Considerações sobre as Medidas de Dispersão

# Aula	13
☑ Preparada	☑
☑ Revisada	☑
☑ Lecionada	☑

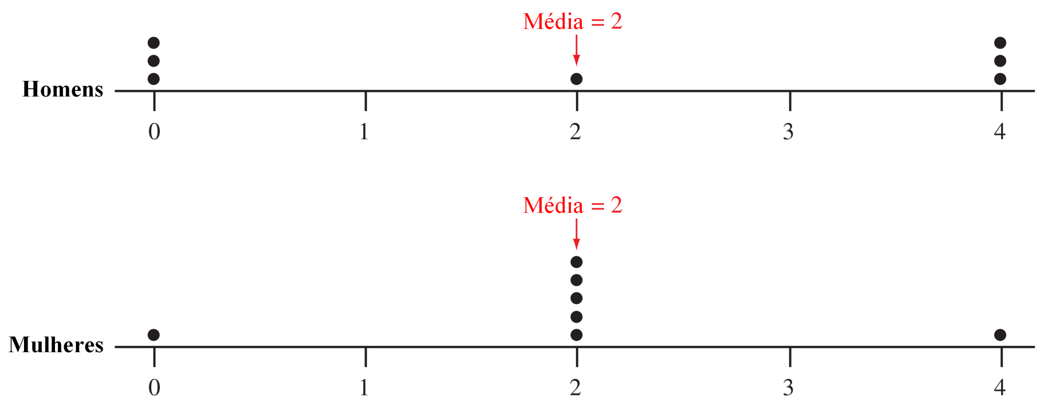
▼ Vamos pensar um pouco… 🤔

▼ Num primeiro exemplo:

Uma pesquisa foi feita numa amostra com 7 mulheres e 7 homens, perguntando o número de filhos considerado ideal:

- Mulheres: 0, 2, 2, 2, 2, 2, 4
- Homens: 0, 0, 0, 2, 4, 4, 4
- Qual é a média nestes dois grupos?

▼ Uma visualização gráfica sempre ajuda…



▼ Quais são a média e o desvio padrão dos dois grupos?

- A média é igual a 2 nos dois grupos
- O desvio padrão é diferente nos dois grupos, $s_{Mulheres}$ = 1,15 e s_{Homens} = 2.

▼ Mas afinal, o que isso significa?

- A questão aqui é que as observações do grupo de homens estão mais dispersas mesmo tendo a mesma média que as observações do grupo de mulheres.
- Isso fica mais visível quando referenciamos os desvios padrão em relação às respectivas médias através dos coeficientes de variação: CV(Mulheres) = 58% e CV (Homens) = 100%

▼ Quais são a amplitude, a mediana e a distância interquartílica dos 2 grupos?

- Amplitude: é igual a 4 nos dois grupos.
- Mediana: é igual a 2 nos dois grupos.
- Primeiro Quartil (Q1): vale 2 no grupo das mulheres e 0 no grupo dos homens
- Terceiro Quartil (Q3): vale 2 no grupo das mulheres e 4 no grupo dos homens
- Distância Interquartílica (IQR): vale 0 no grupo das mulheres e 4 no grupo dos homens
- A mensagem é a mesma, as observações no grupo das mulheres estão mais concentradas do que no grupo dos homens…

▼ Agora imagine que um dado estivesse errado…

A última observação foi coletada errada… Seguem os valores corretos:

- Mulheres: 0, 2, 2, 2, 2, 2, 18
- Homens: 0, 0, 0, 2, 4, 4, 18

▼ Vamos recalcular as medidas de posição e de dispersão novamente?

- Mínimo (M): 0
- Máximo (M): 18
- Amplitude (M): 18
- Média (M): 4
- Desvio Padrão (M): 6,22
- Coeficiente de Variação (M): 155%
- Q1 (M): 2
- Q2 (M): 2
- Q3 (M): 2
- IQR (M): 0

- Mínimo (H): 0
- Máximo (H): 18
- Amplitude (H): 18
- Média (H): 4
- Desvio Padrão (H): 6,43
- Coeficiente de Variação (H): 161%
- Q1 (H): 0
- Q2 (H): 2
- Q3 (H): 4
- IQR (H): 4

▼ O que mudou e o que não mudou?

- As observações revisadas contém um valor extremo que afetou a média, o desvio padrão e a amplitude.
- No entanto, o valor extremo não afetou os quartis nem a distância interquartílica.

▼ Mas afinal, o que isso significa?

Significa que mínimo, máximo e amplitude são totalmente sensíveis a valores extremos.

Significa que média e desvio padrão são sensíveis a valores extremos.

Significa que quartis e a distância interquartílica não são sensíveis a valores extremos.

▼ Afinal, quando usamos cada medida de dispersão?

Ao invés de responder esta pergunta, cabe questionar a própria pergunta…

Uma vez que usamos um computador para fazer essas contas, por que não calcular todas elas? É fácil escrever um código que faça isso. Mas…

▼ A questão fundamental é:

O computador não sabe interpretar os números…

Você precisa fazer essa parte!

Ele calcula, você interpreta!

Aproveite esta diferença para detectar valores extremos!

▼ Recordando: Aplicação Prática em modelos de regressão…

MAE (*Mean Absolute Error* [Erro Absoluto Médio])

MSE (*Mean Squared Error* [Erro Quadrático Médio])

RMSE (*Root Mean Squared Error* [Raiz do Erro Quadrático Médio])

▼ Partindo da média, qual a relação do desvio padrão com a proporção dos dados?

▼ Caso Específico:

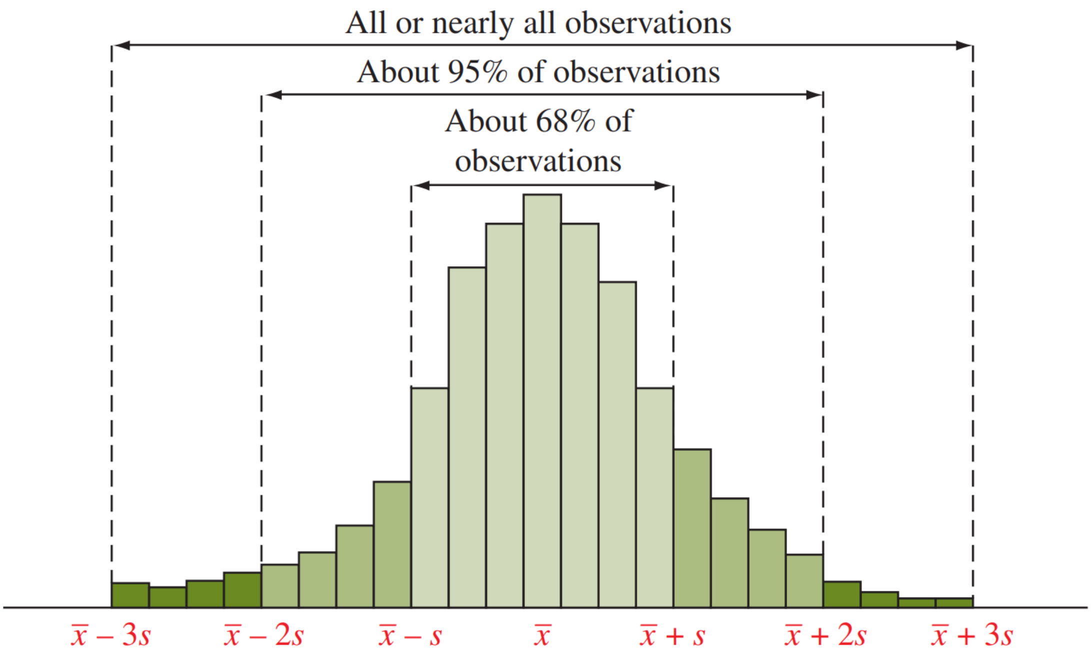
Suponha o caso específico de uma distribuição unimodal que seja aproximadamente simétrica com o formato aproximado de um sino, como a figura abaixo.

Neste caso o desvio padrão assume uma interpretação mais objetiva. A partir da média e do desvio padrão, podemos deduzir intervalos com percentuais aproximados dos dados.

É a chamada regra empírica (tem este nome por que isto é observado na prática), que diz que:

SE a distribuição dos dados tem o formato aproximado de uma curva de sino, então aproximadamente:

- 68% das observações ficam no intervalo de 1 desvio padrão da média, ou seja, entre os valores de $\overline{x} - s$ e $\overline{x} + s$ (escrevemos $\overline{x} \pm s$)
- 95% das observações ficam no intervalo de 2 desvios padrão da média ($\overline{x} \pm 2s$)
- Quase todas as observações ficam no intervalo de 3 desvios padrão da média ($\overline{x} \pm 3s$)



Caso Geral:

O russo *Pafnuti Chebyshev* (1821-1894) propôs o Teorema de Chebyshev, segundo o qual:

A proporção de qualquer conjunto de dados que se situe a K desvios padrão da média é sempre, no mínimo, igual a

$$Proporção = 1 - \frac{1}{K^2}$$

onde K é qualquer número positivo > 1.

- Para K = 1,5 isso quer dizer que, *pelo menos*, 55,56% dos valores se localizam dentro do intervalo de 1,5 desvios padrão da média.
- Para K = 2,0 isso quer dizer que, *pelo menos*, 75,00% dos valores se localizam dentro do intervalo de 2 desvios padrão da média.
- Para K = 2,5 isso quer dizer que, *pelo menos*, 84,00% dos valores se localizam dentro do intervalo de 2,5 desvios padrão da média.
- Para K = 3,0 isso quer dizer que, *pelo menos*, 88,89% dos valores se localizam dentro do intervalo de 3 desvios padrão da média.
- Para K = 3,5 isso quer dizer que, *pelo menos*, 91,84% dos valores se localizam dentro do intervalo de 3,5 desvios padrão da média.
- Para K = 4,0 isso quer dizer que, *pelo menos*, 93,75% dos valores se localizam dentro do intervalo de 4 desvios padrão da média.
- Para K = 4,5 isso quer dizer que, *pelo menos*, 95,06% dos valores se localizam dentro do intervalo de 4,5 desvios padrão da média.
- Para K = 5,0 isso quer dizer que, *pelo menos*, 96,00% dos valores se localizam dentro do intervalo de 5 desvios padrão da média.
- Para K = 5,5 isso quer dizer que, *pelo menos*, 96,69% dos valores se localizam dentro do intervalo de 5,5 desvios padrão da média.
- Para K = 6,0 isso quer dizer que, *pelo menos*, 97,22% dos valores se localizam dentro do intervalo de 6 desvios padrão da média.