	25	750	10	170
30 |- 100	20	65	1300	70	190
100 |- 500	10	300	3000	400	200
Total	200				

A 1ª freqüência acumulada que supera o valor n/2 = 100 é N_i=140. Por isso, o intervalo mediano é [10;20]. Assim,

$$M_{ed} = l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i \quad = 10 + \frac{100 - 60}{80} \times 10 = 15$$

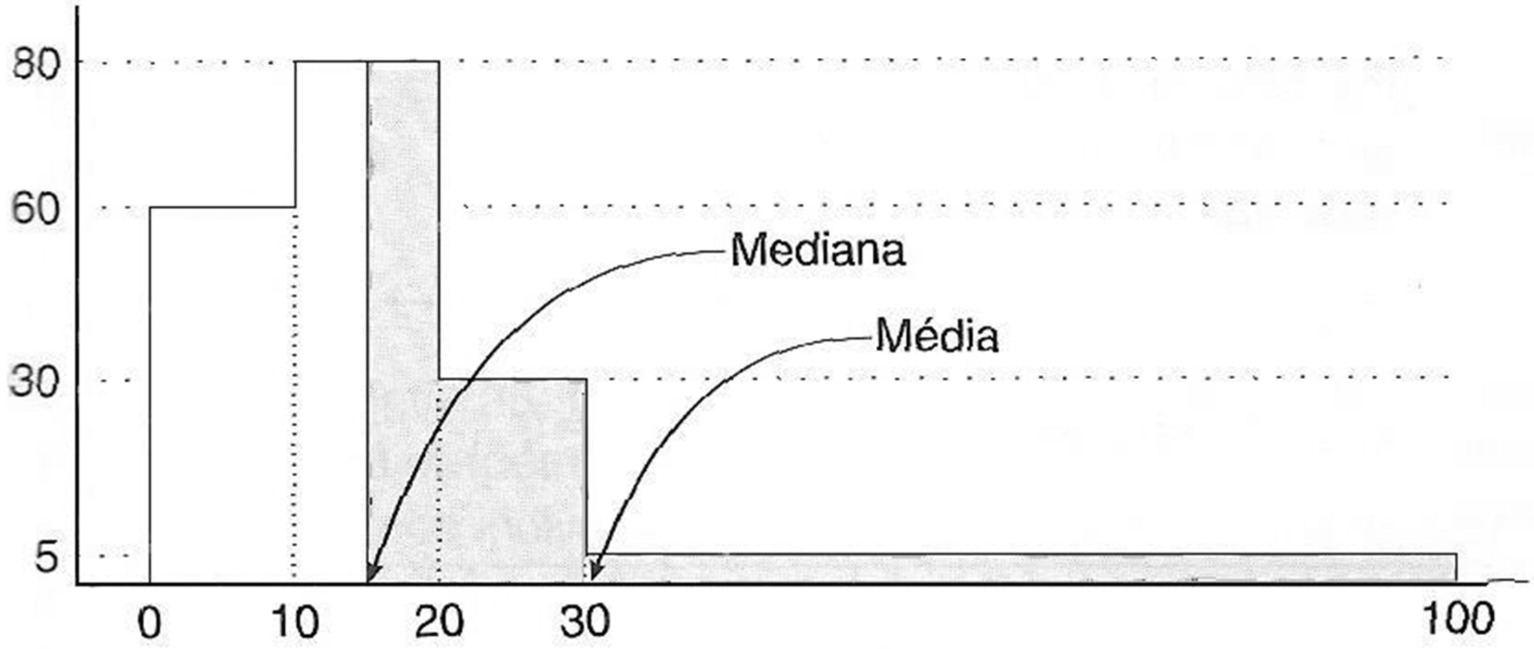


Figura 2.3 *Para essa distribuição de freqüências*, é mais representativo usar, como estatística de *tendência central*, a *mediana* que a *média*.

Moda

- Qualquer máximo valor de uma distribuição
- É fácil de calcular
- Pode não ser única

$Mo = x_i$ de maior freqüência

Método de King

$$Mo = l_{i-1} + a_i \cdot \frac{f_3}{f_1 + f_3}$$

a_i = amplitude da classe onde se localiza a moda

f_1 e f_3 = freqüência das classes adjacentes à moda

f_2 = freqüência da classe onde se encontra a moda

Método de Czuber

$$Mo = l_{i-1} + a_i \cdot \frac{\Delta a}{\Delta a + \Delta p}$$

$\Delta a = (f_1 - f_2)$

$\Delta p = (f_2 - f_3)$

l_{i-1} l_i	n_i	x_i	$x_i n_i$	a_i	N_i
0 |- 10	60	5	300	10	60
10 |- 20	80	15	1200	10	140
20 |- 30	30	25	750	10	170
30 |- 100	20	65	1300	70	190
100 |- 500	10	300	3000	400	200
Total	200				

$$Mo = l_{i-1} + a_i \cdot \frac{f_3}{f_1 + f_3}$$

$$Mo = 10 + 10 \cdot \frac{30}{60 + 30} = 13{,}33$$

17 é o valor que representa a moda neste conjunto de dados agrupados

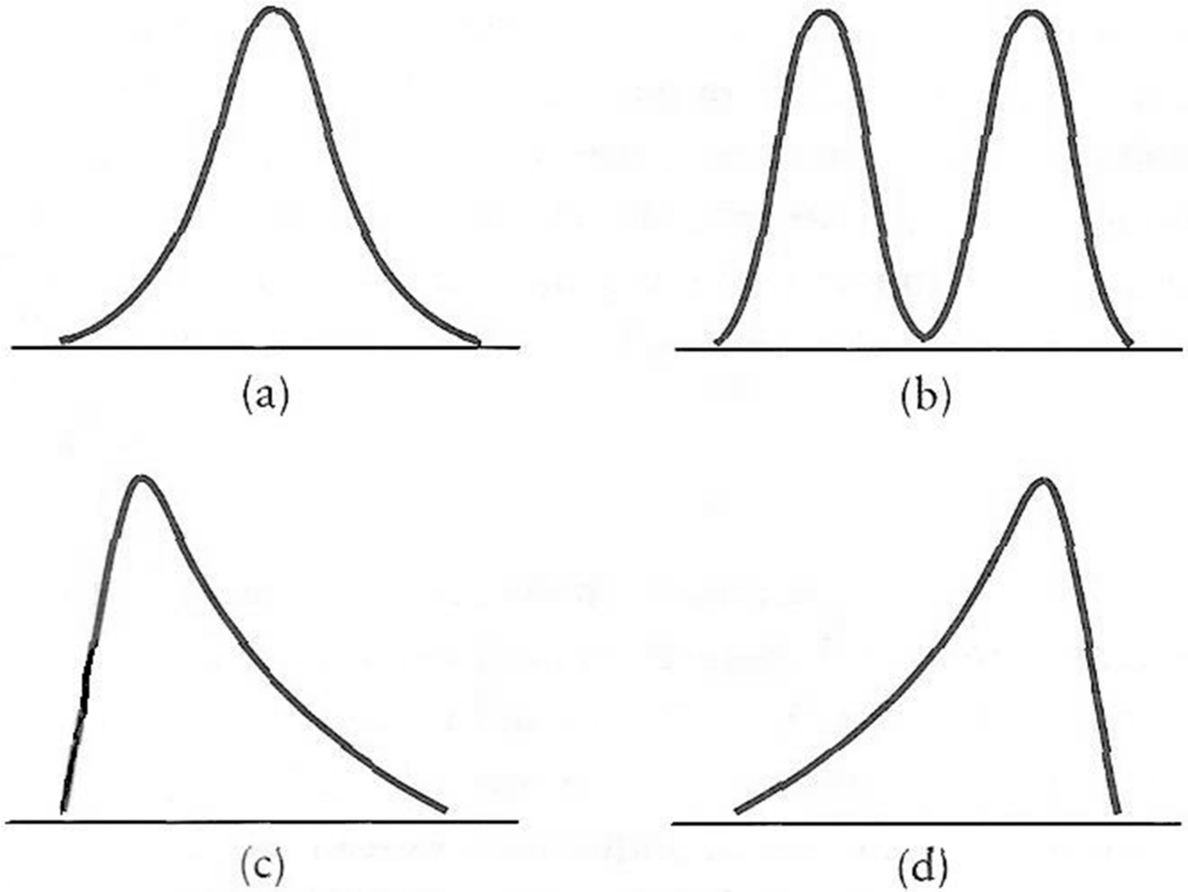


FIGURA 3.1
Possíveis distribuições dos valores de dados.

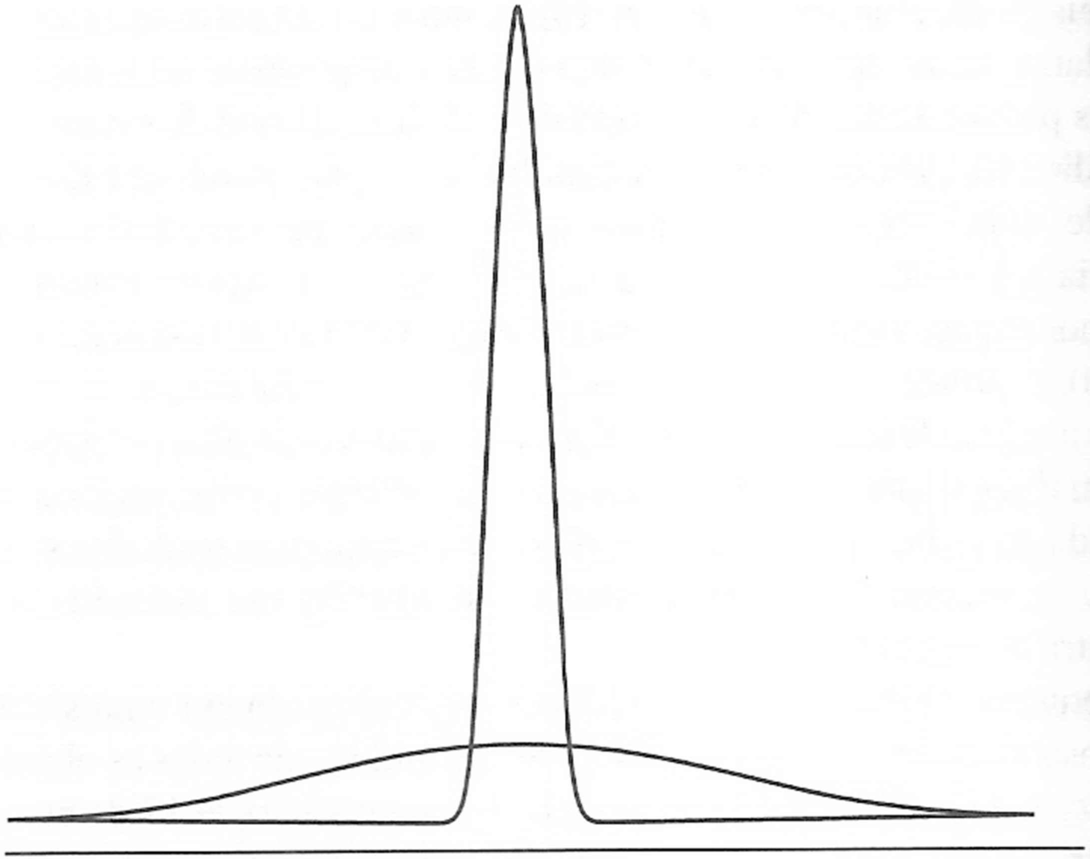


FIGURA 3.2
Duas distribuições com médias, medianas e modas idênticas.

Medidas de variabilidade ou dispersão

AMPLITUDE

- é a diferença entre a maior observação e a menor
- uso limitado, pois só considera os valores extremos de um conjunto de dados
- É altamente sensível aos valores excepcionalmente grandes ou pequenos

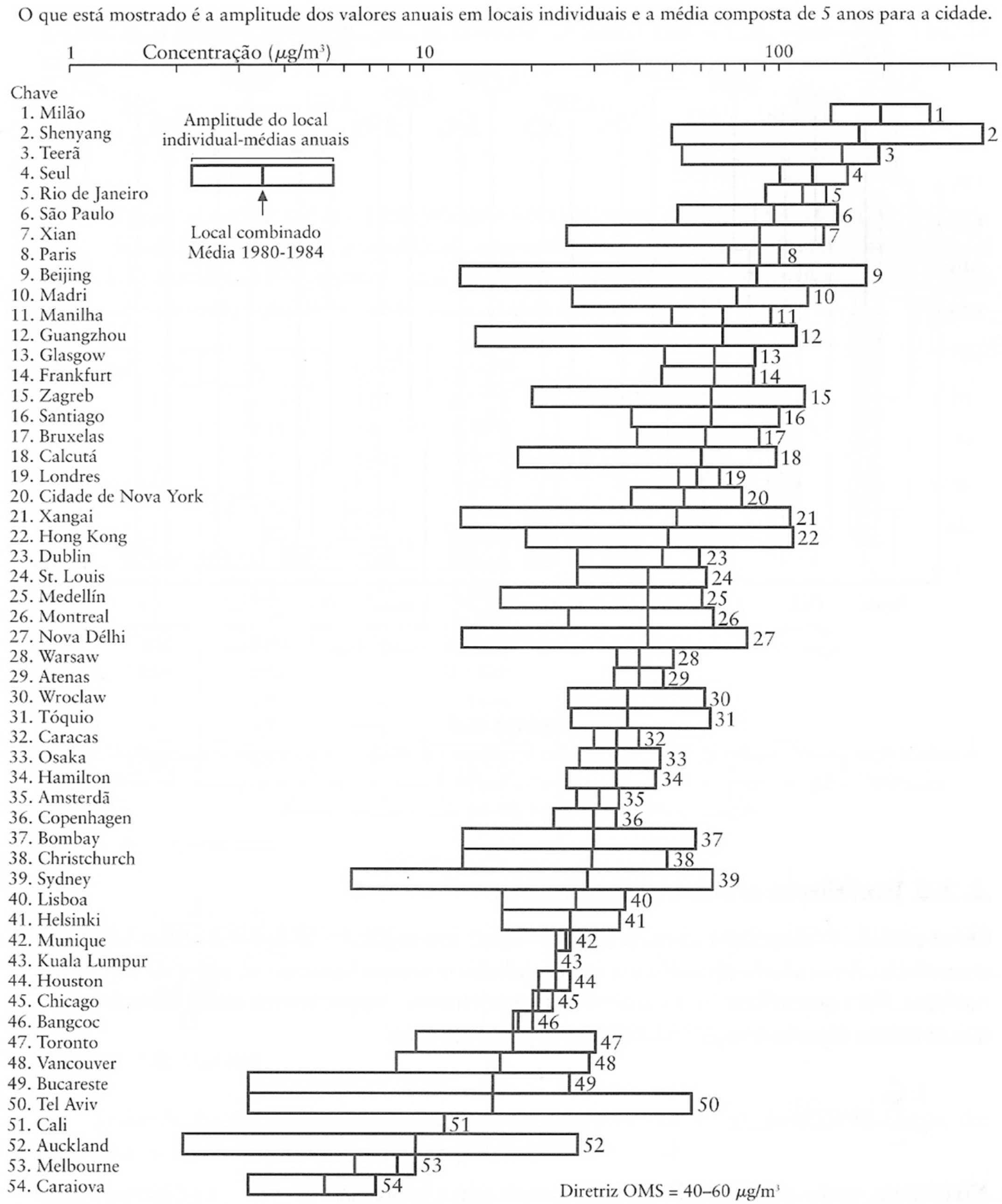


FIGURA 3.3
Resumo das médias de dióxido de enxofre, 1980–1984.

- Os **percentis** são úteis para descrever a forma de uma distribuição
- São valores da variável caracterizados por superar uma certa porcentagem de observações na população.
- Decis dividem a distribuição em dez grupos iguais
- Quartis: dividem as observações em quatro grupos iguais
    - $Q1 = P_{25}$
    - $Q2 = P_{50} = Med$
    - $Q3 = P_{75}$
    - Para cálculo dos quartis, utiliza-se a mesma fórmula da mediana ($Q1 = n/4$; $Q2 = n/2$ e

      $Q3 = 3n/4$)

$$Q_1 = l_{i-1} + \frac{\frac{n}{4} - N_{i-1}}{n_i} \cdot a_i$$

$$Q_2 = l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i$$

$$Q_3 = l_{i-1} + \frac{\frac{3n}{4} - N_{i-1}}{n_i} \cdot a_i$$

Intervalo interquartil ou amplitude quartil

- Subtrai-se o o valor da posição do 25º percentil dos dados do valor da posição do 75º percentil que conseqüentemente engloba os 50% do meio das observações
- Não está sujeita às flutuações dos valores extremos
- Se o nº de observações for ímpar – nk/100 (posição)
- Se o nº de observações for par - a média entre (nk/100) e (nk/100 +1),

ou seja: Q3 – Q1

Amplitude semi-quartil

$$Q = \frac{Q3 - Q1}{2}$$

Vantagem sobre a amplitude quartil

- ❖ A divisão por 2 dá a distância média pela qual os quartis se desviam da mediana

VARIÂNCIA

- ✓ s^2 : é a média das diferenças quadráticas de n valores em relação à sua média aritmética.
- ✓ quantifica a variabilidade ou o espalhamento ao redor da média das medidas

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

Tab 3.1 Volumes expiratórios forçados em 1 segundo para 13 adolescentes que sofrem de asma. Local X, Ano Y.

Indivíduo	X_i	$X_i - \overline{x}$	$(x_i - \overline{x})^2$
1	2,30	- 0,65	0,4225
2	2,15	- 0,80	0,6400
3	3,50	0,55	0,3025
4	2,60	- 0,35	0,1225
5	2,75	- 0,20	0,0400
6	2,82	- 0,13	0,0169
7	4,05	1,10	1,2100
8	2,25	- 0,70	0,4900
9	2,68	- 0,27	0,0729
10	3,00	0,05	0,0025
11	4,02	1,07	1,1449
12	2,85	- 0,10	0,0100
13	3,38	0,43	0,1849
Total	38,35	0,00	4,6596

$$S^2 = \frac{1}{(13-1)} \sum_{i-1}^{13} (x_i - 2,95)^2$$

$$S^2 = \frac{4,6596}{12} = 0,39 \text{ litros}^2$$

DESVIO PADRÃO

- Extrai-se a raiz quadrada da variância para que ela tenha a mesma unidade de medida que a média
- Quanto menor o desvio padrão mais homogêneas são as observações
- A magnitude real dos desvio padrão depende dos valores do conjunto de dados
- A soma de todos os desvios em torno da média aritmética da distribuição das medidas é zero

$$\boxed{s = \sqrt{s^2}}$$

$$\boxed{s = \sqrt{0,39} = 0,62 \text{ litro}}$$

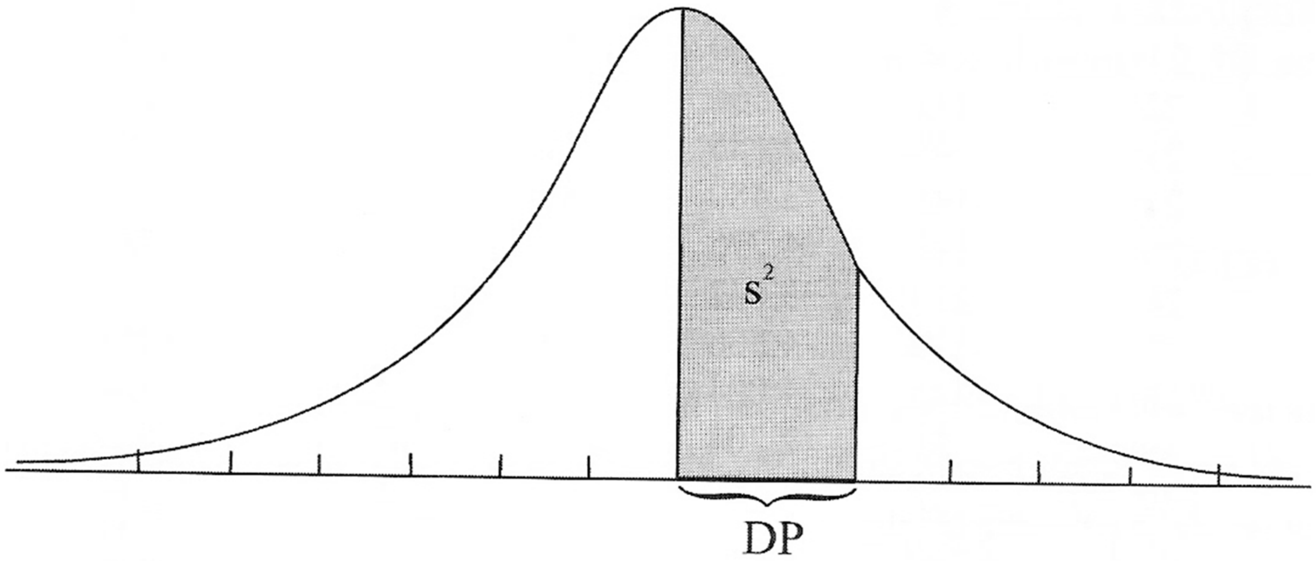


Figura 2-7. Ilustração da variância e do desvio padrão

A variância representa uma área e o DP representa um intervalo

COEFICIENTE DE VARIAÇÃO DE PEARSON

- Permite comparar a variabilidade entre 2 ou mais conjuntos de dados que representam quantidades variadas com diferentes unidades de medida
- é adimensional

$$CV = \frac{s}{\bar{x}} \times 100\%$$

$$\boxed{CV = \frac{0,62}{2,95} \times 100\% = 21\%}$$

COEFICIENTE DE VARIAÇÃO QUARTIL

- utilizada quando a tendência central e a variabilidade são medidas em termos dos quartis

$$V_q = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$$

DISTRIBUIÇÃO DO Nº DE GESTANTES SEGUNDO CENTRO DE SAÚDE E Nº DE FILHOS, 1997

LOCAL B

X	f
0	30
1	-
2	40
3	-
4	30
TOTAL	100

LOCAL C

X	f
0	10
1	20
2	40
3	20
4	10
TOTAL	100

LOCAL D

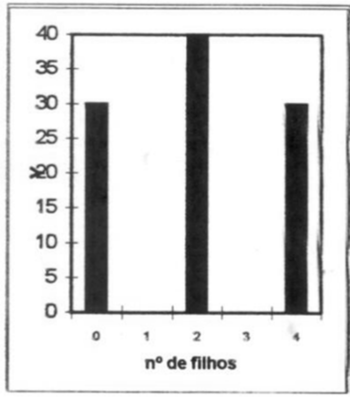
X	f
0	50
1	-
2	-
3	-
4	50
TOTAL	100

LOCAL E

X	f
0	-
1	-
2	100
3	-
4	-
TOTAL	100

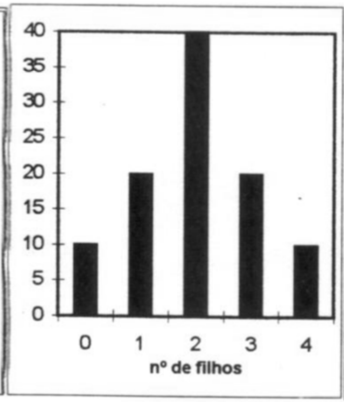
DISTRIBUIÇAO DO Nº DE GESTANTES SEGUNDO CENTRO DE SAUDE E Nº DE FILHOS, 1997

LOCAL B



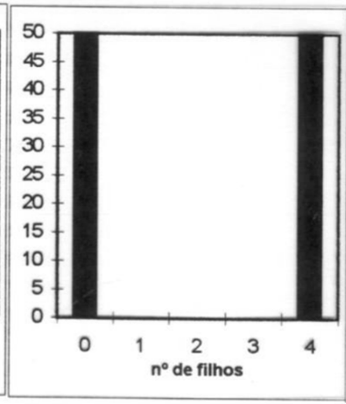
média = 2 filhos
mediana = 2 filhos
moda = 2 filhos
amplitude = 4 filhos
variância = 2,4 (filhos)^2
d. padrão =1,6 filhos
C.V.Pearson = 80%

LOCAL C



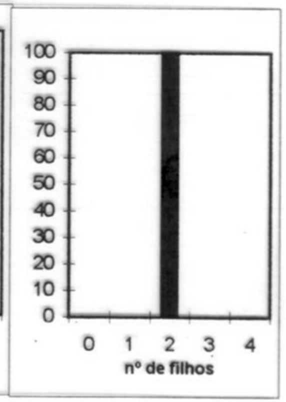
média = 2 filhos
mediana = 2 filhos
moda = 2 ilhos
amplitude = 4 filhos
variância =1,2 (filhos)^2
d. padrão =1,1 filhos
C.V.Pearson = 55%

LOCAL D



média = 2 filhos
mediana = 2 filhos
moda = 0 e 4 filhos
amplitude = 4 filhos
variância = 4 (filhos)^2
d. padrão = 2 filhos
C.V.Pearson = 100%

LOCAL E



média = 2 filhos
mediana = 2 filhos
moda = 2 filhos
amplitude = 0 filhos
variância = 0(filhos)^2
d. padrão = 0 filhos
C.V.Pearson= 0%

GOTLIEB, S.L.D.
HEP-5732 FSP/USP

l_{i-1} l_i	n_i	x_i	$x_i n_i$
80 |-| 119	13	99,5	1.293,5
120 |-| 159	150	139,5	20.925,0
160 |-| 199	442	179,5	79.339,0
200 |- |239	299	219,5	65.630,5
240 |-| 279	115	259,5	29.842,5
280|-| 319	34	299,5	10.183,0
320|-| 359	9	339,5	3.055,5
360|-| 399	5	379,5	1.897,5
Total	1067		Σ_i= 212.166,5

$$\overline{X} = = \frac{212.166,5}{1067} = 198,8/100ml$$

l_{i-1} l_i	n_i	x_i	$X_i - \overline{x}$	$(X_i - \overline{x})^2$	$(X_i - \overline{x})^2 \times n_i$
80 |-| 119	13	99,5	- 99,3	9860,49	128.186,37
120 |-| 159	150	139,5	- 59,3	3516,49	527.473,5
160 |-| 199	442	179,5	- 19,3	372,49	164.640,58
200 |- |239	299	219,5	20,7	428,49	128.118,51
240 |-| 279	115	259,5	60,7	3684,49	423.716,35
280|-| 319	34	299,5	100,7	10.140,49	344.776,66
320|-| 359	9	339,5	140,7	19.796,49	178.168,41
360|-| 399	5	379,5	180,7	32.652,49	163.262,45
Total	1067				2.058.342,8

$$S^2 = \frac{2.058.342,8}{1066} = 1.930,9 \ (\text{mg}/100\text{ml})^2$$

$$s = \sqrt{1.930,9} = 43,9 \ \text{mg}/100\text{ml}$$

$$CV = \frac{43,9}{198,8} \times 100\% = 22\%$$

MEDIDAS DE ASSIMETRIA

- Análise do histograma e da mediana

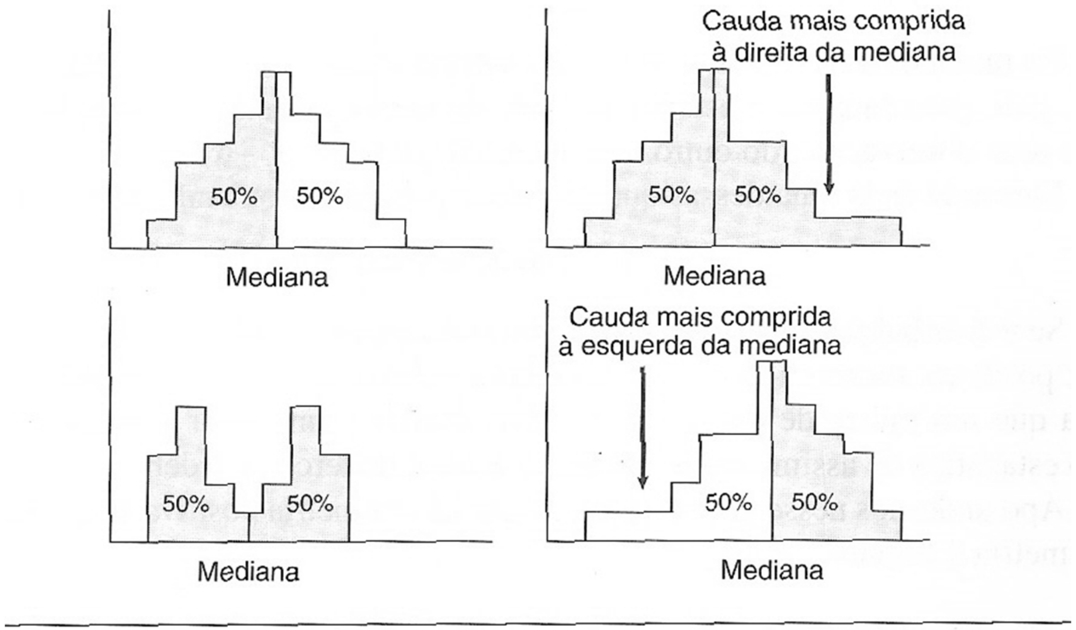
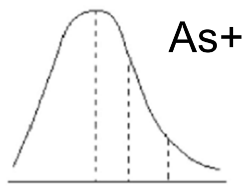
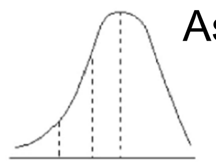


Figura 2.6 Distribuições de freqüências simétricas e assimétricas.



As+



As -

- Assimetria positiva: média $>$ mediana $>$ moda
- Assimetria negativa: média $<$ mediana $<$ moda

Coeficiente de assimetria de Pearson

- Baseia-se na diferença entre média e moda

$$sk = \frac{\bar{x} - M_o}{s}$$

- Substitui-se a moda pelo valor dado por Pearson para distribuições assimétricas, $M_o = 3\,Me - 2\,\bar{x}$

$$sk = \frac{3\,(\bar{x} - Me)}{s}$$

Assimetria positiva $sk>0$ e assimetria negativa $sk<0$

Coeficiente de assimetria de Pearson

- Simétrica, se $|As| < 0{,}15$
- Assimétrica moderada, se $0{,}15 \leq |As| < 1{,}0$
- Assimétrica forte, se $|As| \geq 1{,}0$

Coeficiente quartil de assimetria

- Índice de Yule – Bowley
- Se a distribuição é simétrica $Q_3 - Q_2 = Q_2 - Q_1$
- Se assimétrica positiva $Q_3 - Q_2 > Q_2 - Q_1$

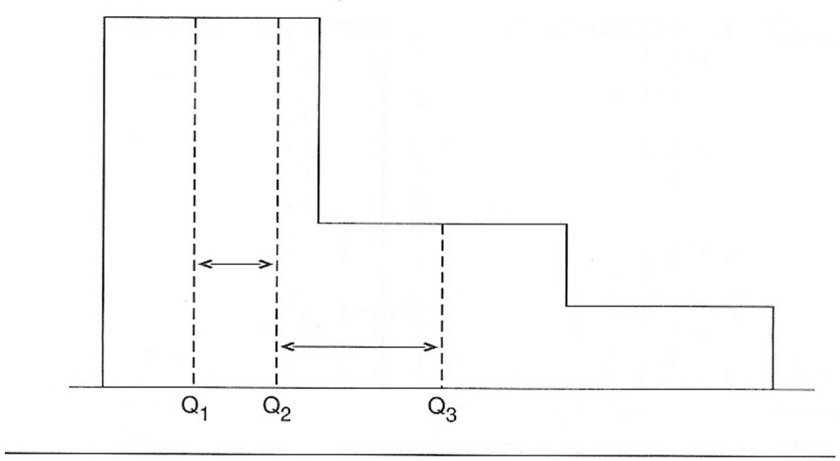


Figura 2.7 Uso dos quartis para medir a assimetria.

$$A_s = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$

$$-1 \leq A_s \leq 1$$

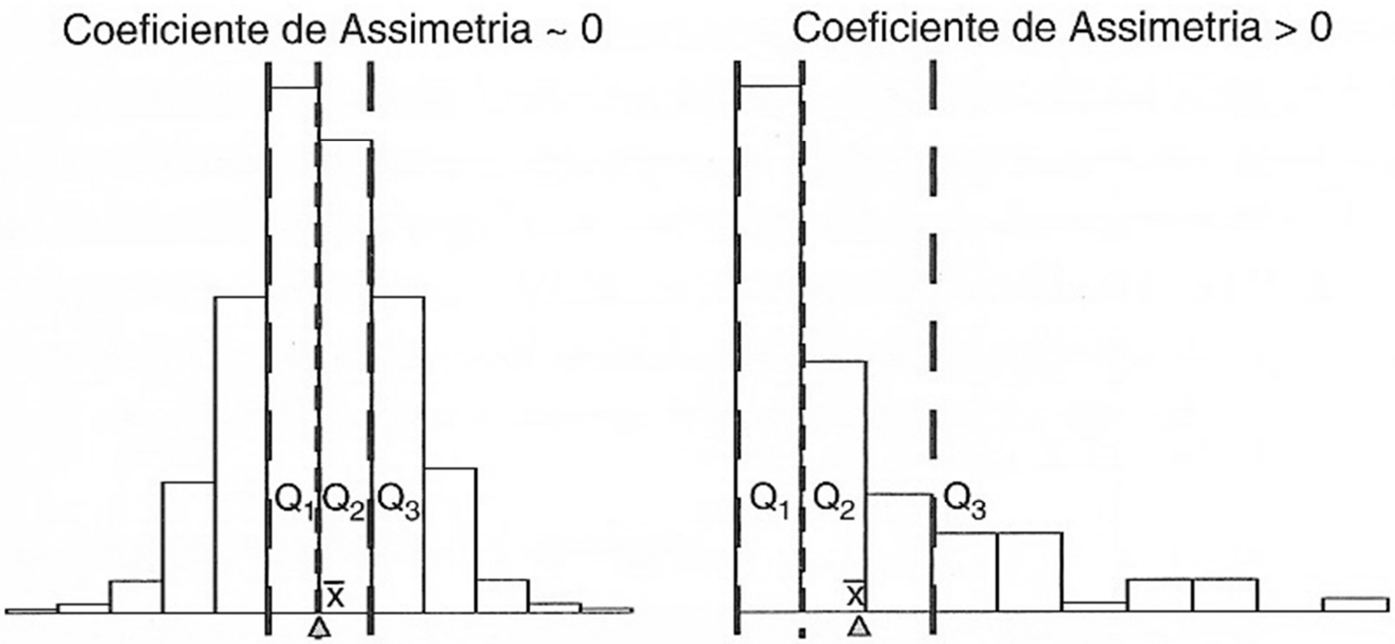


Figura 2.8 Diferenças entre as medidas de tendência central ou as distâncias entre quartis consecutivos que indicam assimetria.

Extraído de Diáz FR, López FJB. Bioestatística. São Paulo: Thomson Learning; 2007:

Medidas de achatamento (curtose)

- duas distribuições com mesma média e variabilidade, porém diferem nas alturas nas vizinhanças da média
- diferem quanto ao achatamento ou curtose

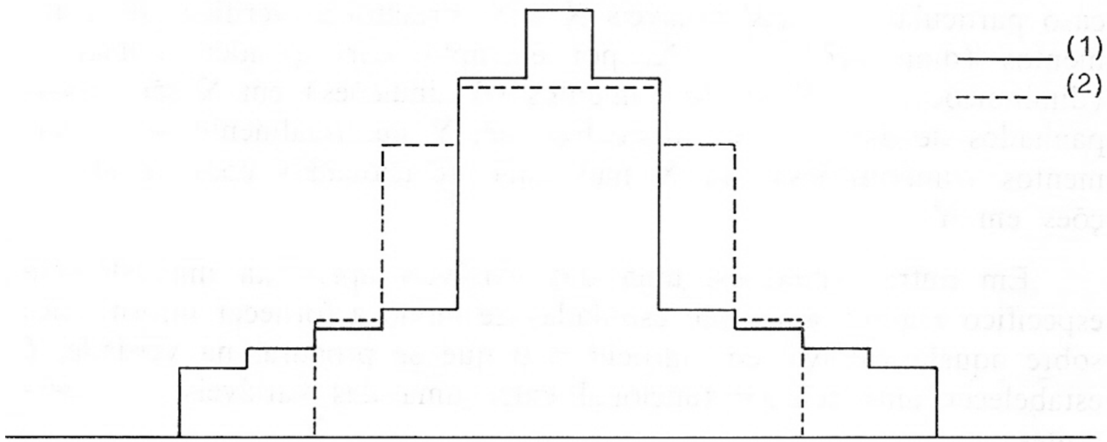


Figura 4.4 Distribuições de freqüências com médias e desvios padrão iguais.

- A soma das quartas potências dos desvios a partir da média em distribuições com o aspecto da distribuição (1) tende a ser maior do que aquela da distribuição (2)

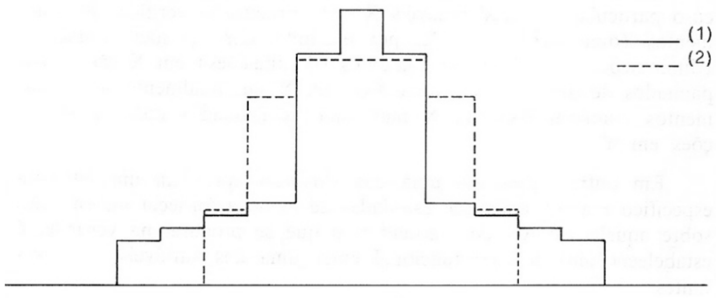


Figura 4.4 Distribuições de freqüências com médias e desvios padrão iguais.

$$g2 = \frac{\frac{1}{n} \sum_{i=1}^{k} (xi - \overline{x})^4 \ fi}{\left\{ \sqrt{\frac{1}{n} \sum_{i=1}^{k} (xi - \overline{x})^2 \ fi} \right\}^4}$$

- Leptocúrtica: distribuição menos achatada do que a normal, $g_2 > 3$
- Mesocúrtica: distribuição tão achatada quanto a normal, $g_2 = 3$
- Platicúrtica: distribuição mais achatada do que a normal, $g_2 < 3$

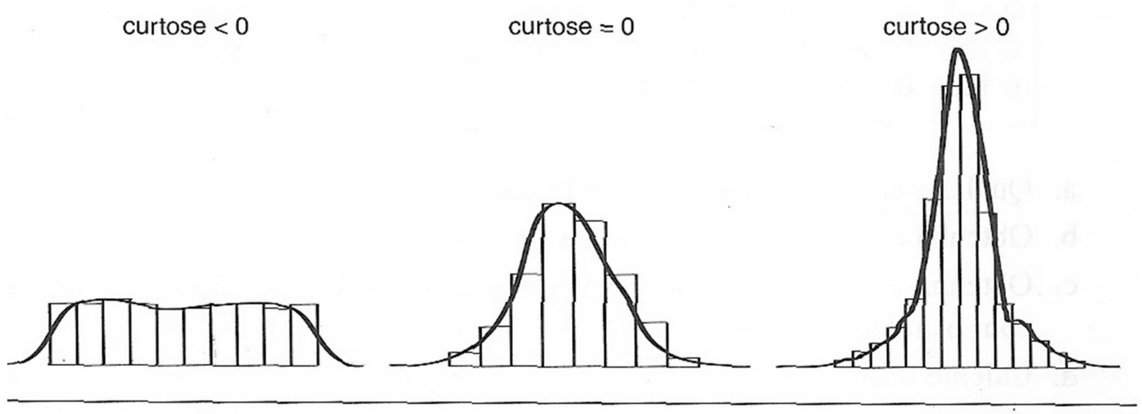


Figura 2.10 Achatamento de distribuições de freqüências.

A medida de curtose nos indica a forma da curva de distribuição em relação ao seu achatamento.

Forma da curva de distribuição:

- Leptocúrtica: distribuição apresenta uma curva de freqüência mais fechada que a normal (ou mais aguda em sua parte superior)
- Platicúrtica: distribuição apresenta uma curva de freqüência mais aberta que a normal (ou mais achatada na sua parte superior)
- Mesocúrtica: curva normal

Coeficiente percentílico de curtose:

$$C = \frac{Q3-Q1}{2\,(P_{90} - P_{10})}$$

Relativamente a curva normal, temos; C = 0,263

Assim:
C = 0,263 ⇒ curva mesocúrtica
C > 0,263 ⇒ curva leptocúrtica
C < 0,263 ⇒ curva platicúrtica

Exercício:

Tabela. Distribuição de idades de um grupo de pessoas, local X, Ano Y

Idade	n
7 |- 9	4
9 |- 11	18
11 |- 12	14
12 |- 13	27
13 |- 14	42
14 |- 15	31
15 |- 17	20
17 |- 19	1
Total	157

Xi	Ni	Xini	$(xi - x)^2 fi$
8	4	32,0	106,08
10	22	180,0	178,56
11,5	36	161,0	38,11
12,5	63	337,5	11,41
13,5	105	567,0	5,15
14,5	136	449,5	56,50
16	156	320,0	162,45
18	157	18,0	23,52
		2065	581,78

$\overline{X} = 13{,}15$ anos

$s^2 = 3{,}78$ anos^2

$CV = \frac{1{,}94}{13{,}15} = 0{,}15 = 15\%$

$s = 1{,}94$ anos

Assimetria de Yule – Bowley

$$Q_1 = 12 + \frac{39{,}25 - 36}{27} \cdot 1 = 12{,}12$$

$$Q_2 = 13 + \frac{78{,}5 - 63}{42} \cdot 1 = 13{,}37$$

$$Q_3 = 14 + \frac{117{,}75 - 105}{31} \cdot 1 = 14{,}41$$

$A_s = - 0{,}09$

$Sk = - 0{,}21$

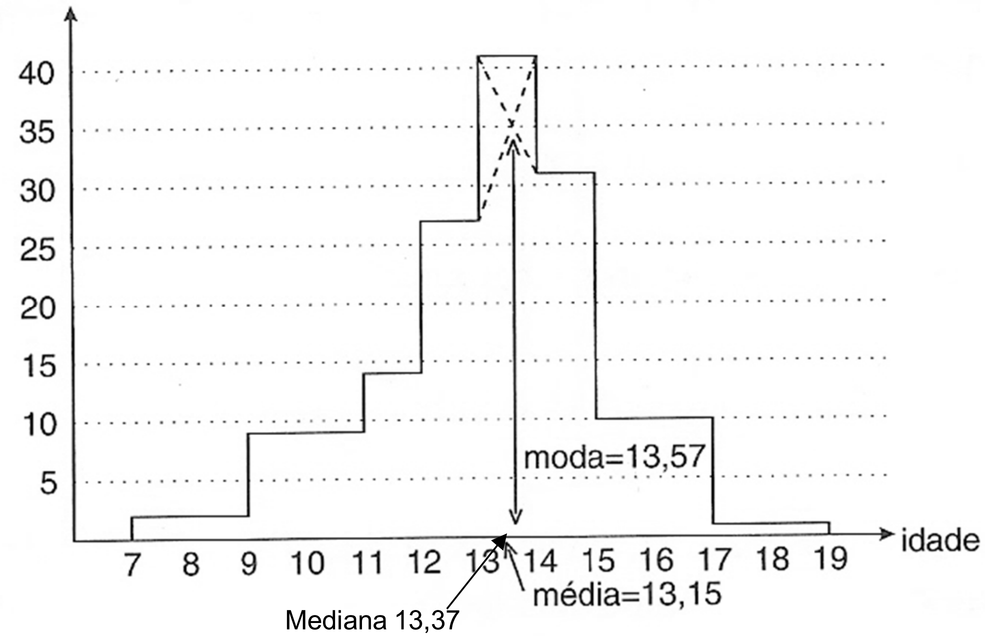


Figura 2.9 A distribuição de freqüências da idade apresenta uma leve assimetria negativa.