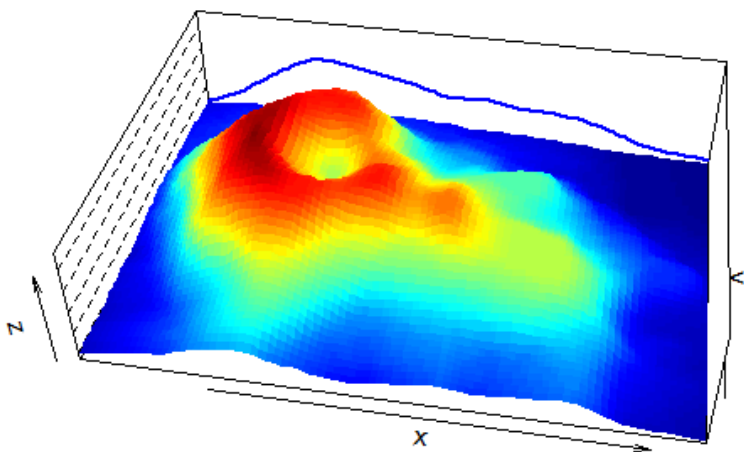




Estatística

Profa. Dra. Juliana Garcia Cespedes
PEDS - Programa de Excelência em Data Science
ITA+ITAÚ



Medidas Resumo

Medidas de posição

Vimos como resumir os dados por meio de tabelas de frequências e gráficos. Desta forma obtemos informações sobre o comportamento de uma variável.

Mas, muitas vezes, queremos resumir ainda mais estes dados, apresentando **um** ou **alguns valores** que sejam **representativos** da série toda.

Quando usamos um só valor, reduzimos drasticamente os dados.

Geralmente, empregam-se as medidas de posição (localização) central: **média, mediana ou moda**

Média

- **Conceito familiar:** É a soma das observações dividida pelo número total delas.
- **Conceito formal:** Se $x_1, \dots, x_n$ são os n valores da variável X , a média aritmética de X pode ser escrita:

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

Média

- Se os dados estiverem resumidos em uma tabela de frequências, então a média de X pode ser escrita:

$$\bar{X} = \frac{\sum_{j=1}^k f_j x_j}{n}$$

- Ou, usando a frequência relativa:

$$\bar{X} = \sum_{j=1}^k fr_j x_j$$

$$\frac{1 + 1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 4}{10}$$

$$= 2,4$$

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

$$\frac{1 + 1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 4}{10}$$

$$\frac{2*1 + 3*2 + 4*3 + 1*4}{10}$$

$$\frac{1 + 1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 4}{10}$$

$$\frac{2*1 + 3*2 + 4*3 + 1*4}{10}$$

$$\frac{1 + 1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 4}{10}$$

$$\frac{2 \times 1 + 3 \times 2 + 4 \times 3 + 1 \times 4}{10}$$

$$\bar{X} = \frac{\sum_{j=1}^k f_j x_j}{n}$$

$$\frac{1 + 1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 4}{10}$$

$$\frac{2 * 1 + 3 * 2 + 4 * 3 + 1 * 4}{10}$$

$$\frac{2 * 1}{10} + \frac{3 * 2}{10} + \frac{4 * 3}{10} + \frac{1 * 4}{10}$$

$$\frac{1 + 1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 4}{10}$$

$$\frac{2 * 1 + 3 * 2 + 4 * 3 + 1 * 4}{10}$$

$$\frac{2 * 1}{10} + \frac{3 * 2}{10} + \frac{4 * 3}{10} + \frac{1 * 4}{10}$$

$$\frac{1 + 1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 4}{10}$$

$$\frac{2 * 1 + 3 * 2 + 4 * 3 + 1 * 4}{10}$$

$$\frac{2 * 1}{10} + \frac{3 * 2}{10} + \frac{4 * 3}{10} + \frac{1 * 4}{10}$$

$$\bar{X} = \sum_{j=1}^k fr_j x_j$$

Média

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

$$\bar{X} = \frac{\sum_{j=1}^k f_j x_j}{n}$$

$$\bar{X} = \sum_{j=1}^k fr_j x_j$$

N. Filhos	Freq. Abs.	Freq. rela.	Freq. Acum.	%
0	4	0.2	0.2	20%
1	5	0.25	0.45	25%
2	7	0.35	0.8	35%
3	3	0.15	0.95	15%
4	0	0	0.95	0%
5	1	0.05	1	5%
Total	20	1		100%

$$\bar{X} = \frac{4*0 + 5*1 + 7*2 + ... + 1*5}{20} = 1,65$$

$$\bar{X} = 0,2*0 + 0,25*1 + 0,35*2 + ... + 0,05*5 = 1,65$$

Salário	Freq	Freq relativa	Freq acum	Porcentagem
[4,00; 8,00)	10	10/36 =0,278	0,278	27,78%
[8,00; 12,00)	12	12/36 =0,333	0,611	33,33%
[12,00; 16,00)	8	8/36 =0,222	0,833	22,22%
[16,00; 20,00)	5	5/36 =0,139	0,972	13,89%
[20,00; 24,00)	1	1/36 =0,029	1,000	2,78%
Total	36	1		100%

$$\bar{X} \approx \frac{10 * 6,00 + 12 * 10,00 + 8 * 14,00 + \dots + 1 * 22,00}{36} = 11,22$$

$$\bar{X} \approx 0,27 * 6,0 + 0,33 * 10,0 + 0,22 * 14,0 + \dots + 0,03 * 22,0 = 11,22$$

Mediana

- A *Mediana* é a realização que ocupa a posição central da série de observações, quando estão ordenadas em ordem CRESCENTE.
- Se o número de observações for ímpar a mediana é a posição central da série:

3, 4, 7, 8, 8

- Se o número de observações for par, usa-se como mediana a média aritmética das duas observações centrais:

3, 4, 7, 8 $(4+7)/2 = 5,5$



Mediana

- Conceito formal:

Considere as estatísticas de ordem $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

- A **mediana** da variável X pode ser definida como:

$$md(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & , \text{se } n \text{ ímpar;} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & , \text{se } n \text{ par.} \end{cases}$$

0 0 0 0 1 1 1 1 1 2 2 2 2 2 2 3 3 3 5

$$md(X) = \frac{X_{\left(\frac{20}{2}\right)} + X_{\left(\frac{20}{2}+1\right)}}{2} = \frac{X_{(10)} + X_{(11)}}{2} = \frac{2 + 2}{2} = 2$$

N. Filhos	Freq. Abs.	Freq. rela.	Freq. Acum.	%
0	4	0.2	0.2	20%
1	5	0.25	0.45	25%
2	7	0.35	0.8	35%
3	3	0.15	0.95	15%
4	0	0	0.95	0%
5	1	0.05	1	5%
Total	20	1		100%

Classe de salários	Freq	Freq relativa	Freq acum	Porcentagem
[4,00; 8,00)	10	10/36 =0,278	0,278	27,78%
[8,00; 12,00)	12	12/36 =0,333	0,611	33,33%
[12,00; 16,00)	8	8/36 =0,222	0,833	22,22%
[16,00; 20,00)	5	5/36 =0,139	0,972	13,89%
[20,00; 24,00)	1	1/36 =0,029	1,000	2,78%
Total	36	1		100%

$$md(X) \approx \frac{Pm_{(18)} + Pm_{(19)}}{2} = \frac{10,00 + 10,00}{2} = 10,00$$

É possível encontrar um valor mais preciso?????

Moda

- A **moda** é definida como a realização mais frequente do conjunto de valores observados.
- Podem haver mais de uma moda, ou seja, as distribuições dos valores podem ser: bimodal, trimodal, ..., multimodal.

0 0 0 0 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 5

$$mo(X) = 2$$

N. Filhos	Freq. Abs.	Freq. rela.	Freq. Acum.	%
0	4	0.2	0.2	20%
1	5	0.25	0.45	25%
2	7	0.35	0.8	35%
3	3	0.15	0.95	15%
4	0	0	0.95	0%
5	1	0.05	1	5%
Total	20	1		100%

Tempo de atendimento	Freq	Freq relativa	Freq acum	Porcentagem
[4,00; 8,00)	10	$10/36 = 0,278$	0,278	27,78%
[8,00; 12,00)	12	$12/36 = 0,333$	0,611	33,33%
[12,00; 16,00)	8	$8/36 = 0,222$	0,833	22,22%
[16,00; 20,00)	5	$5/36 = 0,139$	0,972	13,89%
[20,00; 24,00)	1	$1/36 = 0,029$	1,000	2,78%
Total	36	1		100%

$$mo(X) \approx 10,00$$

Medidas resumo

- Para calcular a **moda** precisamos apenas da tabela de frequências.
(Variáveis qualitativas e quantitativas)
- Para calcular a **mediana**, precisamos ordenar as realizações da variável.
(Variáveis qualitativas ordinal e quantitativas)
- Para a **média** precisamos que as variáveis sejam mensuráveis.
(Variáveis quantitativas)

Medidas de dispersão

- O resumo de um conjunto de dados por uma única medida de posição central esconde toda a informação sobre a variabilidade do conjunto de observações.
- Considere a nota de cinco grupos de alunos:
Grupo A (Variável X): 3,4,5,6,7
Grupo B (Variável Y): 1,3,5,7,9
Grupo C (Variável Z): 5,5,5,5,5
Grupo D (Variável W): 3,5,5,7
Grupo E (Variável V): 3,5,5,6,6

$$\begin{aligned}\bar{X} &= \bar{Y} = \bar{Z} = \\ &= \bar{W} = \bar{V} = 5\end{aligned}$$

Medidas de dispersão

- A média de cada grupo é igual, e com isso não conseguimos informação sobre sua variabilidade.
- Para resumir a variabilidade de um conjunto de dados utiliza-se a dispersão dos dados em torno da média, e as medidas mais usadas são a **variância** e o **desvio padrão**.

Medidas de dispersão

- Considere o grupo A:

3, 4, 5, 6, 7

Média = 5

- Distância das observações em relação à média:

3-5

4-5

5-5

6-5

7-5

-2

-1

0

1

2

 $(x_i - \bar{X})$

- Resumindo essas informações:

$$\frac{-2 -1 +0 +1 +2}{5} = 0$$



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

Medidas de dispersão

- Como resolver esse problema?

Medidas de dispersão

- Módulo das distância :

$$\begin{array}{ccccc} |3-5| & |4-5| & |5-5| & |6-5| & |7-5| \\ 2 & 1 & 0 & 1 & 2 \end{array} \quad \Rightarrow \quad |x_i - \bar{X}|$$

- Resumindo essas informações:

$$\frac{2 + 1 + 0 + 1 + 2}{5} = 1,2 \quad \Rightarrow \quad \frac{\sum_{i=1}^n |x_i - \bar{X}|}{n}$$

Desvio médio

Medidas de dispersão

- Distância ao quadrado :

$$\begin{array}{ccccc} (3-5)^2 & (4-5)^2 & (5-5)^2 & (6-5)^2 & (7-5)^2 \\ 4 & 1 & 0 & 1 & 4 \end{array} \Rightarrow (x_i - \bar{X})^2$$

- Resumindo essas informações:

$$\frac{4 + 1 + 0 + 1 + 4}{5} = 2 \Rightarrow \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$$

Variância

Medidas de dispersão

- Opções:
- Considerar o total dos módulos dos desvios:

$$\sum_{i=1}^n |x_i - \bar{X}|$$

- Considerar o total dos quadrados dos desvios

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

Medidas de dispersão

- Mas como comparar essas medidas quando os conjuntos de dados tem tamanhos diferentes?
- É melhor exprimir as medidas como médias:

$$dm(X) = \frac{\sum_{i=1}^n |x_i - \bar{X}|}{n}$$

Desvio médio

$$\text{var}(X) = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$$

Variância

Medidas de dispersão

As medidas de variabilidade indicam o quão homogêneo é um conjunto de dados.

Para os grupos A e E tem-se:

$$dm(X) = \frac{2+1+0+1+2}{5} = 1,2$$

$$\text{var}(X) = \frac{4+1+0+1+4}{5} = 2,0$$

$$dm(V) = \frac{2+0+0+1+1}{5} = 0,8$$

$$\text{var}(V) = \frac{4+0+0+1+1}{5} = 1,2$$

Medidas de dispersão

A variância é uma medida de dispersão igual ao quadrado da dimensão dos dados, pode causar problemas de interpretação. Costuma-se usar, então o **desvio padrão**, que é definido como a raiz quadrada positiva da variância.

$$dp(X) = \sqrt{\text{var}(X)} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}}$$

Medidas de dispersão

A medida de dispersão, **desvio padrão**, indica em média qual será o “erro” (desvio) cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados (Média).

Se os dados estiverem agrupados em tabelas de frequências, as medidas de dispersão são dadas por:

$$\text{var}(X) = \frac{\sum_{j=1}^k f_j (x_j - \bar{X})^2}{n} = \sum_{j=1}^k fr_j (x_j - \bar{X})^2$$

Medidas de dispersão

Uma maneira equivalente de calcular a variância, computacionalmente mais eficiente, é:

$$\begin{aligned}\text{var}(X) &= \frac{\sum_{i=1}^n x_i^2}{n} - \bar{X}^2 \\ &= \frac{\sum_{j=1}^k f_j x_j^2}{n} - \bar{X}^2 \\ &= \sum_{j=1}^k fr_j x_j^2 - \bar{X}^2\end{aligned}$$

Programa R

`mean(variavel)`

`median(variavel)`

`var(variavel)`

`varp<-sum((variavel - mean(variavel))^2/(length(variavel)))`

`sd(variavel)`

`summary(variavel)`

Exercício

Quer se estudar o número de erros de impressão de um livro. Para isso observou-se 50 páginas, encontrando-se o número de erros por página.

- Qual o número médio de erros? E o número mediano?
- Qual é o desvio padrão?
- Se o livro tem 500 páginas, qual o número total de erros esperado no livro?

Erros	Freq
0	25
1	20
2	3
3	1
4	1

Solução

- Média = 0,66
- Mediana = 0,5
- Desvio padrão = 0,8393
- Valor esperado = 330