Análise Exploratória e Estimação

MONITORIA DE ESTATÍSTICA E PROBABILIDADE PARA COMPUTAÇÃO

Médias

Média Aritmética (valor médio de uma distribuição)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n)$$

Média Aritmética (dados agrupados)

$$\bar{X} = \frac{(f_1X_1 + \cdots + f_kX_k)}{f_1 + \cdots + f_k} = \frac{\sum_{i=1}^{k} f_iX_i}{\sum_{i=1}^{k} f_i}$$

Exemplo

Intervalos de classes	Frequência absoluta
12,51 a 13,50	3
13,51 a 14,50	8
14,51 a 15,50	15
15,51 a 16,50	13
16,51 a 17,50	9
17,51 a 18,50	2

$$\bar{X} = \frac{3 \cdot 13 + 8 \cdot 14 + 15 \cdot 15 + 13 \cdot 16 + 9 \cdot 17 + 2 \cdot 18}{30} = 15{,}46$$

Médias

Média Ponderada: $\bar{X} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$

Média Harmônica: $H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$

Média Geométrica: $G = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$

Mediana

Para valores ordenados crescentemente, dois modos de calcular:

- Se n é ímpar, mediana é o valor central:
  - Na amostra 30 32 35 48 76 a mediana é 35

- Se n é par, mediana é a média simples entre os dois valores centrais:
  - Na amostra 30 32 35 48 76 81 a mediana é $\frac{34+48}{2} = 41{,}5$

Mediana para dados agrupados

1. Calcula-se n/2;
2. Achar qual das classes esse valor se encontra a partir das frequências absolutas;
3. Usar a fórmula

$$Md = l_{Md} + \frac{\left(\frac{n}{2} - \sum f\right) \cdot h}{f_{Md}}$$

Aonde:

l_{Md} é o limite inferior da classe;

f_{Md} é a frequência da classe da mediana;

$\sum f$ é a Soma das frequências anteriores a classe da mediana;

h é a amplitude da classe da mediana.

Moda

Valor que ocorre com maior frequência.

- 2 6 2 9 8 4 3 2 4 5

  2 2 2 3 4 4 5 6 8 9

  Mo = 2

- 45 46 49 52 52 60 60 76 79

  Mo = 52 e 60

Moda para Dados Agrupados

Utiliza-se a fórmula de King:

$$Mo = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot h$$

Aonde:

- l - limite inferior da classe modal = 40
- Δ_1 - diferença entre a frequência da classe e a anterior = 16
- Δ_2 - diferença entre a frequência da classe e a posterior = 7
- h - amplitude da classe modal = 20

Notas	Número de Alunos
0 |- 20	2
20 |- 40	7
40 |- 60	23
60 |- 80	16
80 |- 100	3
Total	51

Amplitude Total

É a diferença entre o maior e menor valor de um conjunto de dados.

Amplitude = (maior valor) - (menor valor)

Exemplo:

30,4 34,7 39,8 40,45 47,9 49,5 51,9 69,7

69,7-30,4 = 39,3

Desvio Padrão

Variação dos valores em torno de uma média dado um conjunto de valores amostrais.

Para uma população de N indivíduos:

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

Para uma amostra de *n* observações, $x_1, \ldots, x_n$:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Aonde:

- x_i é o valor de cada variável;
- $\bar{x}$ é a média amostral e μ é a média populacional.

Coeficiente de Variação

Percentual do desvio padrão com relação à média.

- Para população

$$cv = \frac{\sigma}{\mu}$$

- Para amostra

$$cv = \frac{s}{\bar{x}}$$

Variância

A medida da variação é o quadrado do desvio padrão.

Para a população: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Para a amostra: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Aonde:

- x_i é o valor de cada variável;
- $\bar{x}$ é a média amostral e μ é a populacional.

Obs.: Dado um desvio padrão de unidade “u” a variância do mesmo terá unidade “u²”.

Amplitude Inter-quartílica

É a amplitude do intervalo entre o primeiro e o terceiro quartil. Representada por Q.

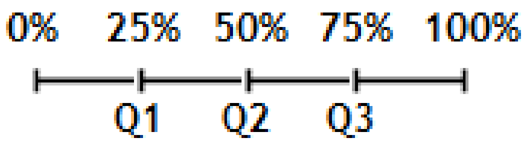
$$Q = Q3 - Q1$$

Obs: Às vezes também é usada a semi-amplitude inter-quartílica, que é a metade da anterior.

Obs2: Q é aproximadamente igual a $\frac{4}{3}\sigma$

Medida de Posição - Quartil

1. Quartil é qualquer um dos três valorres que divide o conjunto em quatro partes iguais.



2. Para dados agrupados.

$$Q_1 = l_{Q_1} + \frac{\left(\frac{n}{4} - \sum f\right) \cdot h}{F_{Q_1}}$$

Obs: Se fosse para calcularmos o Q3, o fariamos na razão de 3n/4 !

Percentil

Valores que dividem o conjunto em partes iguais que representam 1/100 da amostra ou população!

Seja N igual ao tamanho amostral, temos:

$$P_k = \frac{N \cdot k}{100}$$

(arredondar para o inteiro mais próximo)

Percentil para dados agrupados

$$P_i = l_{P_i} + \frac{\left(\frac{in}{100} - \sum f\right) \cdot h}{F_{P_i}}$$
$$i \in \{1,2,3,4, \dots, 96,97,98,99,100\}$$

Aonde:

l_{P_i} é o limite inferior de P_i

$\sum f$ é a soma das frequências anteriores de P_i

h é a amplitude da classe de P_i

F_{P_i} é a frequência da classe P_i

Medida de Assimetria

O calculo da Assimetria resultará em valores sempre entre -1 e 1 e para tal utilizamos a equação de Pearson:

$$Sk = \frac{\bar{X} - Mo}{S}$$

Construção de tabelas de distribuição de frequência

Objetivo: construir tabelas de distribuição de frequência a partir de dados brutos (n observações).

1º Passo: determinar a amplitude total;

2º Passo: estimar o número de intervalos;

- Pode-se utilizar $K = \sqrt{n}$, para $n > 25$ e $K = 5$ para $n < 25$
- Ou a fórmula de Sturges: $K = 1 + 3{,}22 \log n$

3º Passo: estimar a amplitude dos intervalos: $h = \frac{R}{K}$;

4º Passo: esquematizar a tabela de acordo com as informações dos passos anteriores.

Estimação

Estimativa pontual:

- $\bar{x}$ é uma estimativa pontual para μ, onde $(x_1, \dots, x_n)$ é uma amostra.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n)$$

Estimativa intervalar (**intervalo de confiança**):

- Intervalo de valores que contém a média da população com uma determinada probabilidade de acerto
- É necessário calcular a margem de erro do intervalo ($\bar{x} - \mathrm{E}$ e $\bar{x} + \mathrm{E}$) de acordo com o nível de confiança pedido, e dependendo se a variância é conhecida ou não.

Intervalo de confiança

Variância conhecida

O erro é dado por: $E = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Logo, o intervalo de confiança para média μ é: $\bar{x} - \mathrm{E} \leq \mu \leq \bar{x} + E$

Variância desconhecida

É necessário calcular a variância da amostra por:

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

Então, o erro é dado por: $E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ aonde $t_{\alpha/2}$ é o valor correspondente a $\alpha/2$ com n – 1 graus de liberdade.

O intervalo de confiança para média μ é: $\bar{x} - \mathrm{E} \leq \mu \leq \bar{x} + E$

Exercícios

1. Para a distribuição abaixo responda:
   a) Qual a amplitude total?
   b) Ponto médio do terceiro intervalo.
   c) Qual(is) o comprimento dos intervalos?
   d) Qual a porcentagem de internautas que gastam acima de 42 minutos na internet?
   e) Qual o valor: modal, mediano e médio? O que eles representam na distribuição?

Tempo (minutos)	Internautas
7 |-- 18	6
18 |-- 31	10
31 |-- 42	13
42 |-- 54	8
54 |-- 66	5
66 |-- 78	6
78 |-- 90	2

Recife - 2009
Fonte: Fictícia.

Resolução

a) Amplitude total = 90 – 7 = 83

b) Ponto Médio 3º classe = 42+31/2 = 66,5

c) Comprimento dos intervalos = Amplitude de cada intervalo. Exemplo:
1º 18 – 7 = 11; 2º 31 – 18 = 13 [...]

d) Porcentagem de users para > 42min, a partir da 4ª classe: $\frac{8+5+6+2}{50}$= 0,42

e) Moda, Mo = 31—42| , pois aparece com maior frequência.
Média, $\frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i} = \frac{(12{,}5*6+24{,}5*10+36{,}5*13+48*8+60*5+72*6+84*2}{50} = \frac{2082{,}5}{50} = 4{,}65$
Mediana, n/2 = soma das frequencias/2 = 50/2 = 25. Se fizermos a tabela de frequências acumuladas esse valor vai referenciar a 3ª classe. Então:
$\mathrm{Md} = 31 + \frac{(25-16)\cdot 11}{13} = 38{,}61$

Exercícios

2. Considere a seguinte distribuição de frequências.
   a) Calcule a média, a variância e o desvio padrão, a mediana e a moda.
   b) Qual das medidas de tendência central descreve melhor os dados? Justifique

X_i	-4	-3	-2	-1	0	1	2	3	4
f_i	60	120	180	200	240	190	160	90	30

Resolução

a)

MÉDIA: $\bar{X} =$

$$(60 \cdot (-4) + 120 \cdot (-3) + 180 \cdot (-2) + 200 \cdot (-1) + 240 \cdot 0 + 190 \cdot 1 + 160 \cdot 2 + 90 \cdot 3 + 30 \cdot 4) / 1270$$

$$= \underline{\underline{-0,204}}$$

DESVIO PADRÃO =

$$s = \sqrt{3,958} = \underline{\underline{1,9895}}$$

VARIÂNCIA:

$$\frac{1}{n} \sum fi \cdot (x_i - \bar{X})^2 =$$

$$\frac{1}{1270} \big(60 \cdot (-4 - (-0,204))^2 + 120 \cdot (-3 - (-0,204))^2 + 180 \cdot (-2 - (-0,204))^2 + 200 \cdot (-1 - (-0,204))^2 + 240 \cdot (0 - (-0,204))^2 + 190 \cdot (1 - (-0,204))^2 + 160 \cdot (2 - (-0,204))^2 + 90 \cdot (3 - (-0,204))^2 + 30 \cdot (4 - (-0,204))^2 \big)$$

$$s^2 = \frac{1}{1270} \cdot 5026,772 = \underline{\underline{3,958}}$$

Continuação...

$$Mediana = 0 + \frac{\left(\frac{1270}{2} - 560\right) x1}{800}$$

Obs: O limite inferior da classe é o próprio valor.

Obs2: A amplitude da classe é 1, pois só existe um elemento.

$$Mediana = 0,09375$$

MEDIANA → VALOR CENTRAL.

$\frac{1270}{2} = 635$ → ESTÁ NESSA POSIÇÃO.

FREQUÊNCIA ACUMULADA

x_i	f_i
-4	60
-3	180
-2	360
-1	560
0	800
1	990
2	1150
3	1240
4	1270

A MEDIANA ESTÁ NESSA FAIXA

Continuação...

MODA = 0

(VALOR COM A MAIOR FREQUÊN-
CIA)

b) COMO A DISTRIBUIÇÃO DOS
DADOS ESTÁ BEM LOCALIZADA
EM TORNO DA MÉDIA, QUAL-
QUER UMA DAS MEDIDAS CEN-
TRAIS (MÉDIA, MODA OU MEDIA-
NA) É ADEQUADA. PORÉM, COMO
A VARIÁVEL NÃO ASSUME VALORES
DECIMAIS, ENTÃO É MELHOR CONSI-
DERAR A MODA OU A MEDIANA.

Exercícios

3. Seguidamente apresentam-se algumas estimativas para a velocidade da luz, determinadas por Michelson em 1882 (Statistics and Data Analysis, Siegel):

299.88, 299.90, 299.94, 299.88, 299.96, 299.85, 299.94, 299.80, 299.84

a) Determine a média
b) Determine o desvio padrão, utilizando a expressão da definição.
c) Subtraia 299 de cada um dos dados e determine o desvio padrão, dos resultados obtidos, utilizando a fórmula utilizada na alínea anterior. Comente os resultados obtidos.
d) Calcule a média dos valores com que trabalhou na alínea anterior. Adicione à média obtida 299.

Resolução

a) $\bar{x} = \frac{1}{9}(299.88 + 299.90 + 299.94 + 299.88 + 299.96 + 299.85 + 299.94 + 299.80 + 299.84) = 299.8878$

b) $S^2 = \frac{1}{8}(299.88 - 299.877)^2 + \frac{1}{8}(299.90 - 299.877)^2 + \frac{1}{8}(299.94 - 299.877)^2 + \frac{1}{8}(299.88 - 299.877)^2 + \frac{1}{8}(299.96 - 299.877)^2 + \frac{1}{8}(299.85 - 299.877)^2 + \frac{1}{8}(299.94 - 299.877)^2 + \frac{1}{8}(299.80 - 299.877)^2 + \frac{1}{8}(299.84 - 299.877)^2 = 0{,}0028$

(observe que para uma amostra utiliza-se n-1)

Resolução

b) Com a variância, calculamos o desvio padrão:

$$S = \sqrt{0{,}0028} = 0{,}0528$$

c) Precisamos da nova média para calcular o desvio padrão (isso já responde a letra d):

$$\bar{x} = \frac{1}{9}(0.88 + 0.90 + 0.94 + 0.88 + 0.96 + 0.85 + 0.94 + 0.80 + 0.84)$$

$$= 0.8878$$

Calculando a variância...

Resolução

c) $S^2 = \frac{1}{8}(0.88 - 0.877)^2 + \frac{1}{8}(0.90 - 0.877)^2 + \frac{1}{8}(0.94 - 0.877)^2 +$

$\frac{1}{8}(0.88 - 0.877)^2 + \frac{1}{8}(0.96 - 0.877)^2 + \frac{1}{8}(0.85 - 0.877)^2 +$

$\frac{1}{8}(0.94 - 0.877)^2 + \frac{1}{8}(0.80 - 0.877)^2 + \frac{1}{8}(0.84 - 0.877)^2 = 0{,}0028$

Desvio padrão...

$$S = \sqrt{0{,}0028} = 0{,}0528$$

Resolução

c) Comentário:

O desvio padrão foi o mesmo da amostra anterior. Isso significa que a amostra está variando da mesma maneira, apesar de cada valor ter sido diminuído em 299. Observe que, consequentemente, a média também diminuiu 299 quando cada valor da amostra foi diminuído em 299.