ESTATÍSTICA

Uma das funções da Estatística é prever, com a maior precisão possível, o comportamento de experimentos probabilísticos. Para termos uma noção mais precisa desse comportamento, são necessárias algumas medidas que representam certos aspectos do experimento.

Medidas de tendência central

Muitas vezes um único número é suficiente para nos dar ideia sobre um fenômeno probabilístico. Dados como a renda *per capita* de um país são exemplos disso, mesmo que muitas vezes sozinhos esses dados possam nos levar a conclusões falsas.

Média

A principal medida de tendência central é a *média aritmética*. A média de n dados é a sua soma dividida pela quantidade n de dados. Representamos a média pela letra grega μ (lê-se "mi"). Desse modo:

$$\mu = \frac{a_1 + a_2 + ... + a_n}{n}$$

Quando há dados repetidos, podemos agrupá-los: se a_1 aparece p_1 vezes, a_2 aparece p_2 vezes, ..., a_n aparece p_n vezes, a média é $\frac{p_1a_1 + p_2a_2 + ... + p_na_n}{p_1 + p_2 + ... + p_n}$.

Como $\mu = \frac{a_1 + a_2 + ... + a_n}{n} \Leftrightarrow a_1 + a_2 + ... + a_n = n \cdot \mu$, podemos afirmar que a média aritmética é o que substitui todos os valores de uma variável se ela fosse constante. Observe que a soma dos valores é sempre a mesma.

Mediana

A mediana é uma medida de tendência central na distribuição de dados, definida da seguinte maneira: ordenando-se os dados, chama-se mediana o valor, pertencente ou não ao conjunto de dados, que os divide ao meio, ou seja, metade dos valores são maiores ou iguais à mediana e a outra metade são valores menores ou iguais à mediana. Simbolicamente, se os dados são $a_1, a_2, ..., a_n$, a mediana, denotada por *Md*, é calculada primeiramente colocando-se, os dados em ordem (crescente ou decrescente) e, em seguida:

- se n é ímpar, tomamos o termo central;
- se n é par, há dois termos centrais: tomamos a média dos dois termos centrais.

Por exemplo, suponha que as notas de uma sala de 9 alunos em uma prova sejam 2, 4, 5, 6, 2, 4, 6, 8, 7. Ordenando as notas, obtemos 2, 2, 4, 4, *5*, 6, 6, 7, 8 e, portanto, a mediana das notas dessa sala é 5. Em uma outra sala, sendo as notas dos 8 alunos 3, 4, 6, 8, 6, 5, 7, 4, a mediana é $\frac{5+6}{2} = 5{,}5$ (ordenando, obtemos 3, 4, 4, *5*, *6*, 6, 7, 8).

Moda

A moda (denotada por *Mo*) é o dado que aparece mais vezes na população. Quando mais de um dado tem a mesma frequência máxima, o conjunto de dados tem mais de uma moda.

Quando dizemos que certa roupa está na moda, estamos dizendo que essa roupa está sendo usada com muita frequência pela população. Observe em sua sala se a calça *jeans* ainda continua na moda.

A moda é particularmente interessante em se tratando de vendas (qual variante de um produto vende mais?) e *marketing* (que característica do produto o define melhor?). Lembrem-se das pesquisas das marcas que mais são lembradas pelos consumidores. O entrevistador pergunta: "De qual marca você lembra quando alguém fala em refrigerante?". O consumidor pode responder: "Coca-Cola", "Guaraná Antártica", etc.

Trabalhando com dados agrupados

Os dados a seguir foram obtidos em uma pesquisa feita em 45 casas de um bairro da cidade de São Paulo e representam o número de livros existentes em cada casa.

118	247	260	540	396	316	180	275	105
105	129	59	151	404	290	310	120	240
10	72	86	25	580	200	270	240	420
83	513	410	120	420	338	215	478	286
680	327	210	350	360	142	375	236	400

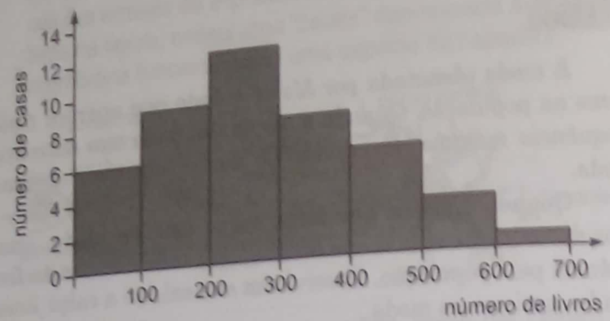
Queremos calcular aqui a média e a mediana de livros em cada casa. Para isso, temos duas opções: uma é calcular esses valores diretamente a partir dos dados anteriores, o que dá certo trabalho. Outra é construir uma tabela de frequências, o *histograma*, e obter *aproximações* desses valores, o que dá menos trabalho. Vejamos como fazer isso.

Dividamos os dados em 7 intervalos (ou classes) de tamanho 100. Obtemos a seguinte tabela de frequências:

Intervalo	Frequência	Ponto médio	Frequência relativa
0 ⊢ 100	6	50	13,3%
100 ⊢ 200	9	150	20%
200 ⊢ 300	12	250	26,7%
300 ⊢ 400	8	350	17,8%
400 ⊢ 500	6	450	13,3%
500 ⊢ 600	3	550	6,7%
600 ⊢ 700	1	650	2,2%

Vamos fazer as contas a partir desses dados. Note que, ao fazermos essa tabela, ganhamos em simplificação, mas perdemos alguns dados: agora só sabemos, por exemplo, que 6 casas têm *menos* que 100 livros, ao passo que antes sabíamos as quantidades de livros em cada casa.

Obtemos então o seguinte histograma:

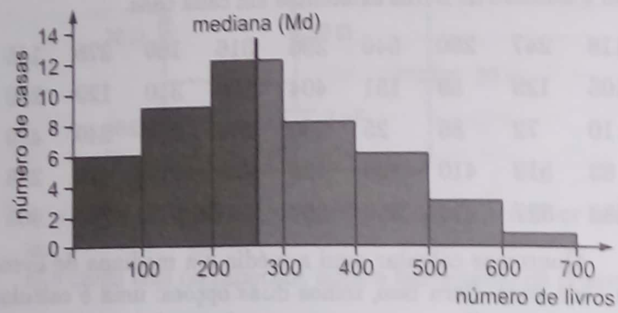


Utilizamos para a estimativa da média dessa amostra os pontos médios das classes:

$$\bar{x} \cong \frac{6 \cdot 50 + 9 \cdot 150 + 12 \cdot 250 + 8 \cdot 350 + 6 \cdot 450 + 3 \cdot 550 + 1 \cdot 650}{6 + 9 + 12 + 8 + 6 + 3 + 1}$$

$\cong$ 276,7 livros

Para o cálculo da mediana, podemos proceder da seguinte maneira: como metade dos valores são menores ou iguais à mediana, ela divide a área do histograma exatamente na metade:



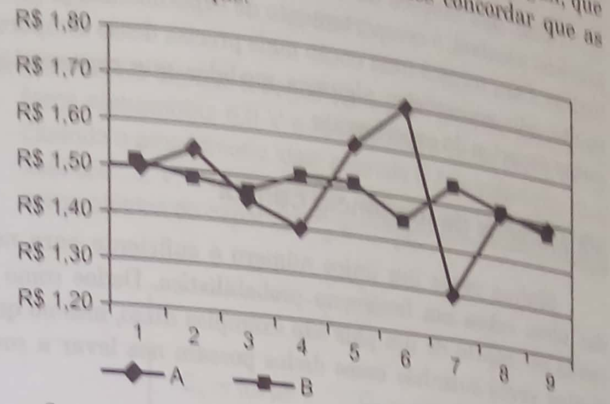
A área total do histograma é $100 \cdot (6 + 9 + 12 + 8 + 6 + 3 + 1) = 45 \cdot 100 = 4\,500$, ou seja, a mediana deve dividi-la em duas áreas de $\frac{4\,500}{2} = 2\,250$. A área das duas primeiras classes é $100 \cdot (6 + 9) = 1\,500$. Assim, sendo x a medida do lado horizontal do retângulo delimitado pelo limite inferior da classe e a mediana, sua área deve ser $2\,250 - 1\,500 = 750$, isto é, $x \cdot 12 = 750 \Leftrightarrow x = 62{,}5$. Assim, $Md = 200 + 62{,}5 = 262{,}5$.

A mediana ser menor do que a média implica o seguinte: mais de 50% das casas têm menos livros do que a média, ou seja, há uma concentração maior de livros com menos pessoas.

Medidas de dispersão

Considere os seguintes valores de duas carteiras de ações cotadas na Bolsa de Valores de São Paulo:

A média do valor de cada carteira é igual a R$ 1,50 e as medianas são respectivamente R$ 1,50 e R$ 1,51, que são valores muito próximos. Mas observando o gráfico a seguir, que representa os valores de ambas, devemos concordar que as carteiras têm diferenças:



O que torna essas duas sequências de dados tão diferentes? A carteira A oscila mais que a carteira B. Em termos financeiros, o valor da carteira A tem *variabilidade* maior e, considerando essa variabilidade, a carteira A tem risco maior. Então, se tivermos que escolher entre as duas carteiras, e quisermos correr menos riscos, devemos escolher a carteira B.

Como medir de modo mais preciso essa variabilidade? Para isso, utilizamos as *medidas de dispersão*.

O cálculo de todas as medidas de dispersão de uma sequência utiliza a seguinte sistemática:

I. Primeiro, tomamos uma medida de tendência central, como a média ou a mediana, para servir como referência.

II. Em seguida, calculamos a *distância* de todos os valores da sequência à média ou mediana.

III. A medida de dispersão é a média dessas distâncias ou de alguma potência dessas distâncias.

Note que quanto maior a distância média, maior a medida de dispersão. Assim, se esse número for muito grande, há grande variabilidade nos dados. Em contrapartida, se o número for pequeno, há pouca variabilidade.

As medidas de dispersão mais utilizadas são a *variância* e o *desvio-padrão*. Vamos defini-las seguindo os passos anteriores.

Variância

Tomamos como ponto de referência a média e os quadrados das distâncias. A variância é definida da seguinte maneira:

Seja x uma variável quantitativa que assume os valores $x_1, x_2, x_3, ..., x_n$, e μ a média correspondente a esses valores. A variância desses valores, denotada por σ^2, é dada por:

$$\boxed{\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_n - \mu)^2}{n}}$$

Note que cada termo do numerador corresponde ao quadrado da diferença entre o valor observado e o valor médio. Essa diferença nos mostra o quanto um valor observado se distancia da média.

Carteira	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7	Dia 8	Dia 9
A	R$1,50	R$1,55	R$1,45	R$1,40	R$1,60	R$1,70	R$1,30	R$1,51	R$1,49
B	R$1,51	R$1,49	R$1,47	R$1,52	R$1,52	R$1,45	R$1,55	R$1,51	R$1,48

No nosso exemplo, as variâncias são

$$\sigma_A^2 = \frac{(1{,}50-1{,}50)^2+(1{,}55-1{,}50)^2+(1{,}45-1{,}50)^2+(1{,}40-1{,}50)^2+(1{,}60-1{,}50)^2+(1{,}70-1{,}50)^2+(1{,}30-1{,}50)^2+(1{,}51-1{,}50)^2+(1{,}49-1{,}50)^2}{9},$$

ou seja, 0,01169, e

$$\sigma_B^2 = \frac{(1{,}51-1{,}50)^2+(1{,}49-1{,}50)^2+(1{,}47-1{,}50)^2+(1{,}52-1{,}50)^2+(1{,}52-1{,}50)^2+(1{,}45-1{,}50)^2+(1{,}55-1{,}50)^2+(1{,}51-1{,}50)^2+(1{,}48-1{,}50)^2}{9},$$

isto é, 0,0008.

Esses resultados nos mostram evidências de que a carteira B tem menos variabilidade que a carteira A.

Desvio-padrão

Qual seria a unidade da variância? Note que elevamos as diferenças entre os valores e a média ao quadrado e obtivemos uma espécie de média. Assim, no nosso exemplo anterior, a unidade seria "reais quadrados".

Para evitar esse engodo com unidade e, mais ainda, para podermos fazer comparações melhores com grandezas de mesma unidade, definimos o desvio-padrão como a raiz quadrada da variância. Assim:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_n-\mu)^2}{n}}$$

No nosso exemplo, o desvio-padrão de A é $\sqrt{0{,}01169} \cong 0{,}11$ real e o desvio-padrão de B é $\sqrt{0{,}0008} \cong 0{,}03$ real.

Veja que tanto a variância quanto o desvio-padrão nos dão evidências mais precisas de que a variabilidade de B é menor que a de A e, portanto, B acarreta menor risco com mesmo valor médio.

◆ Coeficiente de variação

Diferentemente das medidas anteriores, que acabamos de estudar, o coeficiente de variação representa uma medida relativa de variação. Ele é sempre expresso na forma de porcentagem, em vez de ser apresentado em termos das unidades de dados específicos.

O coeficiente de variação, representado pelo símbolo CV, mede a dispersão dos dados em relação à média aritmética e é dado por:

$$CV = \left(\frac{\sigma}{\mu}\right)\cdot 100\%$$

No exemplo das carteiras de ações A e B, temos:

$$CV_A = \left(\frac{0{,}11}{1{,}50}\right)\cdot 100\% = 7{,}3\% \text{ e } CV_B = \left(\frac{0{,}03}{1{,}50}\right)\cdot 100\% = 2\%$$

01. Carlos investiga o preço de certo artigo em 40 lojas de São Paulo e encontra as seguintes informações de preços (em reais):

82 60 60 74 83 74 82 74 82 82 82 60 60 74 82 74 82 74 78 78
60 64 74 74 74 62 74 67 74 82 60 68 74 74 74 74 74 68 68 82

a) Qual a moda dos valores anteriores?
b) Elabore uma tabela de distribuição de frequências.
c) Tomando-se os valores médios de cada classe na tabela, calcule o preço médio do artigo.

02. Dispomos de uma relação de preços de 200 aluguéis de imóveis urbanos na tabela a seguir.

Classes de aluguéis (em reais)	Número de imóveis
200 ⊢ 300	10
300 ⊢ 500	40
500 ⊢ 700	80
700 ⊢ 1 000	50
1 000 ⊢ 1 500	20
Total	200

a) Construa o histograma da distribuição calculando média e mediana.
b) Com base no histograma e nas medidas de tendência central calculadas, podemos afirmar que mais da metade dos aluguéis está com preço abaixo da média?

03. A distribuição de salários de uma empresa é dada pela tabela a seguir:

Salário (em R$)	Número de funcionários
500	10
1 000	5
1 500	1
2 000	10
5 000	4
10 500	1
Total	31

a) Quais são a média e a mediana dos salários dessa empresa?
b) Suponha que sejam contratados dois novos funcionários com salários de R$ 2.000,00 cada. A variância da nova distribuição de salários ficará menor, igual ou maior que a anterior?

04. Um conjunto de 10 valores numéricos $x_1, x_2, x_3, \ldots, x_{10}$ tem média aritmética igual a 100 e variância igual a 20. Se adicionarmos 5 a cada valor, isto é, se obtivermos o conjunto $(x_1+5), (x_2+5), (x_3+5), \ldots, (x_{10}+5)$, qual é:
a) a média do novo conjunto de valores?
b) a variância do novo conjunto de valores?

05. (FGV) As notas de um candidato em suas provas de um concurso foram: 8,4; 9,1; 7,2; 6,8; 8,7; e 7,2. A nota média, a nota mediana e a nota modal desse aluno são, respectivamente:

a) 7,9; 7,8 e 7,2. b) 7,2; 7,8 e 7,9. c) 7,8; 7,8 e 7,9.
d) 7,2; 7,8 e 7,9. e) 7,8; 7,9 e 7,2.

06. Numa pesquisa em determinada cidade foram obtidos os seguintes dados relativos ao número de crianças por família:

Número de crianças por família	Porcentagem de famílias na cidade
0	5
1	25
2	30
3	20
4	10
5 ou mais	10

O número médio de crianças nas famílias com 5 ou mais filhos é 5,8. O número médio de crianças por família nessa cidade é, então, igual a:

a) 2,00 b) 2,30 c) 2,43 d) 2,55 e) 3,00

07. Numa certa empresa, os funcionários trabalham 8 h/dia na segunda, 7 h/dia na terça, 10 h/dia na quarta, 11 h/dia na quinta e 4 h/dia na sexta.

a) Em média, quantas horas eles trabalham por dia de segunda a sexta?

b) Numa dada semana (de segunda a sexta) ocorrerá um feriado de 1 dia. Qual a probabilidade de eles trabalharem ao menos 30 horas nessa semana?

08. Ao procurar emprego, um determinado cidadão teve que optar por uma de duas ofertas dispostas em um classificado. Qual a oferta que representa a melhor opção? Por quê?

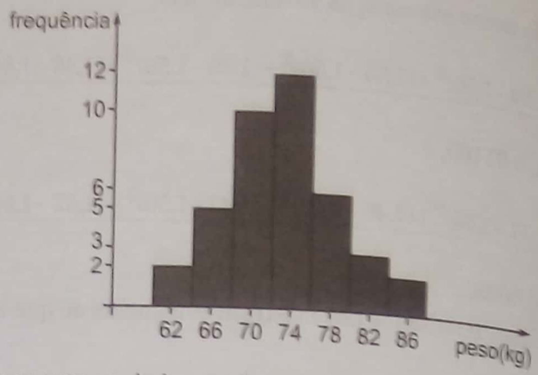
	Oferta 1	Oferta 2
Média salarial na empresa	R$ 890,00	R$ 1.050,00
Mediana	R$ 800,00	R$ 650,00
Desvio padrão	R$ 32,00	R$ 38,00

09. (UFU) Uma equipe de futebol realizou um levantamento dos pesos dos seus 40 atletas e chegou à distribuição de frequência dada pela tabela a seguir, cujo histograma correspondente é também visto em seguida:

Tabela

Peso kg	Frequência
60 ⊢ 64	2
64 ⊢ 68	5
68 ⊢ 72	10
72 ⊢ 76	12
76 ⊢ 80	6
80 ⊢ 84	3
84 ⊢ 88	2
Total de atletas	40

Histograma



Com base nesses dados, pode-se afirmar que o valor da mediana dos pesos é igual a:

a) 75 b) 72 c) 74 d) 73

10. A tabela a seguir mostra o número de cartões recebidos por um time num torneio de 10 jogos.

Número de cartões	Número de jogos
0	2
1	5
2	2
3	1

Calcule a variância referente ao número de cartões recebidos.

RESPOSTAS

01. a) A moda, valor que mais aparece, vale 74.

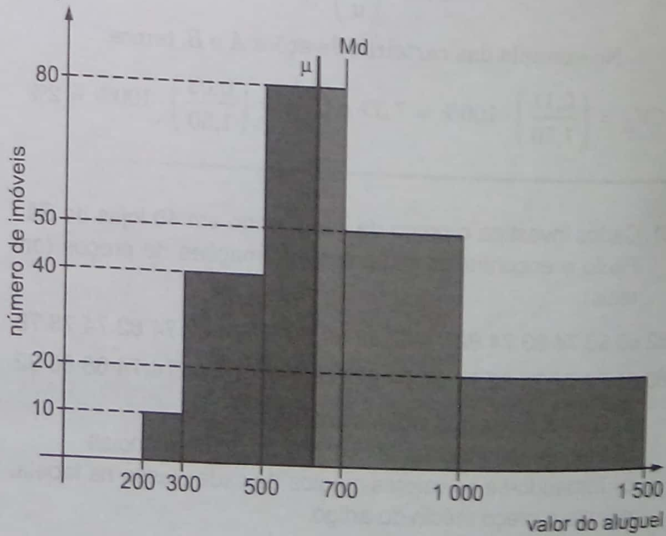
b)

Classes de preços (em reais)	Frequência
60 ⊢ 64	7
64 ⊢ 68	2
68 ⊢ 72	3
72 ⊢ 76	16
76 ⊢ 80	2
80 ⊢ 84	10
Total	40

c) O preço médio pode ser aproximado por

$$\frac{7 \cdot 62 + 2 \cdot 66 + 3 \cdot 70 + 16 \cdot 74 + 2 \cdot 78 + 10 \cdot 82}{40} = 73{,}4.$$

02. a)



Média da zona urbana:

$$\mu = \frac{250 \cdot 10 + 400 \cdot 40 + 600 \cdot 80 + 850 \cdot 50 + 1\,250 \cdot 20}{200} = 670$$

A mediana divide o histograma em duas regiões de mesma área. Como a área total do histograma é $10 \cdot 100 + 40 \cdot 200 + 80 \cdot 200 + 50 \cdot 300 + 20 \cdot 500 = 50\,000$, a área cinza escuro anterior deve ser igual a $\frac{50\,000}{2} = 25\,000$.

Note que a área das três primeiras classes é $100 \cdot 10 + 200 \cdot 40 + 200 \cdot 80 = 25\,000$. Logo a mediana é o limite superior da terceira classe, que vale 700.

b) Como a mediana é superior à média, mais da metade dos imóveis está com preço acima da média.

03. a) A média é igual a R$ 2.000,00 e a mediana igual a R$ 1.500,00.

b) Menor.

04. a) 105 b) 20

05. a

06. c

07. a) 8 h b) $\frac{4}{5}$

08. A opção 1 parece mais viável, pois apesar de a média salarial ser menor, podemos afirmar que metade dos salários é superior ou igual a R$ 800,00 e a variabilidade dos salários é menor.

09. d

10. Temos que:

- $\mu = \frac{0 \cdot 2 + 1 \cdot 5 + 2 \cdot 2 + 3 \cdot 1}{10} = \frac{12}{10} = 1{,}2$
- $\sigma^2 = \frac{2 \cdot (0 - 1{,}2)^2 + 5 \cdot (1 - 1{,}2)^2 + 2 \cdot (2 - 1{,}2)^2 + (3 - 1{,}2)^2}{10}$

$\Leftrightarrow \sigma^2 = \frac{2{,}88 + 0{,}2 + 1{,}28 + 3{,}24}{10}$

$\Leftrightarrow \sigma^2 = 0{,}76$