5. MEDIDAS ESTATÍSTICAS

5.1. Introdução

No tópico anterior foi visto a síntese (ou resumo) de séries de dados sob a forma de apresentações tabulares, apresentações gráficas e as distribuições de freqüências.

Trata-se agora dos cálculos das medidas que possibilitam apresentar e confrontar séries de dados, relativas às observações dos fenômenos, de forma sintética e resumida.

5.2. Medidas de tendência central

Tais medidas orientam quanto aos valores centrais.

Representam os fenômenos pelos seus valores médios, em torno dos quais tendem a se concentrar os dados.

5.2.1. Média aritmética

Medida de tendência central de uso mais comum.

Notação adotada: ($\overline{Y}$ ou μ) para o parâmetro e ($\overline{y}$ ou m) para a estimativa.

5.2.1.1. Dados não agrupados

Sejam y_1, y_2, ..., y_n, portanto (N, n) valores da variável Y. A média aritmética simples de Y representada por ($\overline{Y}$, $\overline{y}$) é definida por:

Parâmetro: $\overline{Y} \; ou \; \mu = \dfrac{\sum y}{N}$

Estimativa: $\overline{y} \; ou \; m = \dfrac{\sum y}{n}$

Exemplo: considerando {3, 7, 8, 10, 11} como uma amostra:

$$\overline{y} = \frac{3+7+8+10+11}{5} = 7{,}8$$

5.2.1.2. Dados agrupados

Quando os dados de uma amostra estiverem agrupados numa distribuição de freqüência a média aritmética dos valores de Y (y_1, y_2, ..., y_n), ponderados pelas respectivas freqüências absolutas: F_1, F_2, ..., F_n é calculada como se segue:

$$\overline{y} \; ou \; m = \frac{\sum y \cdot F}{\sum F}$$

Exemplos:

Y	F_i	YF_i
1	1	1
2	3	6
3	5	15
4	1	4
Σ	10	26

$$\overline{y} \; ou \; m = \frac{\sum y \cdot F}{\sum F} = \frac{26}{10} = 2{,}6$$

Idade	F_i	Y	YF_i
02 ⊢ 04	5	3	15
04 ⊢ 06	10	5	50
06 ⊢ 08	14	7	98
08 ⊢ 10	8	9	72
10 ⊢ 12	3	11	33
Σ	40		268

As classes são representadas pelos seus pontos médios:

$$\overline{y} \; ou \; m = \frac{\sum y \cdot F}{\sum F} = \frac{268}{40} = 6{,}7$$

5.2.1.3. Média geral

Sejam $\overline{y}_1$, $\overline{y}_2$, ..., $\overline{y}_k$ as estimativas das médias aritméticas de K séries e n_1, n_2, ..., n_k os números de termos destas séries, respectivamente. A média aritmética da série formada pelos termos da K séries é dada pela fórmula:

$$\overline{y} \; ou \; m = \frac{n_1\overline{y}_1 + n_2\overline{y}_2 + \ldots + n_k\overline{y}_k}{n_1 + n_2 + \ldots + n_k}$$

Exemplo: Sejam as séries:

1) {4, 5, 6, 7, 8} em que $n_1 = 5$ e $\overline{y}_1 = 6$

2) {1, 2, 3} em que $n_2 = 3$ e $\overline{y}_2 = 2$

3) {9, 10, 11, 12, 13} em que $n_3 = 5$ e $\overline{y}_3 = 11$

$$\overline{Y} = \frac{n_1\overline{y}_1 + n_2\overline{y}_2 + \ldots + n_k\overline{y}_k}{n_1 + n_2 + \ldots + n_k} = \frac{5\cdot 6 + 3\cdot 2 + 5\cdot 11}{5+3+5} = 7$$

5.2.2. Média geométrica

Usada para médias proporcionais de crescimento quando uma medida subseqüente depende de medidas prévias.

Notação adotada: (MG) para o parâmetro e (mg) para a estimativa.

Sejam $y_1, y_2, ..., y_n$, valores de Y associados às respectivas freqüências absolutas $F_1, F_2, ..., F_n$. A média geométrica (MG ou mg) de Y é definida por:

$$MG\ ou\ mg = \sqrt[n]{y_1^{F_1} \cdot y_2^{F_2} \cdot ... \cdot y_n^{F_n}}$$

Exemplo:

Média geométrica de uma amostra {3, 6, 12, 24, 48}

$$mg = \sqrt[5]{3 \cdot 6 \cdot 12 \cdot 24 \cdot 48} = \sqrt[5]{248.832} = 12$$

5.2.3. Média harmônica

Usada para médias de crescimento e proporções de velocidade.

Notação adotada: (MH) para o parâmetro e (mh) para a estimativa.

Sejam $y_1, y_2, ..., y_n$, valores de Y, associados às respectivas freqüências absolutas $F_1, F_2, ..., F_n$. A média harmônica (MH ou mh) de Y é definida por:

$$MH\ ou\ mh = \frac{n}{\frac{F_1}{y_1} + \frac{F_2}{y_2} + ... + \frac{F_n}{y_n}} = \frac{n}{\sum_{i=1}^{n} \frac{F_i}{y_i}}$$

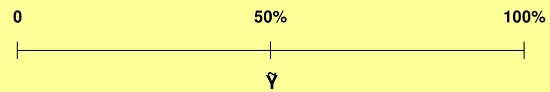
Exemplo:

Média harmônica de uma amostra {2, 5, 8}

$$mh = \frac{3}{\frac{1}{2} + \frac{1}{5} + \frac{1}{8}} = 3{,}64$$

5.2.4. Mediana

Medida de tendência muito usada quando o interesse é a determinação do valor que separa a série de dados em duas partes iguais, 50% situados acima e 50% situados abaixo da medida.



Notação adotada: ($\tilde{Y}$ ou MD) para o parâmetro e ($\tilde{y}$ ou mg) para a estimativa.

Colocados em ordem crescente, a mediana ($\tilde{Y}$, $\tilde{y}$) é o valor que divide a série em duas partes iguais:

5.2.4.1. Cálculo da mediana para variável discreta

Se n for impar, a mediana será o elemento central (de ordem $\frac{n+1}{2}$).

Caso n seja par, a mediana será a média entre os elementos centrais (de ordem $\frac{n}{2}$ e $\frac{n}{2}+1$).

Exemplo 1:

Y_i	F_i	F_{ac}	
1	1	1	
2	3	4	
3	5	9	← contém o 6º elemento
4	2	11	
Σ	11		

n = 11, n é impar, logo $\tilde{y}$ será o elemento de ordem $\frac{n+1}{2}$, ou seja $\frac{11+1}{2} = 6º$

Será, portanto, o sexto elemento.

Por meio das freqüências acumuladas encontra-se o valor y_i correspondente a mediana, que neste exemplo é 3 ($\tilde{y}$ = 3).

Exemplo 2:

Y_i	F_i	F_{ac}	
82	5	5	
85	10	15	
87	15	30	← contém os 21º e 22º elementos
89	8	38	
90	4	42	
Σ	42		

n = 42, n é par, logo $\tilde{y}$ será a média entre os elementos de ordem $\frac{n}{2}$ e $\frac{n}{2}+1$

Ou seja $\frac{42}{2} = 21$ e $\frac{42}{2} + 1 = 22$ (21º e 22º elementos)

Identificam-se os elementos de ordem 21º e 22º pela F_{ac}

Assim, o 21º corresponde a 87 e 22º corresponde a 87, logo:

$$\tilde{y} = \frac{87+87}{2} = 87$$

5.2.4.2. Cálculo da mediana para variável contínua

1º passo: calcula-se a ordem $\frac{n}{2}$. Como a variável é contínua, não importa se n é par ou impar.

2º passo: pela F_{ac} identifica-se a classe que contém a mediana (classe md).

3º passo: usa-se fórmula:

$$\tilde{y} = \ell_{md} + \frac{\left(\frac{n}{2} - \sum f\right) \cdot h}{F_{md}}$$

Em que:

ℓ_{md} = limite inferior da classe md

n = tamanho da série

$\sum f$ = soma das freqüências anteriores à classe md

h = amplitude da classe md

F_{md} = freqüência da classe md

Exemplo:

Classe	F_i	F_{ac}	
35 ⊢ 45	5	5	
45 ⊢ 55	12	17	
55 ⊢ 65	18	35	← classe mediana
65 ⊢ 75	14	49	
75 ⊢ 85	6	55	
85 ⊢ 95	3	58	
Σ	58	268	

1º passo: $\frac{58}{2} = 29^o$

2º passo: classe md = 3ª

3º passo: usa-se a fórmula:

$\ell_{md} = 55$; n = 58; $\sum f = 17$; h = 10; $F_{md} = 18$

$$\tilde{y} = 55 + \frac{\left(\frac{58}{2} - 17\right) \cdot 10}{18} = 61{,}67$$

5.2.5. Moda

Medida de tendência central muito usada quando o interesse é o valor mais freqüente da série.

Notação adotada: (MO) para o parâmetro e (mo) para a estimativa.

A moda pode não existir – o que constitui uma série amodal – ou, mesmo que exista pode não ser única – o que caracteriza uma série multimodal.

Para distribuições simples (sem agrupamento de classes), a identificação da moda é facilitada pela simples observação do elemento que apresenta maior freqüência.

Assim, considerando a distribuição abaixo como uma amostra:

y_i	243	245	248	251	307
F_i	7	17	23	20	8

A moda será 248, e indica-se por mo = 248.

5.2.5.1. Moda para dados agrupados em classes

1º passo: identifica-se a classe modal (maior freqüência).

2º passo: usa-se a fórmula de Czuber

$$mo = \ell + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot h$$

Em que:

ℓ = limite inferior da classe mo
Δ_1 = diferença entre a freqüência da classe modal e a imediatamente anterior
Δ_2 = diferença entre a freqüência da classe modal e a imediatamente posterior
h = amplitude da classe

Exemplo:

Classes	0 ⊢ 1	1 ⊢ 2	2 ⊢ 3	3 ⊢ 4	4 ⊢ 5	Σ
F_i	3	10	17	8	5	43

1º passo: indica-se a classe modal: 3ª (2 ⊢ 3)

2º passo: usa-se a fórmula

$$mo = \ell + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot h$$

$$mo = 2 + \frac{7}{7 + 9} \cdot 1 = 2{,}44$$

5.3. Comparação entre as medidas de tendência central

5.3.1. Média

5.3.1.1. Vantagens

- Fácil de compreender e calcular
- Utiliza todos os valores da série
- É um valor único
- É fácil de ser incluída em expressões matemáticas
- Pode ser determinada nas escalas: intervalar e proporcional.

5.3.1.2. Desvantagens

- Muito afetada por valores extremos
- Necessário conhecer todos os valores da série.

5.3.2. Mediana

5.3.2.1. Vantagens

- Fácil de compreender e aplicar
- Não é afetada por valores extremos
- É um valor único
- É fácil de incluir em expressões matemáticas
- Pode ser determinada nas escalas: ordinal, intervalar e proporcional.

5.3.2.2. Desvantagens

- É difícil de ser incluída em expressões matemáticas
- Não usa todos os valores da série.

5.3.3. Moda

5.3.3.1. Vantagens

- Fácil de compreender e calcular
- Não é afetada por valores extremos
- Pode ser aplicada em todas as escalas: nominal, ordinal, intervalar e proporcional.

5.3.3.2. Desvantagens

- Pode estar afastada do centro dos valores
- É difícil de ser incluída em expressões matemáticas
- Não usa todos os valores da série
- A variável pode ter mais de uma moda (bimodal ou multimodal)
- Algumas variáveis não possuem moda.

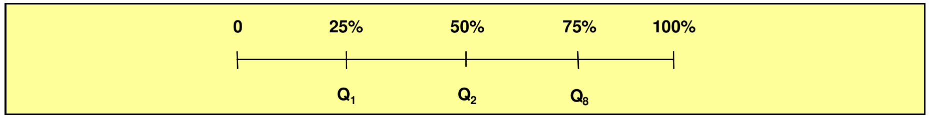
5.4. Medidas de posição ou separatrizes

Genericamente denominadas quantis, orientam quanto à posição na distribuição.

Permitem determinar valores que particionam a série de n observações em partes iguais.

5.4.1. Quartis

Seguindo o mesmo raciocínio da mediana, os três quartis dividem uma série em 4 partes iguais:



Notação adotada: (Q) para o parâmetro e (q) para a estimativa.

$$q_i = \ell_{q_i} + \frac{\left(\frac{i \cdot n}{4} - \sum f\right) \cdot h}{F_{q_i}}$$

Em que:

ℓ_{q_i} = limite inferior da classe q_i (i = 1, ..., 3)

i = 1 para q_1, ..., 3 para q_3

n = tamanho da série

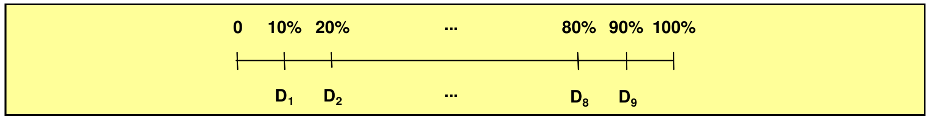
$\sum f$ = soma das freqüências anteriores à classe q_i

h = amplitude da classe q_i

F_{q_i} = freqüência da classe q_i

5.4.2. Decis

Os decis dividem a série em 10 partes iguais.



Notação adotada: (D) para o parâmetro e (d) para a estimativa.

$$d_i = \ell_{d_i} + \frac{\left(\frac{i \cdot n}{10} - \sum f\right) \cdot h}{F_{d_i}}$$

Em que:

ℓ_{d_i} = limite inferior da classe d_i (i = 1, ..., 9)

i = 1 para d_1, ..., 9 para d_9

n = tamanho da série

$\sum f$ = soma das freqüências anteriores à classe d_i

h = amplitude da classe d_i

F_{d_i} = freqüência da classe d_i

5.4.3. Percentis

Os percentis (P para populações e p para amostras) dividem a série em 100 partes iguais.

Notação adotada: (P) para o parâmetro e (p) para a estimativa.

$$p_i = \ell_{p_i} + \frac{\left(\frac{i \cdot n}{100} - \sum f\right) \cdot h}{F_{p_i}}$$

Em que:

ℓ_{p_i} = limite inferior da classe p_i (i = 1, ..., 99)

i = 1 para p_1, ..., 99 para p_{99}

n = tamanho da série

$\sum f$ = soma das freqüências anteriores à classe p_i

h = amplitude da classe p_i

F_{p_i} = freqüência da classe p_i

Os procedimentos para determinar os quartis, decis e percentis são semelhantes aos usados para determinar o valor da mediana.

5.4.4. Situações de uso mais comuns destas medidas

Uma dos usos mais comuns, e importantes, destas medidas na análise exploratória dos dados é o diagrama de caixa (“box plot”), como abaixo:

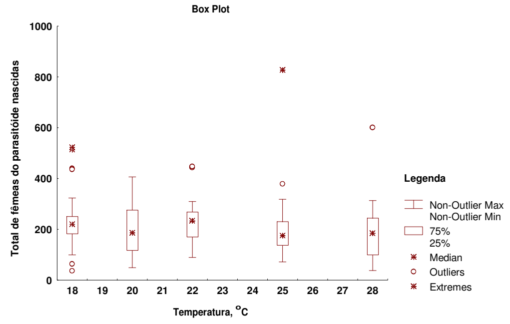


Figura 5.1 – Diagrama de caixa do total de fêmeas do parasitóide nascidas.