Aula 02: Medidas Resumo

Estatística e Probabilidades

André Victor Ribeiro Amaral (sala 3029)

avramaral@gmail.com

Distribuição de frequência e Medidas Resumo

Ao estudarmos a distribuição de frequência de uma variável quantitativa − seja apenas de um grupo, seja comparando grupos −, devemos verificar basicamente três características:

- Tendência central: o que é mais frequente?
- Variabilidade
- Forma: a distribuição é simétrica ou assimétrica?

Medidas Resumo

Medidas resumo são, como o nome sugere, medidas que representam um resumo *brusco* das informações trazidas pela amostra; nesse caso, são representadas por um único número.

O interesse é caracterizar o conjunto de dados através de medidas que resumam a informação com a qual se está trabalhando; representado, por exemplo, a **tendencia central** dos dados ou o quanto eles estão dispersos (medida de **variabilidade**).

Medidas de tendência central

Medidas de tendência central fornecem uma ideia do comportamento central dos dados; ou seja, os valores mais comuns na amostra.

Exemplo: média, moda e mediana.

Usualmente se posicionam nas regiões do gráfico com maior frequência.

Média

Denote as n observações que compõem uma **amostra** por $x_1, \cdots, x_n$; então, a **média amostral** é definida por:

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

onde $\bar{x}$ é estimador*.

Denote as N observações que compõem a **população** por $x_1, \cdots, x_N$; então, a **média populacional** é definida por:

$$\mu = \frac{x_1 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N},$$

onde μ é parâmetro*.

Média

Exemplo: determine a média amostral para o seguinte conjunto de dados

$$1,\ 1,\ 3,\ 5,\ 2,\ 1,\ 2,\ 2,\ 3,\ 2$$

Então,

$$\bar{x} = \frac{1+1+3+5+2+1+2+2+3+2}{10} = 2.2$$

Média

Propriedades:

1. Se multiplicarmos (ou dividirmos) todas as observações por uma constante, a média aritmética também fica multiplicada (ou divida) por essa constante.
2. Somando-se (ou subtraindo-se) uma constante a todos os valores de um conjunto de observações, a média aritmética ficará somada (ou subtraída) desta constante

Média

Demonstração das propriedades "1." e "2.":

1. Seja c constante, então

$$\frac{c \cdot x_1 + \cdots + c \cdot x_n}{n} = \frac{c \cdot (x_1 + \cdots + x_n)}{n} = c \cdot \bar{x}.$$

2. Seja c constante, então

$$\frac{(c + x_1) + \cdots + (c + x_n)}{n} = \frac{n \cdot c}{n} + \frac{x_1 + \cdots + x_n}{n} = c + \bar{x}.$$

Tomando $c = \frac{1}{k}$ ou $c = -k$, para k constante, resolve os problemas da divisão e subtração, respectivamente.

Média

Para calcular a média através da tabela de frequência, basta

$$\bar{x} = \frac{n_1 x_1 + \cdots + n_k x_k}{n} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \sum_{i=1}^{k} f_i x_i$$

No exemplo:

Classe	n_i	f_i	$f_{ac}^{(i)}$
1	3	0.3	0.3
2	4	0.4	0.7
3	2	0.2	0.9
5	1	0.1	1.0
total	$n = 10$	1	–

Onde, $\bar{x} = \frac{3 \cdot 1 + 4 \cdot 2 + 2 \cdot 3 + 1 \cdot 5}{10} = 2.2$.

Média – Exercício

Calcule a média amostral para a seguinte tabela de frequência:

Classe (Idade)	n_i	f_i	$f_{ac}^{(i)}$
17	2	0.06	0.06
18	9	0.25	0.31
19	4	0.11	0.42
20	5	0.14	0.56
21	9	0.25	0.81
22	5	0.14	0.95
23	2	0.05	1.00
total	$n = 36$	1	–

Tabela 1: Tabela de frequência (discreto) para conjunto de idades.

Média – Exercício

No nosso exercício,

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{k} n_i x_i}{n} \\ &= \frac{2 \cdot 17 + 9 \cdot 18 + 4 \cdot 19 + 5 \cdot 20 + 9 \cdot 21 + 5 \cdot 22 + 2 \cdot 23}{36} \\ &= \frac{717}{36} \approx 19.9 \text{ anos.}\end{aligned}$$

Média

Similarmente, é possível definirmos uma *espécie de média amostral* a partir de tabelas de frequência de variáveis contínuas.

Assim, tome:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \cdot \mathrm{PM}_i}{n} = \sum_{i=1}^{k} f_i \cdot \mathrm{PM}_i.$$

onde PM_i é o ponto médio da classe; i.e., $\mathrm{PM}_i = \frac{\max_i + \min_i}{2}$.

Média – Exercício

Para a tabela de frequência a seguir, calcule a média amostral aproximada.

Classe (Peso)	n_i	f_i	$f_{ac}^{(i)}$	PM_i
$[50,\ 60)$	7	0.29	0.29	
$[60,\ 70)$	3	0.12	0.42	
$[70,\ 80)$	7	0.29	0.71	
$[80,\ 90)$	2	0.08	0.79	
$[90, 100]$	5	0.21	1.00	
total	$n = 24$	1	–	–

Tabela 2: Tabela de frequência (contínuo) para conjunto de pesos.

Média – Exercício

Uma aproximação da média amostral, nesse caso, pode ser obtida através de:

Classe (Peso)	n_i	f_i	$f_{ac}^{(i)}$	PM_i
$[50,\ 60)$	7	0.29	0.29	55
$[60,\ 70)$	3	0.12	0.42	65
$[70,\ 80)$	7	0.29	0.71	75
$[80,\ 90)$	2	0.08	0.79	85
$[90, 100]$	5	0.21	1.00	95
total	$n = 24$	1	–	–

$$\bar{x} = 0.29 \cdot 55 + 0.12 \cdot 65 + 0.29 \cdot 75 + 0.08 \cdot 85 + 0.21 \cdot 95 = 72.25\text{kg}.$$

Outros tipos de média

Média aparada: é obtida eliminando-se do conjunto de dados as m maiores e as m menores observações da amostra. Geralmente, $m = 5\%$. Em seguida, calcula-se a média aritmética com a amostra restante.

No nosso exemplo,

$$1,\ 1,\ 3,\ 5,\ 2,\ 1,\ 2,\ 2,\ 3,\ 2$$

teremos $m = 0.05 \cdot n = 0.05 \cdot 10 = 0.5$. Nesse caso, podemos tomar $m = 1$; logo, nossa nova amostra será composta por:

$$1,\ 1,\ 2,\ 2,\ 2,\ 2,\ 3,\ 3.$$

Assim, $\bar{x} = \frac{16}{8} = 2$.

Outros tipos de média

Ainda em relação à **média aparada**, ela oferece a vantagem de desconsiderar valores atípicos no cálculo da média; entretanto, sua aplicação depende da interpretação do problema considerado.

Alternativamente, temos a **média ponderada**, que é definida por:

$$\bar{x}_P = \frac{\sum_{i=1}^{n} p_i \cdot x_i}{\sum_{i=1}^{n} p_i},$$

onde p_i é o peso da i-ésima observação x_i.

Ela se aplica em casos nos quais os valores variam em grau de importância; nesse caso, vamos querer ponderá-los adequadamente.

Outros tipos de média – Exercício

Considere a seguinte tabela de notas de um(a) aluno(a) de Ensino Fundamental:

Disciplina	Nota	Peso
Língua Portuguesa	10	3
Matemática	8	3
Ciências	9	2
Geografia	7	2
História	9	2
Inglês	9	1

Tabela 3: Tabela de notas (com pesos).

Calcule a nota média ponderada pelo(a) aluno(a).

Outros tipos de média – Exercício

Para calcular a média ponderada nesse caso, basta

$$\begin{aligned}\bar{x}_P &= \frac{\sum_{i=1}^{n} p_i \cdot x_i}{\sum_{i=1}^{n} p_i} \\ &= \frac{3 \cdot 10 + 3 \cdot 8 + 2 \cdot 9 + 2 \cdot 7 + 2 \cdot 9 + 1 \cdot 9}{3 + 3 + 2 + 2 + 2 + 1} \\ &= \frac{113}{13} \approx 8.7 \text{ pontos.}\end{aligned}$$

Moda

A **moda** de uma amostra é aquele valor que ocorre com mais frequência; ou seja, aquele que mais se repete.

Observações:

- Se dois valores ocorrem com a mesma frequência máxima, a variável é bimodal;
- Se mais de dois valores ocorrem com a mesma frequência máxima, a variável é multimodal; e
- Quando nenhum valor se repete, a variável não tem moda.

Exemplos:

1. Amostra$_1$: 1, 1, 1, 2, 2, 2, 2, 3, 3, 5; Aqui, $Mo = 2$.
2. Amostra$_2$: 1, 1, 1, 2, 2, 2, 3, 3, 3, 5; Aqui, $Mo = 1, 2$ e 3.

Mediana

A **mediana**, representada por Md, é o valor que ocupa a posição central dos dados *ordenados*.

Definição: a mediana é **qualquer** valor tal que 50% dos observações são menores ou iguais a ele e 50% das observações são maiores ou iguais a ele.

Como regra para cálculo da mediana, podemos adotar:

- se n é ímpar, a mediana será o valor que ocupa a posição do meio dentre no conjunto de dados *ordenado*.
- se n é par, a mediana será o ponto médio entre os dois valores que o ocupam as posições centrais no conjunto de dados *ordenado*.

Mediana – Exercício

Exercício: Para os dois conjuntos de dados abaixo, determine a mediana.

1. Conjunto$_1$: 3, 2, 3, 5, 6, 12, 11, 3, 6
2. Conjunto$_2$: 102, 100, 5, 1000, 97, 1050

Mediana – Exercício

Exercício: Para os dois conjuntos de dados abaixo, determine a mediana.

1. Conjunto$_1$: 3, 2, 3, 5, 6, 12, 11, 3, 6
2. Conjunto$_2$: 102, 100, 5, 1000, 97, 1050

Resposta:

1. Conjunto$_1$ ordenado: 2, 3, 3, 3, 5, 6, 6, 11, 12; logo, $Md = 5$.
2. Conjunto$_2$ ordenado: 5, 97, 100, 102, 1050, 1000; logo, $Md = \frac{100+102}{2} = 101$.

Comparação de medidas de tendência central

Média

- Vantagem: leva em conta todos os valores da amostra e é utilizada em muitos métodos estatísticos.
- Desvantagem: é afetada por valores extremos.

Moda

- Vantagem: não é afetada por valores extremos.
- Desvantagem: não leva em conta todos os valores da amostra; além disso, é raramente utilizada e pode nem existir.

Média

- Vantagem: é utilizada com frequência e não é afetada por valores extremos.
- Desvantagem: não leva em conta todos os valores da amostra.

Percentil de ordem $\alpha\%$

Percentil de ordem $\alpha\%$: é definido como qualquer número tal que $\alpha\%$ das observações são menores ou iguais ao valor do percentil e $(100 - \alpha\%)$ das observações são maiores ou iguais ao valor do percentil.

Observações:

- Os percentis de ordem 25%, 50% e 75% são denotados por primeiro, segundo e terceiro quartil, respectivamente;
- O percentil de ordem 50% é a mediana; e
- Os percentis de ordem 10%, 20%, $\cdots$, 90% também são chamados de decis.

Percentil de ordem $\alpha\%$

Exemplo: Suponha que, durante 10 dias, o tempo (em minutos) que um indivíduo esperou o ônibus para o trabalho foi

$$13, 2, 5, 6, 9, 8, 10, 9, 2, 3.$$

Ordenando os dados, temos

$$2, 2, 3, 5, 6, 8, 9, 9, 10, 13.$$

O percentil de ordem 90% pode ser 11.5 (Notação alternativa: $P_{90\%} = 11.5$).

Interpretação: 90% das vezes, o tempo de espera pelo ônibus foi menor ou igual a 11.5 minutos.

Medida de variabilidade

Medidas **de variabilidade** (ou **de dispersão**) são medidas que tentam quantificar o "espalhamento" dos dados

Nesse sentido, as principais medidas de variabilidade são **variância** e **desvio padrão**. Quanto maior qualquer um desses valores, maior a variação dos dados em torno da média.

Variância e **desvio padrão** para população:

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

onde σ^2 é a variância e σ é o desvio padrão da **população**.

Medida de variabilidade

Similarmente, a **variância** e **desvio padrão** na amostra podem ser definidos por:

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

onde s^2 é a variância e s é o desvio padrão da **amostra**.

[1]O termo "$n-1$" tem a ver com o fato de que s^2 é estimador não viciado para σ^2.

Observação: a interpretabilidade da variância é prejudicada, pois a unidade de σ^2 (ou s^2) é o quadrado da unidade de x_i.

[1]A demostração desse fato pode ser encontrada aqui.

Medida de variabilidade – Exercício

Em relação ao conjunto de dados abaixo, que representa a concentração de álcool no sangue (em ml/l) de 15 motoristas envolvidos em acidentes de fatais e que foram condenados à prisão, determine a **média**, **variância** e **desvio padrão** da amostra.

$$0.27,\ 0.17,\ 0.17,\ 0.16,\ 0.13,\ 0.24,\ 0.29,\ 0.24,$$
$$0.14,\ 0.16,\ 0.12,\ 0.16,\ 0.21,\ 0.17,\ 0.18.$$

A média é determinada por

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0.27 + 0.17 + \cdots + 0.18}{15} \approx 0.187 \text{ ml/l}.$$

Para calcular variância e desvio padrão, podemos organizar o nosso cálculo na tabela a seguir.

Medida de variabilidade – Exercício

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.27	$(0.27 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.17	$(0.17 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.17	$(0.17 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.16	$(0.16 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.13	$(0.13 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.24	$(0.24 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.29	$(0.29 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.24	$(0.24 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.14	$(0.14 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.16	$(0.16 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.12	$(0.12 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.16	$(0.16 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.21	$(0.21 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.17	$(0.17 - 0.187) = \square$	$(\ \square\ )^2 = \square$
0.18	$(0.18 - 0.187) = \square$	$(\ \square\ )^2 = \square$
	$\sum$	$\square$

Medida de variabilidade – Exercício

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.27	$(0.27 - 0.187) = 0.083$	$(0.083)^2 = 0.0069$
0.17	$(0.17 - 0.187) = -0.017$	$(-0.017)^2 = 0.0003$
0.17	$(0.17 - 0.187) = -0.017$	$(-0.017)^2 = 0.0003$
0.16	$(0.16 - 0.187) = -0.027$	$(-0.027)^2 = 0.0007$
0.13	$(0.13 - 0.187) = -0.057$	$(-0.057)^2 = 0.0032$
0.24	$(0.24 - 0.187) = 0.053$	$(0.053)^2 = 0.0028$
0.29	$(0.29 - 0.187) = 0.103$	$(0.103)^2 = 0.0106$
0.24	$(0.24 - 0.187) = 0.053$	$(0.053)^2 = 0.0028$
0.14	$(0.14 - 0.187) = -0.047$	$(-0.047)^2 = 0.0022$
0.16	$(0.16 - 0.187) = -0.027$	$(-0.027)^2 = 0.0007$
0.12	$(0.12 - 0.187) = -0.067$	$(-0.067)^2 = 0.0045$
0.16	$(0.16 - 0.187) = -0.027$	$(-0.027)^2 = 0.0007$
0.21	$(0.21 - 0.187) = 0.023$	$(0.023)^2 = 0.0005$
0.17	$(0.17 - 0.187) = -0.017$	$(-0.017)^2 = 0.0003$
0.18	$(0.18 - 0.187) = -0.007$	$(-0.007)^2 = 0.0001$
	$\sum$	0.0367

Medida de variabilidade – Exercício

Então, a variância amostral é:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{0.0367}{14} = 0.0026 \text{ (ml/l)}^2.$$

O desvio padrão é dado por:

$$s = \sqrt{s^2} = \sqrt{0.0026} = 0.0512 \text{ ml/l}.$$

Coeficiente de variação

O **coeficiente de variação** para um conjunto de dados amostrais (ou populacionais), expresso como percentual, descreve o desvio padrão relativo à média, e é definido por:

$$CV = \frac{s}{\bar{x}} \times 100\%, \text{ para amostra.}$$

$$CV = \frac{\sigma}{\mu} \times 100\%, \text{ para população.}$$

O coeficiente de variação não tem unidade de medida e, por isso, é útil para se comparar a variabilidade de grupos que tem unidade de medida diferentes.

Coeficiente de variação – Exercício

Considerando o tempo (em minutos) de espera na fila para os bancos A e B, calcule o coeficiente de variação.

Banco A: 6.5, 6.6, 7.7, 7.1, 6.7, 7.4

Banco B: 5.4, 6.7, 4.2, 7.7, 5.8

Coeficiente de variação – Exercício

Considerando o tempo (em minutos) de espera na fila para os bancos A e B, calcule o coeficiente de variação.

Banco A: 6.5, 6.6, 7.7, 7.1, 6.7, 7.4
Banco B: 5.4, 6.7, 4.2, 7.7, 5.8

Aqui, temos $\bar{x}_A = 7.0$, $\bar{x}_B = 5.96$, $s_A = 0.482$ e $s_B = 1.324$; assim:

$$CV_A = \frac{s_A}{\bar{x}_A} \times 100\% = \frac{0.428}{7.00} \times 100\% = 6.90\%$$

$$CV_B = \frac{s_B}{\bar{x}_B} \times 100\% = \frac{1.324}{5.96} \times 100\% = 22.0\%$$

Medida de assimetria

Uma distribuição de dados é assimétrica quando se estende mais para um lado do que para o outro.

Uma das medida para esse tipo de comportamento é chamada de "**Coeficiente de Assimetria de Pearson**", definida por:

$$A_p = \frac{3(\bar{x} - Md)}{s}.$$

Aqui, se $A_p \geq 1$ ou $A_p \leq -1$, então os dados podem ser considerados fortemente assimétricos.

A ideia vem do fato de que, se a distribuição é assimétrica à esquerda, então $\bar{x} < Md$; em contrapartida, se a distribuição é assimétrica à direita, então $\bar{x} > Md$.

Medida de assimetria – Exercício

Para o exemplo da quantidade de álcool no sangue, no Slide 27, calcule o Coeficiente de Assimetria de Person. Lembre-se de:

$$0.27, 0.17, 0.17, 0.16, 0.13, 0.24, 0.29, 0.24,$$
$$0.14, 0.16, 0.12, 0.16, 0.21, 0.17, 0.18.$$

E que $\bar{x} = 0.187$ ml/l e $s = 0.0512$ ml/l.

Medida de assimetria – Exercício

Para o exemplo da quantidade de álcool no sangue, no Slide 27, calcule o Coeficiente de Assimetria de Person. Lembre-se de:

$$0.27, 0.17, 0.17, 0.16, 0.13, 0.24, 0.29, 0.24,$$
$$0.14, 0.16, 0.12, 0.16, 0.21, 0.17, 0.18.$$

E que $\bar{x} = 0.187$ ml/l e $s = 0.0512$ ml/l.

Ordenando o conjunto de dados, temos:

$$0.12, 0.13, 0.14, 0.16, 0.16, 0.16, 0.17, 0.17,$$
$$0.17, 0.18, 0.21, 0.24, 0.24, 0.27, 0.29.$$

Logo, $Md = 0.17$ ml/l. Dessa forma, $A_p = \frac{3(0.187-0.17)}{0.0512} \approx 1$.

Medida de assimetria – Exercício

Plotando o histograma do conjunto de dados que acabamos de analisar, temos:

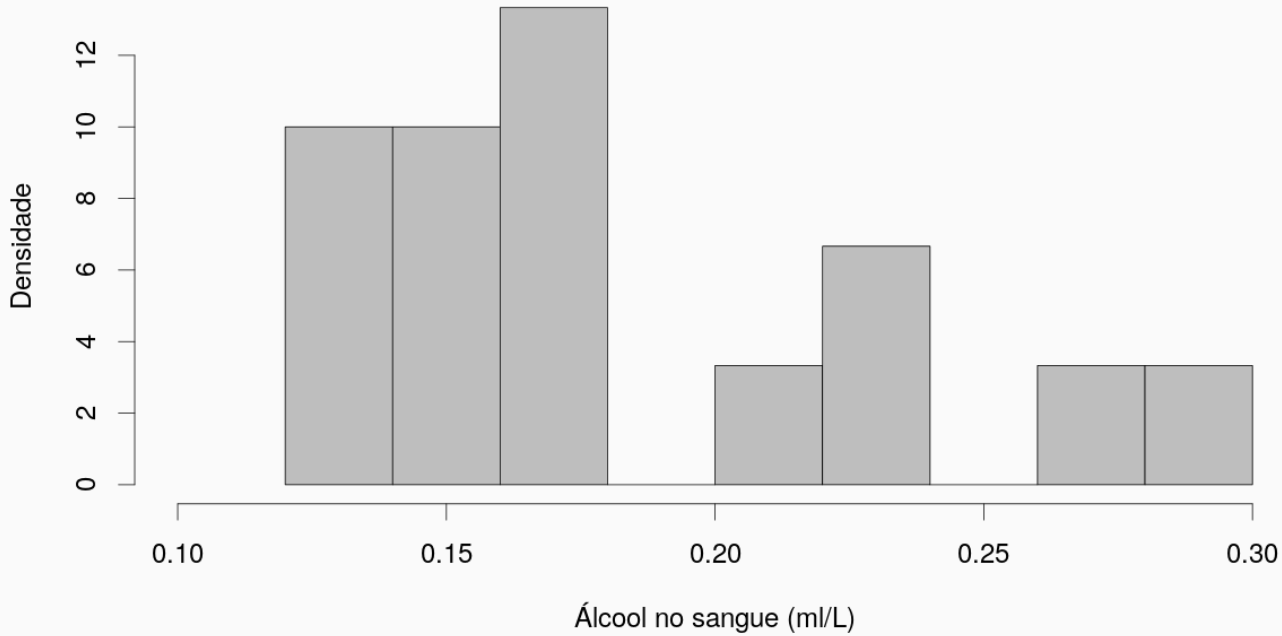

Figura 1: Histograma quantidade de álcool no sangue (em ml/l).