Universidade Federal de Mato Grosso - UFMT
Probabilidade e Estatística

1 Somatório

Um somatório é um operador matemático que nos permite representar facilmente somas muito grandes ou até infinitas. É representado com a letra grega sigma Σ, e é definido por:

$$\sum_{i=1}^{n} x_i$$

em que corresponde a soma dos termos "x_i", em que o índice i varia de 1 a n.

Regras do somatório:

- Somatório de uma constante
  Se k é uma constante, então

$$\sum_{i=1}^{n} k = k + k + k + ... + k = nk$$

- Somatório do produto de uma constante por uma variável
  Se k é uma constante e x_i uma variável

$$\sum_{i=1}^{n} kx_i = kx_1 + kx_2 + kx_3 + ... + kx_n = k(x_1 + x_2 + x_3 + ... + x_n) = k\sum_{i=1}^{n} x_i$$

- Somatório de uma soma algébrica
  O somatório de uma soma de variáveis é igual à soma dos somatórios de cada variável

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

  Se a e b são constantes e x_i uma variável

$$\sum_{i=1}^{n} (a + bx_i) = \sum_{i=1}^{n} a + \sum_{i=1}^{n} bx_i = na + b\sum_{i=1}^{n} x_i$$

Observações:

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i &\neq \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i^2 &\neq \left(\sum_{i=1}^{n} x_i\right)^2 \end{aligned}$$

Exemplos: Seja $X = \{4, 7, 9, 12, 3\}$, Obter

$\sum_{i=1}^{5} x_i = 35$, $\sum_{i=1}^{4} 2x_i = 64$, $\sum_{i=2}^{5} 3x_i = 93$

Sabendo que $\sum_{i=1}^{3} x_i = 6$, $\sum_{i=1}^{3} x_i^2 = 14$, determinar

a) $\sum_{i=1}^{3} (x_i + 1) = \sum_{i=1}^{3} x_i + \sum_{i=1}^{3} 1 = 6 + 3 = 9$

b) $\sum_{i=1}^{3} (x_i - 1)^2 = \sum_{i=1}^{3} \left(x_i^2 - 2x_i + 1\right) = \sum_{i=1}^{3} x_i^2 - 2\sum_{i=1}^{3} x_i + \sum_{i=1}^{3} 1 = 14 - 12 + 3 = 5$

2 Medidas de Posição

Medidas de Posição - São medidas de tendência central, ou seja, representativas do valor central, ao redor do qual se agrupam a maioria dos valores.

2.1 Média Aritmética

A média de uma população ou amostra é a soma de todos os elementos da população (amostra) dividida pelo número de elementos. Esta medida apresenta a mesma unidade dos dados.

- Para a população a média é representada por

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

em que N é o tamanho da população

- Para a amostra a média é representada por

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

em que n é o tamanho da amostra.

A média calculada dos dados originais e dados agrupados podem ser diferentes, devido ao erro de agrupamento. O erro de agrupamento é obtido fazendo a diferença entre o valor obtido pelos dados originais e o valor obtido pelos dados agrupados.

Exemplo: O tempo de vida útil (em horas) de uma amostra de 6 lâmpadas incadescentes é: 612, 983, 623, 883, 666 , 970. A média amostral do tempo de vida é dado por:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{612 + 983 + 623 + 883 + 666 + 970}{6} = \frac{4737}{6} = 789,5$$

2.1.1 Propriedades da média

A média aritmética de uma amostra apresenta um conjunto vasto de propriedades, todas elas, sem dúvida, de grande utilidade no cálculo do seu valor.

1. Adição ou Subtração por uma constanteSeja $(X_1, X_2, X_3, ..., X_n)$ uma amostra aleatória de tamanho n, k uma constante e $\overline{X}$ a média da amostra. Se somarmos ou subtrairmos todos os valores de uma variável X pela constante k, o valor de $\overline{X}$ MÉDIA fica multiplicada ou dividida pela constante.

$$\begin{aligned}
\overline{X}^* &= \frac{\sum_{i=1}^{n}(X_i + k)}{n} \\
&= \frac{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} k}{n} \\
&= \frac{\sum_{i=1}^{n} X_i}{n} + \frac{\sum_{i=1}^{n} k}{n} \\
&= \overline{X} + \frac{nk}{n} \\
&= \overline{X} + k
\end{aligned}$$

   Se no exemplo das lâmpadas somarmos a constante 2 a cada um dos valores da variável temos 614, 985, 625, 885, 667,972

$$\overline{X}^* = \frac{614 + 985 + 625 + 885 + 668 + 972}{6} = \frac{4749}{6} = 791,5$$

   Utilizando a propriedade,

$$\overline{X}^* = \overline{X} + k = 789,5 + 2 = 791,5$$

2. Multiplicação ou divisão por uma constante
   Seja $(X_1, X_2, X_3, ..., X_n)$ uma amostra aleatória de tamanho n, k uma constante e $\overline{X}$ a média da amostra. Se multiplicarmos ou dividirmos todos os valores de uma variável X pela constante k, o valor de $\overline{X}$ MÉDIA fica multiplicada ou dividida pela constante.

$$\begin{aligned}
\overline{X}^* &= \frac{\sum_{i=1}^{n} kx_i}{n} \\
&= k\frac{\sum_{i=1}^{n} x_i}{n} \\
&= k\overline{X}
\end{aligned}$$

Se no exemplo das lâmpadas multiplicarmos a constante 2 a cada um dos valores da variável temos 1224, 1966, 1246, 1766, 1332, 1940.

$$\overline{X}^* = \frac{1224 + 1966 + 1246 + 1766 + 1332 + 1940}{6} = \frac{9474}{6} = 1579$$

Utilizando a propriedade,

$$\overline{X}^* = k\overline{X} = 2 \times 789,5 = 1579$$

3. Soma dos desvios
   Seja $(X_1, X_2, X_3, ..., X_n)$ uma amostra aleatória de tamanho n e $\overline{X}$ a média da amostra. Se subtrairmos cada valor da variável X pelar média obtemos os desvios. A soma algébrica dos desvios é igual a zero

$$\begin{aligned}
\frac{\sum_{i=1}^{n}(X_i - \overline{X})}{n} &= \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\overline{X}}{n} \\
&= \frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n}\overline{X}}{n} \\
&= \overline{X} - \frac{n\overline{X}}{n} \\
&= \overline{X} - \overline{X} = 0
\end{aligned}$$

   No exemplo da lampâda, temos:

Amostra	$\overline{X}$	Desvio
612	789,5	-177,5
983	789,5	193,5
623	789,5	-166,5
883	789,5	93,5
666	789,5	-123,5
970	789,5	180,5
	soma dos desvios	0

2.2 Mediana

Num conjunto de dados ordenados, a mediana (M_d) é o valor que deixa metade da freqüência abaixo dele. A mediana, como a média, possui a mesma unidade de cada observação.

A mediana pode ser obtida por meio da expressão:

$$M_d = \begin{cases} X_{\frac{n+1}{2}} & \text{se } n \text{ for ímpar} \\ \\ \frac{X_{\frac{n}{2}} + X_{\frac{n+2}{2}}}{2} & \text{se } n \text{ for par} \end{cases}$$

Exemplo: Considere o conjunto de dados: 5, 2, 6, 13, 9, 15, 10.

Primeiro é necessário ordenar os dados: 2, 5, 6, 9, 10, 13, 15. Como se de uma conjunto com $n = 7$ (ímpar), então:

$$Md = X_{\frac{n+1}{2}} = X_{\frac{7+1}{2}} = X_4$$

Logo a Mediana é igual ao elemento que está na quarta posição do conjunto de dados, assim

$$Md = 9$$

Exemplo: Considere o conjunto de dados: 1, 3, 8, 6, 2, 4.

Primeiro é necessário ordenar os dados: 1, 2, 3, 4, 6, 8. Como se de uma conjunto com $n = 6$ (par), então

$$Md = \frac{X_{\frac{n}{2}} + X_{\frac{n+2}{2}}}{2} = \frac{X_{\frac{6}{2}} + X_{\frac{6+2}{2}}}{2} = \frac{X_3 + X_4}{2}$$

Logo para obter a mediana é necessário obter os elementos que estão na terceira e quarta posição do conjunto de dados, assim:

$$Md = \frac{3+4}{2} = 3,5$$

2.3 Moda

A moda M_o de um conjunto de dados é o valor mais freqüente e também tem a mesma unidade dos dados. Para obter a moda basta observar qual o dado que mais se repete.

Exemplo: No conjunto de dados 7 , 8 , 9 , 10 , 10 , 10 , 11 , 12 a moda é igual a 10, pois é único que se repete.

Exemplo: No conjunto de dados 3 , 5 , 8 , 10 , 12 não apresenta moda. O conjunto é amodal

Exemplo: No conjunto de dados 2 , 3 , 4 , 4 , 4 , 5 , 6 , 7 , 7 , 7 , 8 , 9 temos duas modas: 4 e 7. O conjunto é bimodal.

2.4 Comparação entre Média, Mediana e Moda

- Média
  - Definição: Soma de todos os valores dividido pelo total de elementos do conjunto.
  - Vantagens: Reflete cada valor; Possui propriedades matemáticas atraentes.
  - Limitações: É influenciada porvalores externos.
  - Quando usar:
    1. Deseja-se obter a medida de posição que possui a maior estabilidade;
    2. Houver necessidade de um tratamento algébrico posterior.
- Mediana
  - Definição: Valor que divide o conjunto em duas partes iguais.
  - Vantagens: Menos sensível a valores extremos que a média.
  - Limitações: Difícil de determinar para grande quantidade de dados

- Quando usar:
  1. Deseja-se obter o ponto que divide o conjunto em partes iguais;
  2. Há valores extremos que afetam de maneira acentuada a média.

- Moda
  - Definição: Valor mais freqüente.
  - Vantagens: Valor "típico"; Maior quantidade de valores concentrados neste ponto
  - Limitações: Não se presta a análise matemática; Pode não haver moda para certos conjuntos de dados.
  - Quando usar:
    1. Deseja-se obter uma medida rápida e aproximada da posição;
    2. A medida de posição deve ser o valor mais típico da distribuição.

2.5 Simetria

A determinação das medidas de posição permite discutir sobre a simetria da distribuição dos dados.

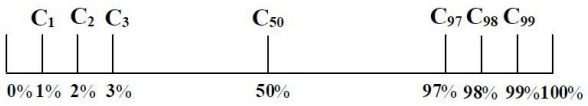
- Distribuição simétrica - $\overline{X} = M_d = M_o$
- Distribuição assimétrica - ocorrem diferenças entre os valores da média, mediana e moda. A assimetria pode ser:
  - à direita - $\overline{X} > M_d > M_o$
  - à esquerda - $\overline{X} < M_d < M_o$

2.6 Separatrizes

Além das medidas de posição que estudamos, há outras que, consideradas individualmente, não são medidas de tendência central, mas estão ligadas à mediana relativamente à sua característica de separar a série em duas partes que apresentam o mesmo número de valores. Essas medidas - os quartis, os decis e os percentis - são, juntamente com a mediana, conhecidas pelo nome genérico de separatrizes.

2.6.1 Percentis ou Centis

São as medidas que dividem a amostra em 100 partes iguais. Assim:



-Calculando o p-ésimo percentil

1. Ordene os dados (Rol)

2. Calcule o índice i

$$i = \left(\frac{p}{100}\right) n,$$

onde p é o percentil de interesse e n é o número de observações.

3. a) Se i **não for um inteiro**, arredonde para cima. O próximo inteiro maior que i denota a posição do p-ésimo percentil.
b) Se i **é um inteiro**, o p-ésimo percentil é a média dos valores de dados nas posições i e $i + 1$.

Exemplo: Salários mensais iniciais para uma amostra de 12 graduados de administração, determinar o 85º percentil.

2350	2450	2550	2380
2255	2210	2390	2630
2440	2825	2420	2380

O primeiro passo a ser dado é o da ordenação dos valores (Rol): 2210, 2255, 2350, 2380, 2380, 2390, 2420, 2440, 2450, 2550, 2630, 2825.

O segundo passo é o cálculo do índice i:

$$i = \left(\frac{p}{100}\right) n = \left(\frac{85}{100}\right) 12 = 10.2$$

Como i não é um inteiro, arredonde para cima. A posição do 85º percentil é o inteiro maior do que 10.2, a 11º posição. Logo, 85º percentil é 2630

- Calcule o 50º percentil.

O primeiro passo a ser dado é o da ordenação dos valores (Rol): 2210, 2255, 2350, 2380, 2380, 2390, 2420, 2440, 2450, 2550, 2630, 2825.

O segundo passo é o cálculo do índice i:

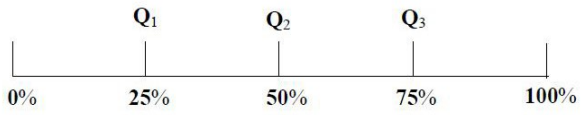
$$i = \left(\frac{p}{100}\right) n = \left(\frac{50}{100}\right) 12 = 6$$

Como i é um inteiro o 50º percentil é a média do 6º com o 7º elementos. Assim,

$$50^{\text{o}} \text{ percentil} = \frac{2390 + 2420}{2} = 2405$$

2.6.2 Quartis

Denominamos quartis os valores de uma série que a dividem em quatro partes iguais.



$Q1$: 1º quartil (ou 25º percentil). Deixa 25% dos elementos antes do seu valor;

$Q2$: 2º quartil (ou 50º percentil). Deixa 50% dos elementos antes do seu valor, coincide com a mediana;

$Q3$: 3º quartil (ou 75º percentil). Deixa 75% dos elementos antes do seu valor (consequentemente, 25% dos elementos acima do seu valor).

Utilizando o exemplo dos salários mensais iniciais para uma amostra de 12 graduados de administração. Calcule:

- Calcule o 1º quartil.

O primeiro passo a ser dado é o da ordenação dos valores (Rol): 2210, 2255, 2350, 2380, 2380, 2390, 2420, 2440, 2450, 2550, 2630, 2825.

O segundo passo é o cálculo do índice i:

$$i = \left(\frac{p}{100}\right) n = \left(\frac{25}{100}\right) 12 = 3$$

Como i é um inteiro o 1º quartil (ou 25º percentil) é a média do 3º com o 4º elementos. Assim,

$$Q_1 = \frac{2350 + 2380}{2} = 2365$$

- Calcule o 3º quartil.

O primeiro passo a ser dado é o da ordenação dos valores (Rol): 2210, 2255, 2350, 2380, 2380, 2390, 2420, 2440, 2450, 2550, 2630, 2825.

O segundo passo é o cálculo do índice i:

$$i = \left(\frac{p}{100}\right) n = \left(\frac{75}{100}\right) 12 = 9$$

Como i é um inteiro o 3º quartil (ou 75º percentil) é a média do 9º com o 10º elementos. Assim,

$$Q_3 = \frac{2450 + 2550}{2} = 2500$$

2.7 Dados agrupados

2.7.1 Média

Quando os dados são agrupados (Distribuição de freqüência) a média é representada por

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

em que

- para variáveis contínuas x_i é o ponto médio da classe

- fa_i é o freqüência absoluta de x_i

A média calculada dos dados originais e dados agrupados podem ser diferentes, devido ao erro de agrupamento. O erro de agrupamento é obtido fazendo a diferença entre o valor obtido pelos dados originais e o valor obtido pelos dados agrupados.

2.7.2 Mediana

Para calcular a mediana em dados agrupados é necessário observar a freqüência acumulada para definir a classe mediana.

A posição da mediana EM_d é definida da seguinte forma

$$EM_d = \begin{cases} \frac{n+1}{2} & \text{se } n \text{ for ímpar} \\ \\ \frac{n}{2} & \text{se } n \text{ for par} \end{cases}$$

Definida a classe mediana utiliza-se a expressão abaixo para obter a mediana

$$M_d = LI_i + \frac{n_1}{n_2}c$$

em que:

- LI_i é o limite inferior da classe mediana
- c é a amplitude da classe mediana
- n_1 é a diferença entre a Posição da mediana e a freqüência acumulada da classe anterior a classe mediana
- n_2 é a freqüência absoluta da classe mediana

2.7.3 Moda

A moda M_o de um conjunto de dados é o valor mais freqüente e também tem a mesma unidade dos dados. Para obter a moda basta observar qual o dado que mais se repete.

Para dados agrupados de variáveis continuas a moda se localiza na classe de maior freqüência (classe modal) e é obtida por meio da expressão:

$$M_o = LI_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}c$$

- LI_i é o limite inferior da classe modal;
- c é a amplitude da classe modal;
- Δ_1 é a diferença da freqüência da classe modal e a freqüência da classe imediatamente anterior;
- Δ_2 é a diferença da freqüência da classe modal e a freqüência da classe imediatamente posterior.

2.7.4 Quartil

A posição da classe quantílica EQ_i é definida da seguinte forma:

$$EQ_i = \frac{in}{4}$$

Definida a classe quartílica utiliza-se a expressão abaixo para obter o quartil

$$Q_i = LI_i + \frac{n_1}{n_2}c$$

em que:

- LI_i é o limite inferior da classe quartílica
- c é a amplitude da classe quartílica
- n_1 é a diferença entre a posição do quartil e a freqüência acumulada da classe anterior a classe quartílica
- n_2 é a freqüência absoluta da classe quartílica

2.7.5 Exemplo

Tabela 1: Dados ordenados, relativos ao tempo em segundos para carga de um aplicativo num sistema compartilhado (30 observações).

6,94	7,27	7,46	7,97	8,03	8,37
8,56	8,66	8,88	8,95	9,30	9,33
9,55	9,76	9,80	9,82	9,98	9,99
10,14	10,19	10,42	10,44	10,66	10,88
10,88	11,16	11,80	11,88	12,25	12,34

Tabela 2: Resumo da distribuição de freqüências, relativa ao ao tempo em segundos para carga de um aplicativo num sistema compartilhado.

Classes			x	Frequencia Absoluta (fa)	$fa \times x$	Frequencia Acumulada (FA)
6,27	$\vdash$	7,62	6,94	3	20,82	3
7,62	$\vdash$	8,97	8,29	7	58,03	10
8,97	$\vdash$	10,32	9,64	10	96,4	20
10,32	$\vdash$	11,67	10,99	6	65,94	26
11,67	$\vdash$	13,02	12,34	4	49,36	30
Total				30	290,55	

Assim,

$$\overline{X} = \frac{\sum_{i=1}^{n} fa_i x_i}{\sum_{i=1}^{n} fa_i} = \frac{290,55}{30} = 9,685 \cong 9,68$$

Para dados agrupados, primeiro vamos obter a classe mediana

$$\frac{n}{2} = \frac{30}{2} = 15$$

Assim a classe mediana é a que contém a freqüência acumulada 15, ou seja é a classe $8,97 \vdash 10,32$. Então temos:

- $LI_i = 8,97$
- c=1,35
- $n_1 = 15 - 10 = 5$
- $n_2 = 10$

Substituindo nas formula, temos

$$M_d = LI_i + \frac{n_1}{n_2}c = 8,97 + \frac{5}{10}1,35 = 8,97 + 0,67 = 9,64$$

Para obter a moda, primeiro vamos obter a classe modal.
A maior freqüência absoluta é 10, assim a classe modal é $8,97 \vdash 10,32$. Assim, temos

$$M_o = LI_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}c$$

- $LI_i = 8,97$;
- $c = 1,35$;
- $\Delta_1 = 10 - 7 = 3$;
- $\Delta_2 = 10 - 6 = 4$

$$M_o = LI_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}c = 8,97 + \frac{3}{3+4}1,35 = 8,97 + 0,58 = 9,55$$

3 Boxplot

O gráfico Boxplot (ou desenho esquemático) é uma análise gráfica que oferece a idéia da posição, dispersão, assimetria, caudas e dados discrepantes. Para construí-lo, desenhamos uma "caixa"com o nível superior dado pelo terceiro quartil (Q_3) e o nível inferior pelo primeiro quartil (Q_1). A mediana (Q_2) é representada por um traço no interior da caixa e segmentos de reta são colocados da caixa até dos limites inferior (LI) e superior (LS), dados por

$$\begin{aligned} LI &= Q_1 - 1.5dq \\ LS &= Q_3 + 1.5dq \end{aligned}$$

em que $dq = Q_3 - Q_1$ denominando diferença quartilica.

Para traçarmos o boxplot utilizamos as seguintes etapas:

- Contruir um retângulo de tal maneira que suas bases têm alturas correspondentes aos primeiro e terceiro quartis da distribuição.
- Cortar o retângulo por um segmento paralelo às bases, na altura correspondente à mediana;
- Traçar um segmento paralelo ao eixo, partindo do ponto médio da base superior do retângulo até o maior valor observado que NÃO supere LS;
- Traçar um segmento paralelo ao eixo, partindo do ponto médio da base inferior do retângulo, até o menor valor que NÃO é menor LI;
- Case tenha valores que superior a LS ou inferior a LI, marcar os pontos, este valores são considerados observações discrepantes.
- Podemos opcionalmente marca o valor da média;

Para o conjunto de dados do tempo de carga de um aplicativo temos:

$$\begin{aligned} Md &= 9,81 \\ Q_1 &= 8,71 \\ Q_3 &= 10,61 \\ dq &= 10,61 - 8,71 = 1,9 \\ LI &= 8,71 - 1,5 \times 1,9 = 5,86 \\ LS &= 10,61 + 1,5 \times 1,9 = 13,46 \end{aligned}$$

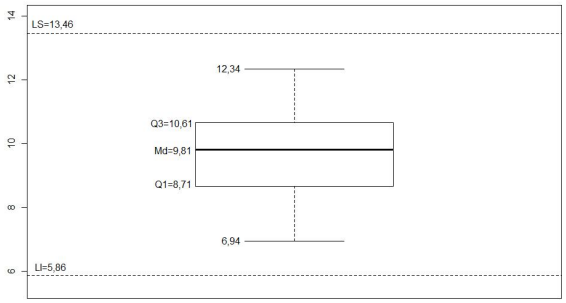


Figura 1: Boxplot para o tempo em segundos para carga de um aplicativo num sistema compartilhado