Medidas de Variação

Curso: Estatística e Probabilidade

Prof. Neemias Martins

PUC Campinas

neemias.silva@puc-campinas.edu.br

Notações

Na aula passada definimos a média aritmética $\overline{x}$. A partir de agora vamos convencionar o seguinte:

- Quando a média for calculada a partir de uma amostra de dados, usaremos a notação $\overline{x} = \frac{\sum x_i}{n}$, em que n é o número de elementos da amostra.
- Quando a média for calculada a partir da população inteira, usaremos a notação $\mu = \frac{\sum x_i}{N}$ em que N é o número de elementos da população.

Medidas de Variação

Medidas de variação ou dispersão

As medidas de tendência central não são suficientes para medir o grau de homogeneidade ou de heterogeneidade de um conjunto de dados. Para tal existem as *medidas de variação* ou de *dispersão*.

As medidas de variação nos permitem avaliar o quanto estão dispersos ou concentrados os valores de uma distribuição de frequência.

Exemplo

Um conjunto de dados contém as vendas diárias de uma empresa separadas por semanas. Os valores a seguir representam 3 amostras distintas.

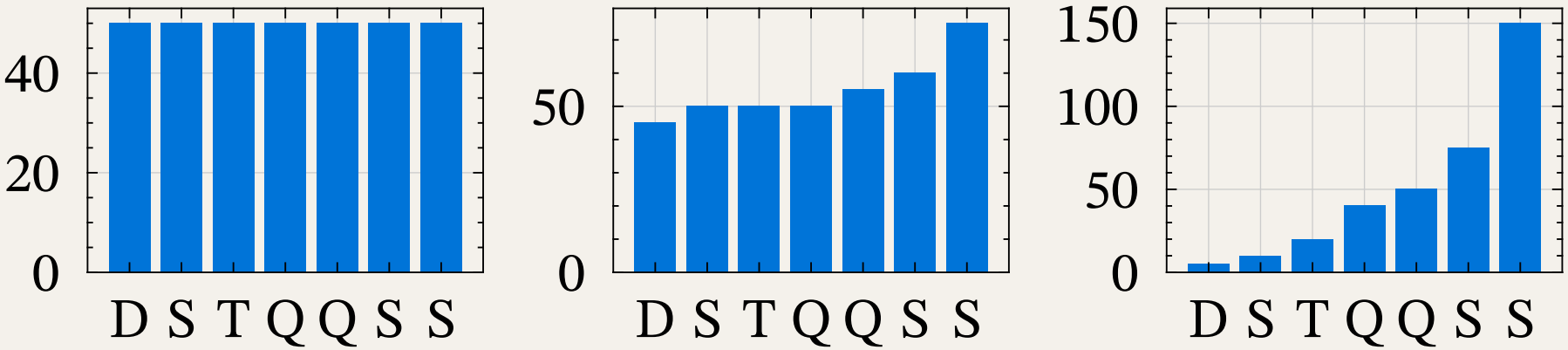
- $A = \{50, 50, 50, 50, 50, 50, 50\}$
- $B = \{45, 50, 50, 50, 55, 60, 75\}$
- $C = \{5, 10, 20, 40, 50, 75, 150\}$

As médias de vendas das 3 amostras são iguais:

$$\overline{a} = \frac{\sum_{i=1}^{7} a_i}{7} = 50, \quad \overline{b} = \frac{\sum_{i=1}^{7} b_i}{7} = 50, \quad \overline{c} = \frac{\sum_{i=1}^{7} c_i}{7} = 50.$$

Exemplo

No entanto, as vendas das amostras são bem distintas em termos de como os dados estão distribuídos:



- O conjunto A é mais homogêneo do que os conjuntos B e C. Como todos as entradas são iguais, sua dispersão é nula.
- O conjunto B possui menor variação entre cada valor e a média do que o conjunto C. A dispersão de B é menor do que a de C.

Medidas de variação ou dispersão

As principais medidas de variação são:

- Amplitude total
- Desvio médio absoluto
- Desvio padrão
- Variância
- Coeficiente de Variação

Amplitude

A *amplitude* é a mais simples e a menos importante medida de dispersão.

- Para dados não agrupados: A amplitude é a diferença entre o maior e o menor valor da série de dados.
- Para dados agrupados: a amplitude é a diferença entre o limite superior da última classe e o limite inferior da primeira classe.

Quanto maior a amplitude, maior a variação dos valores.

Exemplo - Amplitude de dados não agrupados

Considere o conjunto de dados $S = \{2, 5, 8, 14, 20, 55, 122\}$. Então

$$\text{amplitude} = x_{\max} - x_{\min} = 122 - 2 = 120.$$

Exemplo - Amplitude de dados agrupados

Considere a tabela de frequências

Classes	x_i	f_i
$100 \vdash 150$	125	87
$150 \vdash 200$	175	45
$200 \vdash 250$	225	23
$250 \vdash 300$	275	15

Então,

$$\text{amplitude} = 300 - 100 = 200$$

Desvio médio absoluto

O *desvio médio absoluto* é a média aritmética dos valores absolutos dos desvios em relação à média.

$$\text{desvio médio absoluto} = \frac{\sum_{i=1}^{n} |d_i|}{n} = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

Exemplo - Desvio médio absoluto

Se a amostra de dados é $S = \{5, 10, 30\}$, então $\overline{x} = 15$, e os desvios em relação à média são dados por

$$d_1 = 5 - 15 = -10, \quad d_2 = 10 - 15 = -5, \quad d_3 = 30 - 15 = 15.$$

Então,

$$\begin{aligned} \text{desvio médio absoluto} &= \frac{\sum_{i=1}^{3} |d_i|}{3} \\ &= \frac{|-10| + |-5| + |15|}{3} \\ &= \frac{10 + 5 + 15}{3} \\ &= 10. \end{aligned}$$

Observação

Os desvios em relação à media $d_i = x_i - \overline{x}$ possuem a propriedade de que sua soma é nula: $\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = 0$.

Para usar os desvios como medida de variação, é então natural calcular a média usando os valores absolutos $|x_i - \overline{x}|$:

$$\text{desvio médio absoluto} = \frac{\sum_{i=1}^{n}|x_i - \overline{x}|}{n} \geq 0$$

O desvio médio absoluto usa valores absolutos, que é uma operação não "algébrica" e leva a algumas complicações em estatística inferencial. (São operações algébricas: adição, subtração, multiplicação, potenciação...). Por isto, outro método de avaliar os desvios é mais recomendado: **o desvio padrão**.

Desvio padrão

O *desvio padrão amostral* é a medida de variação mais usada em estatística.

O desvio padrão amostral de um conjunto de dados, denotado por s, mede algebricamente o quanto os valores se desviam da média:

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n-1}}.$$

Observações

- É comum expressar o desvio padrão com uma casa decimal a mais do que os dados iniciais.
- Se somarmos uma constante a todos os valores de uma variável, o desvio padrão não muda.
- Se multiplicarmos todos os valores de uma variável por uma constante não-nula, o desvio padrão será multiplicado por essa constante.

Observações

- O valor do desvio padrão s nunca é negativo. Ele é zero somente quando todos os valores de dados são exatamente os mesmos.
- Valores maiores de s indicam maiores quantidades de variação.
- Os desvios padrão podem aumentar drasticamente com um ou mais valores atípicos.
- As unidades do desvio padrão s (min, km, kg, ...) são as mesmas que as unidades dos valores de dados originais

Exemplo

Considere a amostra $S = \{10, 20, 30, 40\}$. Calcularemos o desvio padrão.

x_i	$\overline{x}$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
10	25	-15	225
20	25	-5	25
30	25	5	25
40	25	15	225
Total			**500**

Então, $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{500}{3}} = \sqrt{166.667} = 12.909$. Portanto o desvio padrão é de 12.9

Regra prática para interpretar o desvio padrão

Para uma amostra grande, a seguinte regra prática pode ser usada para interpretar valores:

- Significativamente baixo: valores menores ou iguais a $\overline{x} - 2s$
- Significativamente alto: valores maiores ou iguais a $\overline{x} + 2s$
- Não significativos: Valores entre $\overline{x} - 2s$ e $\overline{x} + 2s$.

Exemplo

Uma amostra contém dados de 150 medições de temperatura de um determinado ambiente. A temperatura média registrada foi de 22ºC. O desvio padrão foi de 3.2°C. Temos:

- $\overline{x} - 2 \cdot s = 22 - 2 \cdot 3.2 = 15.6$
- $\overline{x} + 2 \cdot 2s = 22 + 2 \cdot 3.2 = 28.4$

Conclui-se:

- Temperaturas normais (ou não significativas): entre 15.6ºC e 28.4ºC.
- Temperaturas significativamente baixas: 15.6ºC ou menos
- Temperaturas significativamente altas: 28.4ºC ou mais.

Desvio padrão populacional

O desvio padrão populacional, calculado sobre toda a população, denotado por σ, é ligeiramente diferente do desvio padrão amostral s. Ele é definido por

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Variância

A *variância* de um conjunto de valores é uma medida de variação igual ao quadrado do desvio padrão:

- variância amostral: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$
- variância populacional: $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$

Observações

- As unidades da variância são os quadrados das unidades dos valores dos dados originais. (Se os valores dos dados originais estiverem em minutos, a variância terá unidades em min^2).
- O valor da variância pode aumentar drasticamente com a inclusão de valores atípicos. (A variância não é uma medida de variação resistente).
- O valor da variância nunca é negativo. É zero somente quando todos os valores dos dados são iguais.

Exemplo

Considere a amostra $S = \{10, 20, 30, 40\}$. Calcularemos a variância.

x_i	$\overline{x}$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
10	25	-15	225
20	25	-5	25
30	25	5	25
40	25	15	225
Total			**500**

Então, $s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \frac{500}{3} = 166.667$. Portanto a variância é de 166.7

Exemplo

A tabela a seguir apresenta as medidas de tendência central e as medidas de variação de dois conjuntos de dados.

	Média	Mediana	Moda	Desvio Padrão	Variância
A	35.2	33	36	12.0	144.0
B	35.2	33	36	30.0	900.0

Os dados A e B possuem as mesmas medidas de tendência central, mas as medidas de variação são drasticamente distintas.

Seria um grave erro comparar os dados analisando apenas as medidas de tendência central.

Observação

- É usual se comparar dois desvios-padrão de amostras apenas quando as médias das amostras são aproximadamente as mesmas.
- Ao comparar a variação em amostras ou populações com médias muito diferentes, é melhor usar o **coeficiente de variação.**
- Use também o coeficiente de variação para comparar a variação de duas amostras ou populações com diferentes escalas ou unidades de valores, como a comparação da variação de alturas de homens adultos e pesos de homens adultos.

Coeficiente de variação

O *coeficiente de variação* (CV) de uma amostra ou população é expressa como um percentual e descreve o desvio padrão relativo à media:

- Coeficiente de variação amostral: $\text{CV} = \frac{S}{\overline{x}} \cdot 100$
- Coeficiente de variação populacional: $\text{CV} = \frac{\sigma}{\mu} \cdot 100$

Observação: Use uma casa decimal para arrendondar o coeficiente de variação. (Exemplo: 14.5%)

Exemplo

Faturamento mensal de duas empresas de tecnologia em milhões de reais.

	Média	Mediana	Moda	Desvio Padrão	Variância
TI Soluções	5	7	4	3.0	9.0
ND Soluções	15	9	10	2.0	4.0

Como as médias são muito distintas, vamos comparar a variação dos dados através do coeficiente de variação das amostras.

Exemplo

- T.I. Soluções:

$$\text{CV} = \frac{s}{\overline{x}} \cdot 100 = \frac{3.0}{5} \cdot 100 = 0.6 \cdot 100 = 60.0\%$$

- N.D. Soluções:

$$\text{CV} = \frac{s}{\overline{x}} \cdot 100 = \frac{2.0}{15} \cdot 100 = 0.1333 \cdot 100 = 13.3\%$$

O faturamento mensal da T.I. Soluções varia consideravelmente mais do que o faturamento de N.D. Soluções.

Exercícios

Exercícios

1. A tabela a seguir compara o faturamento diário e o lucro mensal de um time de vendas em mil reais.

	Média	Mediana	Moda	Desvio Padrão	Variância
Fat. diário	20	8	13	7.0	49.0
Lucro mensal	350	100	120	100	10000.0

Compare a variação dos dados das duas amostras.

Exercícios

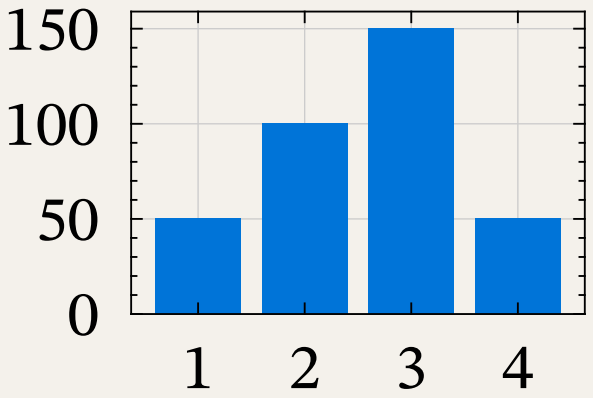
2. A partir da tabela de frequências, calcule:

a) média aritmética
b) moda
c) variância
d) desvio padrão

Classes	f_i
$0 \vdash 20$	1
$20 \vdash 40$	4
$40 \vdash 60$	2
$60 \vdash 80$	3

Exercícios

A partir do histograma, obtenha tabela de frequências calcule a média, moda, desvio padrão e variância.



Bons estudos!