



ME414 - Estatística para Experimentalistas

Parte 2

Análise Descritiva

Análise Descritiva Univariada

Na aula passada trabalhamos na análise descritiva univariada para variáveis categóricas (nominal e ordinal) e quantitativas (discreta e contínua).

Relembrando, a análise descritiva univariada consiste em:

- classificar a variável quanto a seu tipo
- obter tabela, gráfico e/ou medidas resumo apropriados

Na aula de hoje, falaremos sobre **estatísticas sumárias ou medidas resumo**, tanto de posição quanto de dispersão.



Medidas Resumo

Vimos na aula anterior como usar gráficos e tabelas de frequência para resumir os dados.

Podemos também usar **estatísticas sumárias** (ou **medidas resumo**): quantidades numéricas calculadas a partir dos dados.

Por exemplo, podemos estar interessados em encontrar qual seria um valor “típico” do conjunto de dados.

Podemos então usar uma estatística que descreva o centro da distribuição dos dados.

Objetivo: resumir os dados, através de valores que representem o conjunto de dados em relação à alguma característica (posição, dispersão).



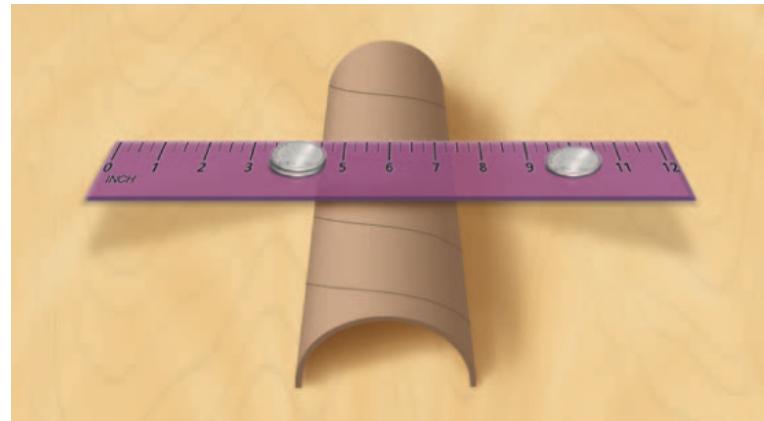
Medidas de Posição Central

Média Aritmética

Se x_1, x_2, \dots, x_n são as n observações, a média aritmética é:

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

A média pode ser interpretada como o ponto de equilíbrio de uma distribuição.



Exemplo: Cereais matinais

Temos cereais matinais de várias marcas e observamos a quantidade de calorias e carboidratos em porções de 30g.

Calorias e Carboidratos (Porções de 30g)

| Cereal | Calorias | Carboidratos |
|-------------|----------|--------------|
| Sucrilhos | 109 | 26.0 |
| All Bran | 81 | 13.5 |
| Nesfit | 102 | 21.0 |
| Nescau | 115 | 23.0 |
| Snow | 113 | 25.0 |
| Crunch | 119 | 23.0 |
| Moça | 113 | 25.0 |
| Fibra Mais | 84 | 15.0 |
| Froot Loops | 113 | 25.0 |



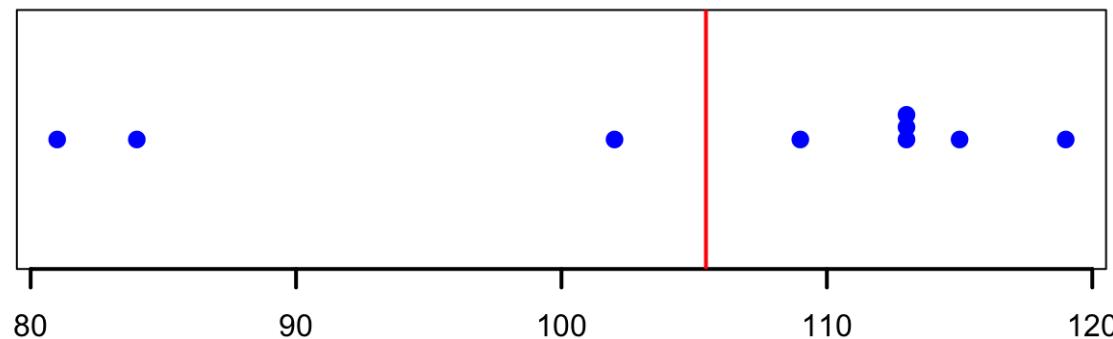
Exemplo: Cereais matinais

Calorias dos 9 cereais: 109, 81, 102, 115, 113, 119, 113, 84, 113

x_i : calorias do cereal i .

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{9} \sum_{i=1}^9 x_i = 105.44$$

No gráfico de pontos abaixo, os pontos azuis representam as observações e a linha vermelha representa a média.



Mediana

Mediana: valor que separa os dados em dois grupos de tamanhos iguais, ou seja, 50% das observações em cada, de acordo com seus valores ordenados.

Para determinar a mediana (também conhecida como Q_2), ordene as n observações: $x_{(1)}, x_{(2)}, \dots, x_{(k)}, \dots, x_{(n)}$

- Se n é ímpar: a mediana é o valor do meio, na sequência ordenada.
- Se n é par: a mediana, por convenção, é a média aritmética das duas observações que caem no meio da sequência ordenada.

A fórmula da mediana pode ser escrita como:

$$Q_2 = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ é ímpar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ é par} \end{cases}$$

Exemplo: Cereais matinais

Calorias dos 9 cereais:

109, 81, 102, 115, 113, 119, 113, 84, 113

Calorias em ordem crescente:

81, 84, 102, 109, 113, 113, 113, 115, 119

A mediana é 5^a observação, ou seja, 113.



Se descartássemos o maior valor, 119, teríamos oito observações e aí a mediana seria:

$$\text{mediana} = \frac{109 + 113}{2} = 111.$$

Moda

A moda é o valor com maior número de ocorrências nos dados.

Calorias dos 9 cereais:

109, 81, 102, 115, 113, 119, 113, 84, 113

Tabela de frequências:

| | | | | | | |
|----|----|-----|-----|-----|-----|-----|
| 81 | 84 | 102 | 109 | 113 | 115 | 119 |
| 1 | 1 | 1 | 1 | 3 | 1 | 1 |

Portanto, a moda de calorias dos cereais é 113.



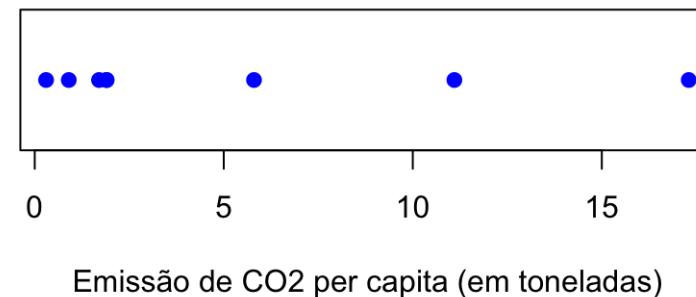
Exemplo: Emissão de CO_2

Veja a tabela com dados da emissão de CO_2 per capita (em toneladas) para 8 países, em 2009.

| País | Emissão CO2 |
|------------|-------------|
| China | 5.8 |
| Índia | 1.7 |
| EUA | 17.3 |
| Indonésia | 1.9 |
| Brasil | 1.9 |
| Rússia | 11.1 |
| Paquistão | 0.9 |
| Bangladesh | 0.3 |

Fonte:

<http://data.worldbank.org>



$$\text{Média: } \bar{x} = \frac{40.9}{8} \approx 5.11$$

Dados Ordenados: 0.3, 0.9, 1.7, 1.9, 1.9, 5.8, 11.1, 17.3
Mediana = 1.9

A mediana é bem menor do que a média.

Exemplo: Emissão de CO_2

Veja a tabela com dados da emissão de CO_2 per capita (em toneladas) para 8 países, em 2009.

| País | Emissão CO2 |
|------------|-------------|
| China | 5.8 |
| Índia | 1.7 |
| EUA | 17.3 |
| Indonésia | 1.9 |
| Brasil | 1.9 |
| Rússia | 11.1 |
| Paquistão | 0.9 |
| Bangladesh | 0.3 |

Fonte:
<http://data.worldbank.org>

Se desconsiderarmos os EUA (maior valor):

$$\text{Média: } \bar{x} = \frac{23.6}{7} \approx 3.37$$

Dados Ordenados: 0.3, 0.9, 1.7, 1.9, 1.9, 5.8, 11.1
Mediana = 1.9

Mediana é menos afetada por valores muito extremos (muito diferentes do resto das observações) que a média.

Dizemos que a mediana é mais **robusta** que a média.

Exemplo: SleepStudy

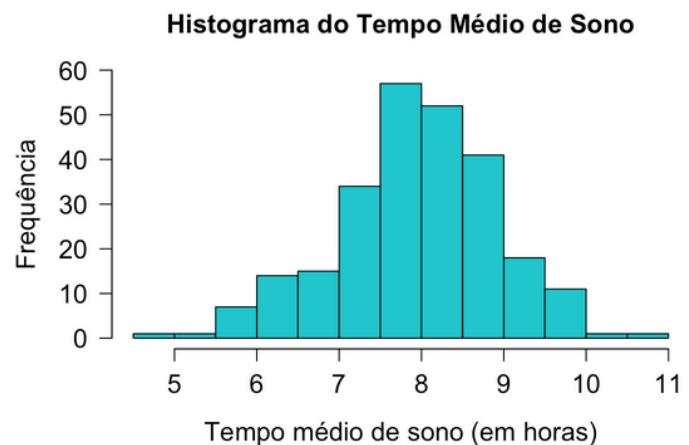
Vamos voltar no exemplo `SleepStudy`: amostra de 253 alunos universitários que fizeram testes para medir função cognitiva, além de outras informações sobre hábitos relacionados ao sono.

Considere a variável `AverageSleep`, a média de horas de sono em todos os dias.

Como são muitos valores, podemos usar um software (R, Excel, etc) para calcular a média e mediana:

Média = 7.97 e Mediana = 8

Será que o fato desses valores serem próximos é uma coincidência?



Exemplo: número de casamentos

Os dados abaixo referem-se ao número de vezes que homens e mulheres se casaram.

| Número de Casamentos | Mulheres | Homens |
|----------------------|----------|--------|
| 0 | 5861 | 7074 |
| 1 | 2773 | 1561 |
| 2 | 105 | 43 |
| Total | 8739 | 8678 |

Você acha que existe diferença entre homens e mulheres quanto ao número de casamentos?

Qual medida de posição você usaria para apresentar a diferença entre homens e mulheres: média, mediana ou moda?

Moda

A moda entre os homens é: 0

A moda entre as mulheres é: 0

Fonte: <http://www.census.gov/prod/2002pubs/p70-80.pdf>

Exemplo: número de casamentos

| Número de Casamentos | Mulheres | Homens |
|----------------------|----------|--------|
| 0 | 5861 | 7074 |
| 1 | 2773 | 1561 |
| 2 | 105 | 43 |
| Total | 8739 | 8678 |

Mediana

Para as mulheres, a amostra ordenada é:

$$\underbrace{000 \dots 0}_{5861 \text{ 0's}} \quad \underbrace{111 \dots 1}_{2773 \text{ 1's}} \quad \underbrace{222 \dots 2}_{105 \text{ 2's}}$$

Como $n = 8739$ é ímpar, a observação central está na posição $(1 + 8739)/2 = 4370$ e essa observação é 0. Portanto, a mediana é 0 para as mulheres.

Para os homens, analogamente, a mediana também é 0.

Para dados discretos com poucos valores diferentes, a mediana ignora muita informação.

Exemplo: número de casamentos

| Número de Casamentos | Mulheres | Homens |
|----------------------|----------|--------|
| 0 | 5861 | 7074 |
| 1 | 2773 | 1561 |
| 2 | 105 | 43 |
| Total | 8739 | 8678 |

Média

Para as mulheres, a média é:

$$\bar{x} = \frac{0 \times 5861 + 1 \times 2773 + 2 \times 105}{8739} = 0.34$$

Para os homens, a média é:

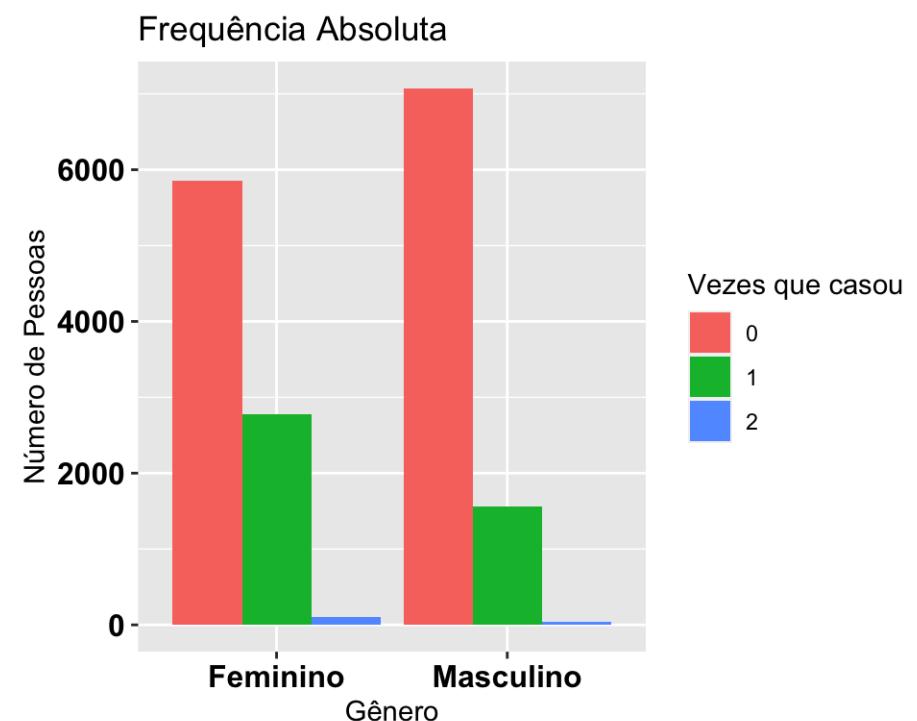
$$\bar{x} = \frac{0 \times 7074 + 1 \times 1561 + 2 \times 43}{8678} = 0.19$$

Nesse caso, temos que a média é a medida de posição que consegue diferenciar homens e mulheres quanto ao número de casamentos.

Exemplo: número de casamentos

Como o número de casamentos assume apenas os valores 0, 1 e 2, podemos apresentar os dados usando um gráficos de barra.

| Número de Casamentos | Mulheres | Homens |
|----------------------|----------|--------|
| 0 | 5861 | 7074 |
| 1 | 2773 | 1561 |
| 2 | 105 | 43 |
| Total | 8739 | 8678 |



Mediana é resistente a observações discrepantes

Considere os três conjuntos de dados abaixo:

$$A : 8, 9, 10, 11, 12$$

$$B : 8, 9, 10, 11, 100$$

$$C : 8, 9, 10, 11, 1000$$

Para cada conjunto, calcule a média e a mediana e compare-as.

Média de A : 10

Mediana de A : 10

Média de B : 27.6

Mediana de B : 10

Média de C : 207.6

Mediana de C : 10

Exemplo: Transporte

Uma empresária cuja empresa está localizada na Av. Paulista, em São Paulo, está preocupada com a quantidade de gasolina gasta pelos seus funcionários. Ela quer promover o uso de transporte público entre seus funcionários. Ela decide investigar a extensão, em km, do trajeto percorrido por cada funcionário caso usassem transporte público durante um dia típico.

Para seus 10 funcionários, os valores são:

1, 1, 4, 1, 1, 1, 10, 1, 6, 1

Encontre a média, a mediana e a moda.

Média é 2.7.

Valores ordenados: 1,1,1,1,1,1,1,4,6,10.

Mediana e moda são iguais a 1.

Exemplo: Transporte

A empresária acabou de contratar um novo funcionário. Ele percorre 90 km em transporte público. Recalcule a média e a mediana.

1, 1, 4, 1, 1, 1, 10, 1, 6, 1, 90

Valores ordenados: 1, 1, 1, 1, 1, 1, 1, 4, 6, 10, 90.

Mediana é 1.

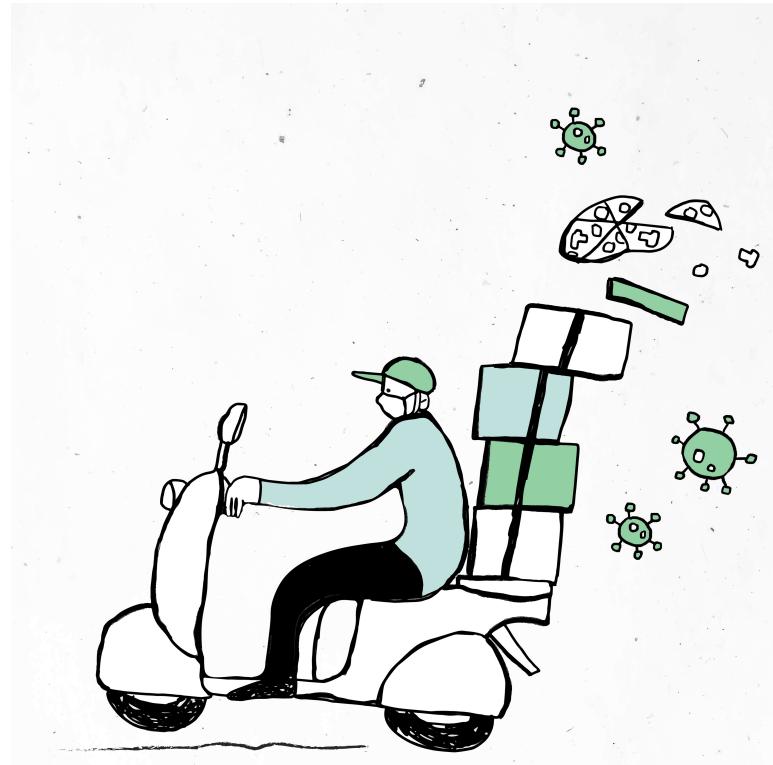
Média é 10.64.

Qual medida de posição representa melhor a distância do grupo de funcionários?

Exemplo: Acidentes com Moto

Dados: entrevistas com 60 pessoas, em que cada uma relata o número de acidentes com moto que sofreu no último ano.

Por que a média seria provavelmente mais útil do que a mediana para resumir os dados?



Exemplo: Salários

A **média** salarial anual em 1998 nos EUA para pessoas com ensino superior era \$528.200.

A **mediana** do salário anual em 1998 nos EUA para pessoas com ensino superior era \$146.400.

Reflita sobre as estatísticas sumárias apresentadas e responda:

1. Por que a média e a mediana diferem tanto?
2. Qual medida de posição você acredita que retrata de maneira mais realística um salário típico de pessoas com ensino superior nos EUA em 1998?

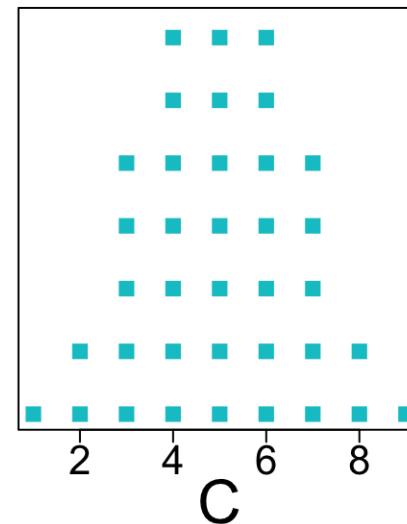
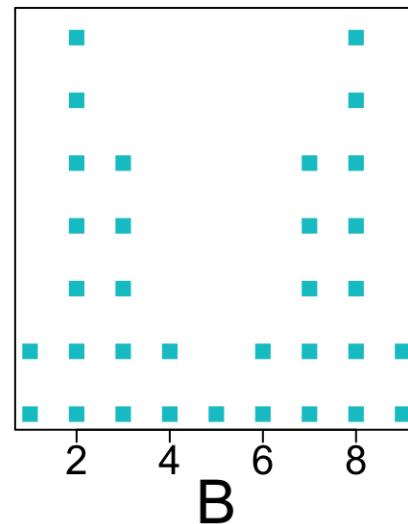
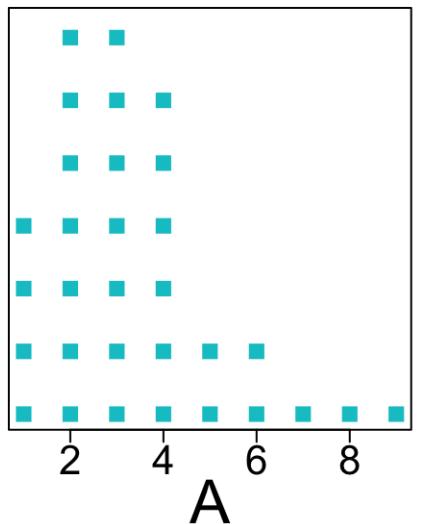
Exemplo: Sindicato

O sindicato dos trabalhadores está reivindicando aumento de salário em uma certa fábrica.

Explique por que o sindicato poderia usar a mediana dos salários de todos os empregados para justificar um aumento, enquanto que o gerente da fábrica poderia usar a média para argumentar que um aumento não é necessário?

Média, mediana e a distribuição dos dados

A figura a seguir mostra gráficos para três conjuntos de dados: A, B e C.



O que você esperaria da relação entre média e mediana para esses dados?

Média, mediana e a distribuição dos dados

Para cada uma das distribuições (A, B, C), qual medida seria maior: média ou mediana?

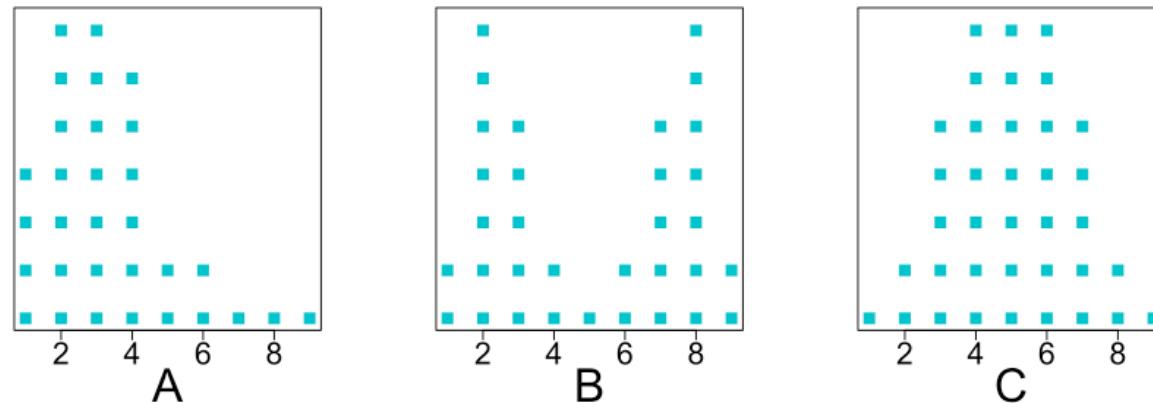
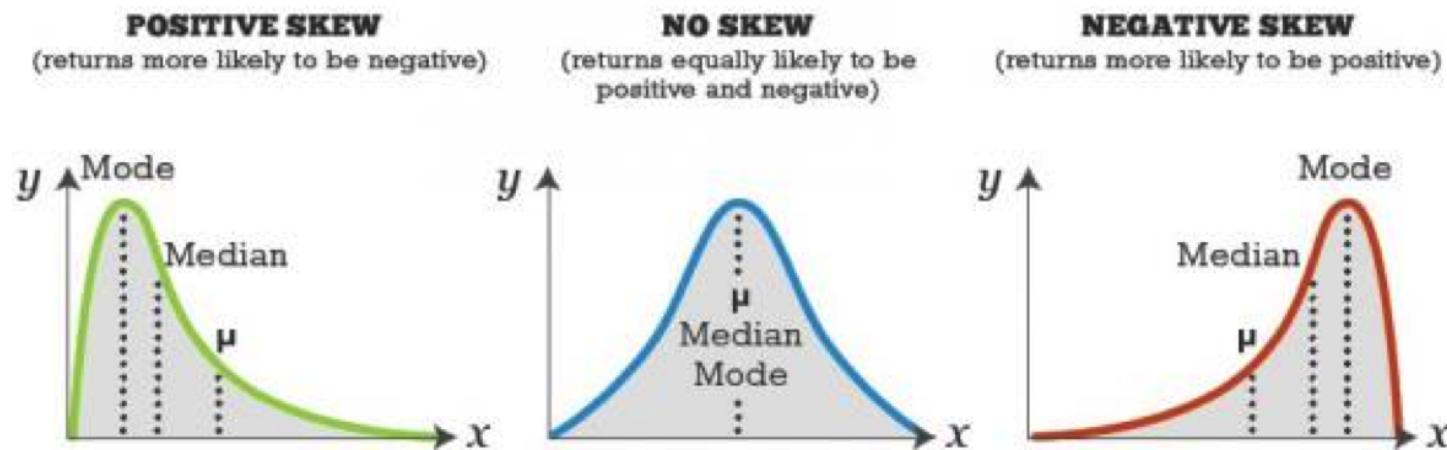


Gráfico A: média é 3.36, mediana é 3.

Gráfico B: média é 5, mediana é 5.

Gráfico C: média é 5, mediana é 5.

Assimetria (Caso Unimodal)



Se os dados são simétricos, a média coincide com a mediana e a moda.

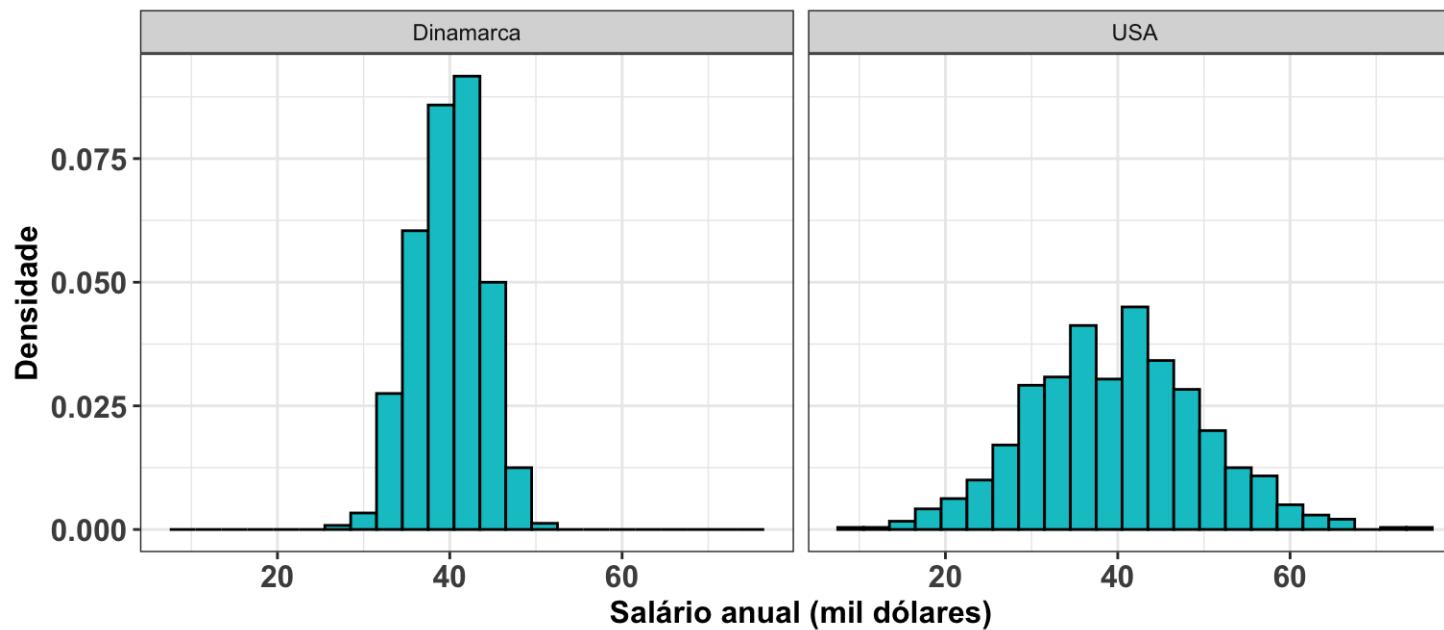
Assimetria à direita (positiva): Média > Mediana > Moda

Assimetria à esquerda (negativa): Média < Mediana < Moda

Medidas de Dispersão

Exemplo: Salário professor de música

Salário anual hipotético de professores de música na Dinamarca (esquerda) e nos EUA (direita).



Média salarial Dinamarca: 40.02. Média salarial EUA: 39.87.

Amplitude

Uma medida de dispersão é **amplitude**: a diferença entre o maior e o menor valor observado na amostra.

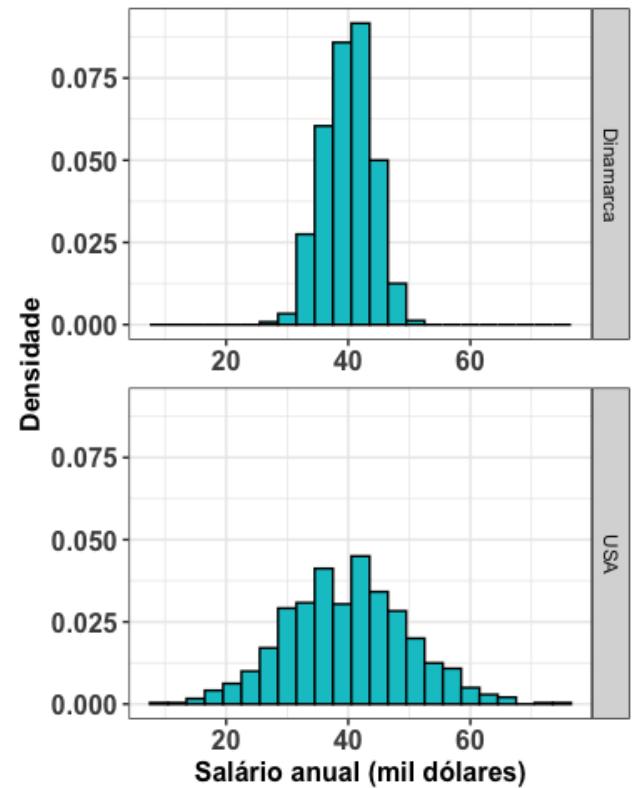
Na Dinamarca:

- Salários variam de 27 a 52.
- Amplitude = $52 - 27 = 25$.

Nos EUA:

- Salários variam de 9 a 75.
- Amplitude = $75 - 9 = 66$.

Problema com a amplitude: utiliza apenas duas observações (a máxima e a mínima).



Medidas de Dispersão

Considere dois conjuntos de dados:

$$A = \{1, 2, 5, 6, 6\} \text{ e } B = \{-40, 0, 5, 20, 35\}$$

Ambos com média 4 e mediana 5.

No entanto, claramente temos que os valores de B são mais dispersos do que em A .

Que medida podemos usar para considerar essa característica dos dados?

Medidas de Dispersão

Podemos observar quão afastadas de uma determinada medida de posição estão as observações.

Desvio de uma observação x_i da média \bar{x} é a diferença entre a observação e a média dos dados: $(x_i - \bar{x})$.

- O desvio é negativo quando a observação tem valor menor do que a média.
- O desvio é positivo quando a observação tem valor maior do que a média.

Estamos interessados nos desvios de todos os pontos x_i 's, então poderia-se propor a seguinte medida de dispersão: $\sum_{i=1}^n (x_i - \bar{x})$.

Qual o problema?

- A média representa o ponto de balanço dos dados, então os desvios irão se contrabalancear, ou seja: $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

Medidas de Dispersão

Além do mais, uma medida de dispersão onde os desvios positivos e negativos se cancelam, não seria útil.

Queremos que se leve em conta cada desvio, independente do sinal.

Alternativas:

$$\sum_{i=1}^n |x_i - \bar{x}| \quad \text{ou} \quad \sum_{i=1}^n (x_i - \bar{x})^2$$

Ambas alternativas evitam que desvios iguais em módulo, mas com sinais opostos, se anulem.

Nota: Veja que $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$.

Variância e Desvio padrão

A média dos desvios ao quadrado é denominada **variância**:

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

Desvio padrão é a raiz quadrada da variância:

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

Interpretação: distância típica entre uma observação e a média dos dados.

Quanto maior s , maior a dispersão dos dados.

Exemplo A

Conjunto de dados $A : \{1, 2, 5, 6, 6\}$.

A média é $\bar{x} = \frac{20}{5} = 4$.

| x_i | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ |
|-------|-----------------|---------------------|
| 1 | -3 | 9 |
| 2 | -2 | 4 |
| 5 | 1 | 1 |
| 6 | 2 | 4 |
| 6 | 2 | 4 |

Então, a variância é:

$$s^2 = \frac{9 + 4 + 1 + 4 + 4}{5 - 1} = 5.5,$$

e o desvio padrão:

$$s = \sqrt{s^2} = \sqrt{5.5} = 2.35.$$

Exemplo B

Conjunto de dados $B : \{-40, 0, 5, 20, 35\}$.

A média é $\bar{x} = \frac{20}{5} = 4$.

| x_i | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ |
|-------|-----------------|---------------------|
| -40 | -44 | 1936 |
| 0 | -4 | 16 |
| 5 | 1 | 1 |
| 20 | 16 | 256 |
| 35 | 31 | 961 |

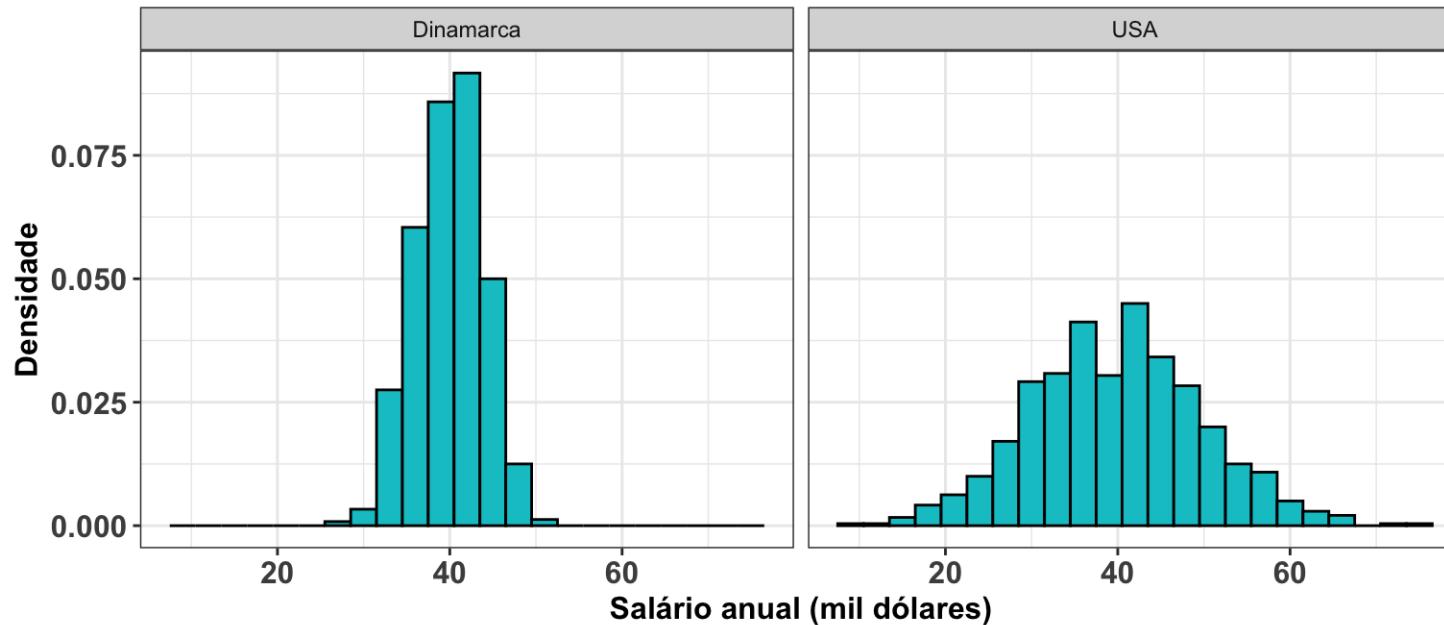
Então, a variância é:

$$s^2 = \frac{1936 + 16 + 1 + 256 + 961}{5 - 1} = 792.5,$$

e o desvio padrão:

$$s = \sqrt{s^2} = \sqrt{792.5} = 28.15.$$

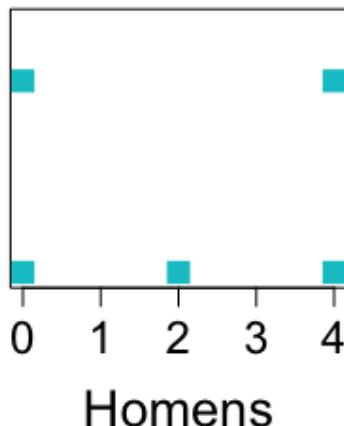
Exemplo: Salário professor de música



Salários na Dinamarca: média = 40.02 e variância= 15.76.

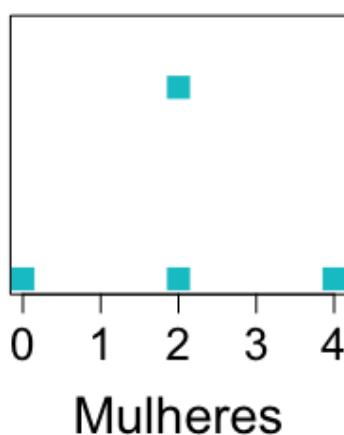
Salários nos EUA: média = 39.87 e variância= 99.5.

Exemplo: “Qual o número ideal de filhos?”



Homens: 0, 0, 0, 4, 4, 4, 2
Mulheres: 0, 2, 2, 2, 2, 2, 4

Média: 2 (para ambos os sexos)
Amplitude: 4 (para ambos os sexos)



Para homens, desvio típico da média parece estar em torno de 2

Desvio padrão: $s = 2$

Para mulheres, desvio típico da média é menor do que o dos homens, pois a grande maioria das observações coincide com a própria média.

Desvio padrão: $s = 1.15$

Exemplo: Prova 1 de ME414

A primeira prova de ME414 teve um total de 100 pontos. Suponha que a média tenha sido 80.

Qual seria um valor plausível para o desvio padrão das notas da classe? s : 0, 10 ou 50.

- $s = 0$: todos os alunos tiraram a mesma nota.
- $s = 50$: uma nota típica da classe estaria 50 pontos distante da média, ou seja, 30 ou 130 pontos.
- $s = 10$: notas típicas seriam de 70 ou 90.



Leitura

- Ross: seções 3.1, 3.2, 3.3, 3.4, 3.5
- Khan Academy: Por que $n - 1$ no cálculo de s^2 ?
- Simulação mostrando o porque de usarmos $n - 1$ no cálculo de s^2

Slides produzidos pelos professores:

- Samara Kiihl
- Tatiana Benaglia
- Benilton Carvalho
- Larissa Matos

