

Medidas resumo

Luan D. Fiorentin

Universidade Federal do Paraná
Departamento de Estatística
Laboratório de Estatística e Geoinformação

02/03/2020



Sumário

1 Introdução

2 Medidas de posição

- Medidas de posição para um conjunto de dados.
- Medidas de posição para VAs discretas

3 Medidas de dispersão

- Medidas de dispersão para um conjunto de dados
- Medidas de dispersão para VAs discretas

4 Exercícios recomendados

Introdução

- Características importantes de qualquer conjunto de dados ou de uma variável aleatória:
 - Centro;
 - Variação;
 - Distribuição;
 - Valores atípicos.

Introdução

- Características importantes de qualquer conjunto de dados ou de uma variável aleatória:
 - Centro;
 - Variação;
 - Distribuição;
 - Valores atípicos.
- Classificaremos as medidas descritivas em dois grupos:
 - Medidas de posição;
 - Medidas de dispersão.

Sumário

1 Introdução

2 Medidas de posição

- Medidas de posição para um conjunto de dados.
- Medidas de posição para VAs discretas

3 Medidas de dispersão

- Medidas de dispersão para um conjunto de dados
- Medidas de dispersão para VAs discretas

4 Exercícios recomendados

Definição

- Medidas de posição central:
 - Úteis para **resumo** e **análise** de dados;
 - As principais medidas são a média, mediana e moda.
- Outras medidas de posição:
 - Extremos: mínimo e máximo;
 - Quantis: 1º quartil, 3º quartil, percentil 5%, entre outras, ...

Moda

- Valor **mais frequente** em um conjunto de dados;
- Dependendo do conjunto de dados, ele pode ser
 - **Sem moda** quando nenhum valor se repete;
 - **Unimodal** quando existe apenas um valor repetido com maior frequência;
 - **Bimodal** quando existem dois valores com a mesma maior frequência;
 - **Multimodal** quando mais de dois valores se repetem com a mesma frequência.

Moda

- Valor **mais frequente** em um conjunto de dados;
- Dependendo do conjunto de dados, ele pode ser
 - **Sem moda** quando nenhum valor se repete;
 - **Unimodal** quando existe apenas um valor repetido com maior frequência;
 - **Bimodal** quando existem dois valores com a mesma maior frequência;
 - **Multimodal** quando mais de dois valores se repetem com a mesma frequência.
- Valor com **maior probabilidade** de ocorrer numa **VA discreta**.
 - Exemplo: lançamento de duas moedas:
 - X : número de caras, $X = \{0, 1, 2\}$;
 - $P(x) = 0.25, 0.5$ e 0.25 , respectivamente;
 - Moda: 1.

Mediana

- O **valor do meio** da amostra **ordenada**;
- Separa o conjunto de dados em duas partes iguais, 50% abaixo e 50% acima;
- Considerando as observações ordenadas, denotamos:
 - A menor observação por $x_{(1)}$, a segunda por $x_{(2)}$, e assim por diante:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$$

- As observações ordenadas são chamadas de **estatísticas de ordem**:
 - $x_{(1)}$ é o mínimo da amostra;
 - $x_{(n)}$ é o máximo da amostra.

Média de dados brutos

- Divide-se a soma de todos os dados pelo número total deles

$$\bar{x}_{obs} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}.$$

Média de dados agrupados

- Soma dos produtos dos valores pelas respectivas frequências e divide pela frequência total:

$$\bar{x}_{obs} = \frac{n_1x_1 + n_2x_2 + \cdots + n_kx_k}{n_1 + n_2 + \cdots + n_k} = \frac{\sum_{i=1}^k n_i x_i}{n}.$$

Exemplo: média de dados discretos agrupados

- Considere a tabela de frequência abaixo:

Número	n_i	f_i
0	4	0,20
1	5	0,25
2	7	0,35
3	3	0,15
5	1	0,05
Total	20	1

A média é calculada por:

$$\begin{aligned}\bar{x}_{obs} &= \frac{0 \cdot 4 + 1 \cdot 5 + 2 \cdot 7 + 3 \cdot 3 + 5 \cdot 1}{4 + 5 + 7 + 3 + 1} = \frac{33}{20} \\ &= 1,65.\end{aligned}$$

Exemplo: média de dados agrupados em classes

- Usar **ponto médio** de cada classe e respectivas frequências:

Classe	PM = x_i	n_i	f_i
[4, 8)	6	10	0,278
[8, 12)	10	12	0,333
[12, 16)	14	8	0,222
[16, 20)	18	5	0,139
[20, 24)	22	1	0,028
Total	36	1	

Considerando os **pontos médios** de cada classe, a média é calculada por

$$\begin{aligned}\bar{x}_{obs} &= \frac{(6 \cdot 10 + 10 \cdot 12 + \dots + 22 \cdot 1)}{10 + 12 + 8 + 5 + 1} = \frac{404}{36} \\ &= 11,22.\end{aligned}$$

Exemplo 4.1 do livro

Suponha que parafusos a serem utilizados em tomadas elétricas são embalados em caixas rotuladas como contendo 100 unidades. Em uma construção, 10 caixas de um lote tiveram o número de parafusos contados, fornecendo os valores:

$$amostra = (98, 102, 100, 100, 99, 97, 96, 95, 99, 100)$$

- Calcular média, mediana e moda:
 - $\bar{x}_{obs} = 98.6$;
 - $md_{obs} = 99$;
 - $mo_{obs} = 100$.

Média e mediana

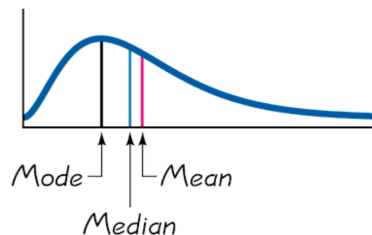
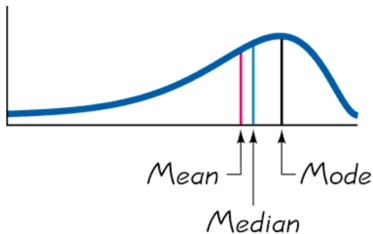
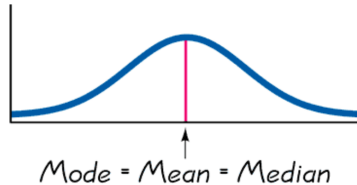
- Notar a influência de valores extremos na média (se ao invés de 95, o valor fosse 45):

$$95 \ 96 \ 97 \ 98 \ 99 \ 99 \ 100 \ 100 \ 100 \ 102 \Rightarrow \bar{x}_{obs} = 98,6 \text{ e } Md = 99$$

$$45 \ 96 \ 97 \ 98 \ 99 \ 99 \ 100 \ 100 \ 100 \ 102 \Rightarrow \bar{x}_{obs} = 93,6 \text{ e } Md = 99$$

- Devido a esse fato, a mediana é uma medida de posição central **robusta**, ou seja, *não é influenciada por valores extremos*.

Média, mediana e moda



Exemplo 4.3 do livro

Foram coletadas 150 observações da variável X , representando o número de vestibulares FUVEST (um por ano) que um mesmo estudante prestou. Com os dados da tabela abaixo, calcule as medidas de posição de X .

X	n_i
1	75
2	47
3	21
4	7

Ainda, suponha que o interesse é estudar o gasto dos alunos associado com as despesas do vestibular. Para simplificar, suponha que se atribui, para cada aluno, uma despesa fixa de R\$ 1300,00 relativa a preparação e mais R\$ 50 para cada vestibular prestado. Calcule as medidas de posição central para a variável D (despesa com vestibular).

Exemplo 4.4 do livro

Um estudante está procurando um estágio para o próximo ano. As companhias A e B têm programas de estágios e oferecem uma remuneração por 20 horas semanais com as seguintes características:

Companhia	A	B
média	2,5	2,0
mediana	1,7	1,9
moda	1,5	1,9

- Qual companhia você escolheria?

Medidas de posição para VAs discretas

- Sabemos que a descrição completa do comportamento de uma VA discreta é feita através de sua **função de probabilidade**;
- Assim como fizemos para um conjunto de dados qualquer, podemos obter as medidas de posição para qualquer variável aleatória;
- Considerando que os possíveis valores de uma VA X são x_1, x_2, \dots, x_k , com correspondentes probabilidades p_1, p_2, \dots, p_k , então as medidas de posição podem ser definidas a seguir.

Medidas de posição para VAs discretas:

- A Média é chamada de **valor esperado** ou **esperança**:

$$E(X) = \sum_{i=1}^k x_i p_i;$$

- A Mediana é o valor Md que satisfaz as seguintes condições:

$$P(X \leq Md) \geq 1/2 \quad \text{e} \quad P(X \geq Md) \geq 1/2;$$

- A Moda é o valor (ou valores) com maior probabilidade de ocorrência:

$$P(X = Mo) = \max\{p_1, p_2, \dots, p_k\}.$$

Exemplo 4.5 do livro

Considere a VA X com a seguinte função discreta de probabilidade:

X	-5	10	15	20
p_i	0.3	0.2	0.4	0.1

- Calcule as medidas de tendência central.

Exemplo 4.6 do livro

Considere uma VA X com função de probabilidade dada por

X	2	5	8	15	20
p_i	0.1	0.3	0.2	0.2	0.2

- Calcule as medidas de posição para a VA Y , em que $Y = 5X - 10$.

Sumário

1 Introdução

2 Medidas de posição

- Medidas de posição para um conjunto de dados.
- Medidas de posição para VAs discretas

3 Medidas de dispersão

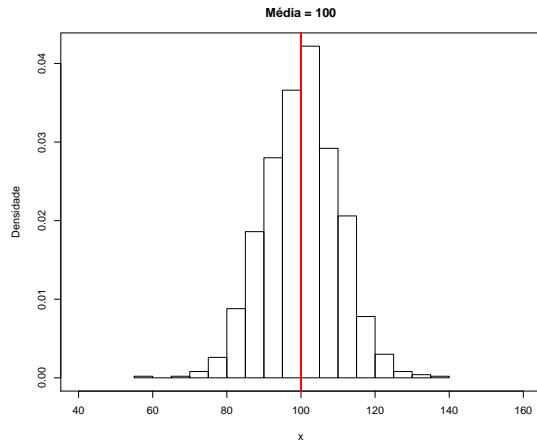
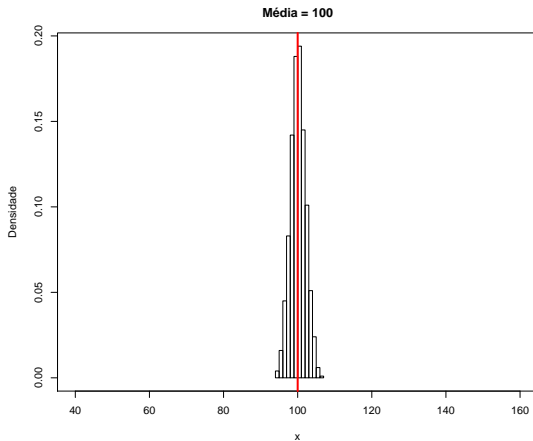
- Medidas de dispersão para um conjunto de dados
- Medidas de dispersão para VAs discretas

4 Exercícios recomendados

Introdução

- O resumo de um conjunto de dados exclusivamente por uma medida de centro, **esconde** toda a informação sobre a variabilidade do conjunto de observações;
- Não é possível analisar um conjunto de dados apenas através de uma medida de tendência central;
- Por isso precisamos de medidas que resumam a **variabilidade** dos dados em relação à um valor central.

Exemplo: mesma média, diferente dispersão



Exemplo 3.2 do livro do Bussab e Morettin

Cinco grupos de alunos se submeteram a um teste, obtendo as seguintes notas

Grupo	Notas	\bar{x}
A	3, 4, 5, 6, 7	5
B	1, 3, 5, 7, 9	5
C	5, 5, 5, 5, 5	5
D	3, 5, 5, 7	5
E	3, 5, 5, 6, 6	5

- O que a média diz a respeito das notas quando comparamos os grupos?

Definição

- São medidas estatísticas que caracterizam o quanto um conjunto de dados está disperso em torno de sua tendência central;
- Ferramentas para **resumo** e **análise** de dados:
 - Amplitude;
 - Desvio-médio (ou mediano);
 - Variância;
 - Desvio-padrão;
 - Coeficiente de Variação.

Amplitude

- A **amplitude** de um conjunto de dados é a diferença entre o maior e o menor valor:

$$\Delta = \max - \min = x_{(n)} - x_{(1)}$$

Grupo	Notas	Δ
A	3, 4, 5, 6, 7	4
B	1, 3, 5, 7, 9	8
C	5, 5, 5, 5, 5	0
D	3, 5, 5, 7	4
E	3, 5, 5, 6, 6	3

- **Apenas** usar máximo e mínimo torna a medida **sensível** a valores extremos.
 - Melhor medida de variabilidade: considerar **todos os dados disponíveis**;
 - **Desvio** de cada valor em relação à uma medida de posição central (média ou mediana).

Desvio médio e mediano

- Um **resumo** da variabilidade: **média** dos desvios **absolutos**;
- **Desvio mediano**: a **mediana** como medida de posição central:

$$\text{desvio mediano} = \frac{1}{n} \sum_{i=1}^n |x_i - md_{obs}|.$$

- **Desvio médio**: a **média** como medida de posição central:

$$\text{desvio médio} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}_{obs}|.$$

Exemplo: Desvio médio

Considere as notas do grupo A do exemplo anterior ($\bar{x}_{obs} = 5$).

O desvio médio (DM) pode ser calculado da seguinte forma:

Grupo A	$x_i - \bar{x}$	$ x_i - \bar{x} $
3	-2	2
4	-1	1
5	0	0
6	1	1
7	2	2
Soma	0	6

$$DM = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}_{obs}| = \frac{6}{5} = 1,2.$$

- O desvio médio é baseado em uma operação **não algébrica** (módulo), o que torna mais difícil o estudo de suas propriedades.

Variância de um conjunto de dados

- Uma alternativa melhor é usar a **soma dos quadrados dos desvios**, que dá origem à **variância** de um conjunto de dados:

$$var_{obs} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_{obs})^2.$$

- Uma expressão alternativa da variância (mais fácil de calcular) é

$$var_{obs} = \hat{\sigma}^2 = s^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}_{obs}^2.$$

Desvio-padrão de um conjunto de dados

- Para manter a mesma unidade de medida dos dados originais, definimos o **desvio padrão** como:

$$dp_{obs} = \hat{\sigma} = s = \sqrt{var_{obs}};$$

- O desvio padrão é mais interpretável em um primeiro momento porque é dado na mesma unidade de medida dos dados originais.

Exemplo

No exemplo anterior:

Grupo A	$x_i - \bar{x}$	$ x_i - \bar{x} $	$(x_i - \bar{x})^2$	x_i^2
3	-2	2	4	9
4	-1	1	1	16
5	0	0	0	25
6	1	1	1	36
7	2	2	4	49
Soma	0	6	10	135

- A variância é

$$var_{obs} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_{obs})^2 = \frac{10}{5} = 2.$$

Usando a fórmula alternativa, temos que

$$var_{obs} = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}_{obs}^2 = \frac{135}{5} - 5^2 = 2.$$

Coeficiente de variação

- O **coeficiente de variação** para um conjunto de dados é definido por:

$$cv_{obs} = \frac{dp_{obs}}{\bar{x}_{obs}};$$

- É uma medida **adimensional**, e geralmente apresentada na forma de porcentagem como:

$$cv_{obs} = \frac{dp_{obs}}{\bar{x}_{obs}} \cdot 100.$$

Exemplo

No exemplo anterior, temos que $dp_{obs} = \sqrt{var_{obs}} = \sqrt{2} = 1,414214$.

- O desvio padrão é:

$$cv_{obs} = \frac{dp_{obs}}{\bar{x}_{obs}} = \frac{1,414214}{5} = 0,2828427 \approx 28,3\%.$$

Variância em tabelas de frequência

- Assim como no caso da média, se tivermos n observações da variável X , das quais n_1 são iguais a x_1 , n_2 são iguais a x_2 , \dots , n_k são iguais a x_k , então a variância pode ser definida por

$$\text{var}_{obs}(X) = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x}_{obs})^2;$$

- Pela fórmula alternativa, temos que

$$\text{var}_{obs}(X) = \frac{1}{n} \sum_{i=1}^k n_i x_i^2 - \bar{x}_{obs}^2.$$

Exemplo

Como exemplo, considere a tabela de frequência abaixo, com média $\bar{x} = 1,65$:

Número	n_i	f_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0	4	0,20	-1,65	2,72
1	5	0,25	-0,65	0,42
2	7	0,35	0,35	0,12
3	3	0,15	1,35	1,82
5	1	0,05	3,35	11,22
Total	20	1		

- A variância é calculada por:

$$\begin{aligned}
 var_{obs} &= \frac{(4 \cdot 2,72 + 5 \cdot 0,42 + \dots + 1 \cdot 11,22)}{4 + 5 + 7 + 3 + 1} = \frac{30,55}{20} \\
 &= 1,528.
 \end{aligned}$$

Exemplo

Considere a seguinte tabela de distribuição de frequência, com média $\bar{x} = 11,22$:

Classe	PM = x_i	n_i	f_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
[4, 8)	6	10	0,278	-5,222	27,272
[8, 12)	10	12	0,333	-1,222	1,494
[12, 16)	14	8	0,222	2,778	7,716
[16, 20)	18	5	0,139	6,778	45,938
[20, 24)	22	1	0,028	10,778	116,160
Total		36	1		

Considerando os **pontos médios** de cada classe como os valores x_i , a variância pode ser calculada por

$$\begin{aligned}
 var_{obs} &= \frac{(10 \cdot 27,272 + 12 \cdot 1,494 + \dots + 1 \cdot 116,160)}{10 + 12 + 8 + 5 + 1} = \frac{698,22}{36} \\
 &= 19,395.
 \end{aligned}$$

Exemplo 4.9 do livro

No Exemplo 4.3, definimos a quantidade D , despesa no vestibular, obtida a partir de X pela expressão $D = 50X + 1300$, com X indicando o número de vestibulares prestados.

X	n_i
1	75
2	47
3	21
4	7

- Calcule a variância de D ;
- Calcule a variância do Exemplo 4.10 do livro.

Variância de uma VA discreta

- Calcular o valor esperado:

$$\mu = E(X) = \sum_{i=1}^k x_i p_i;$$

- Multiplicar o quadrado dos desvios em torno do valor esperado pela probabilidade e somar

$$Var(X) = \sum_{i=1}^k (x_i - \mu)^2 p_i;$$

- Alternativamente, podemos usar

$$Var(X) = E[(X - \mu)^2] = E(X^2) - E(X)^2$$

com $E(X^2) = \sum_{i=1}^k x_i^2 p_i$.

Exemplo 4.11 do livro

Uma pequena cirurgia dentária pode ser realizada por três métodos diferentes cujos tempos de recuperação (em dias) são modelados pelas variáveis X_1 , X_2 e X_3 . Admita suas funções de probabilidades dadas por

X_1	0	4	5	6	10
p_i	0.2	0.2	0.2	0.2	0.2

X_2	1	5	9
p_i	1/3	1/3	1/3

X_3	4	5	6
p_i	0.3	0.4	0.3

- Calcule as medidas de posição central e dispersão para cada VA e decida sobre o método mais eficiente.

Propriedades da média e da variância

Conjunto de Dados	Variável Aleatória
$Y = a \cdot X + b$	$Y = a \cdot X + b$
$\bar{y} = a \cdot \bar{x} + b$	$E[Y] = a \cdot E[X] + b$
$s^2(Y) = a^2 \cdot s^2(X)$	$s^2(Y) = a^2 \cdot s^2(X)$

Sumário

1 Introdução

2 Medidas de posição

- Medidas de posição para um conjunto de dados.
- Medidas de posição para VAs discretas

3 Medidas de dispersão

- Medidas de dispersão para um conjunto de dados
- Medidas de dispersão para VAs discretas

4 Exercícios recomendados

Exercícios recomendados

- Seção 4.2: Ex. 1, 2, 3, 4 e 6.
- Seção 4.3: Ex. 1, 2, 3, 4 e 5.
- Extras: Seção 4.4: Ex. 1, 2, 4, 7, 10 e 19.