MAT02018 - Estatística Descritiva

Medidas de variabilidade

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2022

Introdução

Introdução

- As medidas de tendência central são insuficientes para representar adequadamente conjuntos de dados, pois nada revelam sobre sua **variabilidade**.
- Tome como exemplo as notas de 5 avaliações realizadas por 2 alun@s.
  - **Alun@ A:** 6, 6, 6, 6, 6.
    - **Total de pontos:** 30; **Média:** 6.
  - Alun@ B: 7, 5, 6, 4, 8.
    - **Total de pontos:** 30; **Média:** 6.

Introdução

- Para mostrar a diversidade de desempenho de ambos, necessita-se de um valor que meça a **dispersão** ou a **variabilidade** dos valores nos dois casos.
- Nestas notas de aula vamos apresentar algumas das medidas mais utilizadas para descrever a variabilidade de um conjunto de dados.

Amplitude de variação

Amplitude de variação

- A medida mais simples de dispersão é a **amplitude de variação (a)**, que é a diferença entre os valores extremos do conjunto de dados.

$$a = Max(x) - Min(x),$$

em que $Min(x)$ e $Max(x)$ representam os valores de **mínimo** e **máximo** de um conjunto de dados, respectivamente.

Amplitude de variação

- **Exemplo:** suponha que observamos a **largura da pétala** (em centímetros) de 10 flores

1.3	0.2	0.2	1.9	2.0
1.3	1.8	0.4	0.6	0.1

Amplitude de variação

- Note que,
  - $Min(x) = 0.1$;
  - $Max(x) = 2$;
- Logo, a amplitude é $a = 2 - 0.1 = 1.9$

Amplitude de variação

Comentário

- Os valores de mínimo e máximo de um conjunto de dados podem ser utilizados para a **verificação de inconsistências nos dados** com respeito a variável de interesse.
  - Exemplos:
    - $Max(\text{idade}) = 170$ (anos).
    - $Min(\text{temperatura}) = -20^{\circ}$ (na série histórica de temperaturas na cidade de Porto Alegre).

Amplitude de variação

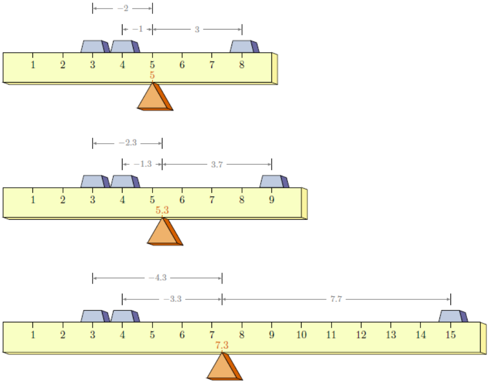
Duas limitações

1. A amplitude de variação só utiliza os valores extremos do conjunto de dados.
2. Quando avaliada em amostras, frequentemente *subestima* a amplitude populacional.

Variância

Variância

- As principais medidas de dispersão envolvem os **desvios em relação a média:** $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$.
- Ou seja, os desvios da média são obtidos pela subtração de $\bar{x}$ de cada uma das n observações da amostra, conforme mostra a figura a seguir.



Variância

- Um desvio será positivo se a observação for maior que a média (à direita da média no eixo das medidas) e negativo se a observação for menor que a média.
- Se todos os desvios forem pequenos em magnitude, todos os x_i estarão próximos à média e haverá pouca dispersão.
- Por outro lado, se alguns desvios forem grandes, alguns x_i estarão distantes de $\bar{x}$, indicando maior dispersão.

Variância

- Uma forma simples de combinar os desvios em uma única quantidade é calcular a sua média.
  - No entanto, temos que a soma dos desvios é igual a zero, conforme mostramos a seguir

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x}) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \\
&= \sum_{i=1}^{n} x_i - n\bar{x} \\
&= \sum_{i=1}^{n} x_i - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = 0.
\end{aligned}
$$

Variância

- Como podemos evitar desvios negativos e positivos ao neutralizar um ao outro quando eles são combinados?
- Uma possibilidade é trabalhar com os valores absolutos e calcular o \structure{desvio médio absoluto $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$.
  - Porém, a operação em valor absoluto conduz a dificuldades teóricas, e uma alternativa é considerar o \structure{quadrado dos desvios $(x_i - \bar{x})^2$ (também conhecidos como desvios quadráticos).
  - E dessa forma temos a definição da **variância**.

Variância

- A **variância** é a média dos desvios quadráticos com respeito a média.
- É representada pelo símbolo σ^2 **na população** e por s^2 **na amostra**.

Observação

Note que, por definição (a média de dos quadrados dos desvios), **a variância é uma medida não negativa**.

Variância

- Na população calculamos a variância da seguinte forma:

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

- Na amostra, calculamos a variância da seguinte forma:

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

Variância

- **Exemplo:** considere mais uma vez valores observados da largura da pétala em centímetros (de 4 flores)

1.3	1.3	0.2	1.8

Variância

- Note que $\bar{x} = 1.15$, e assim a variância da largura da pétala é

$$
\begin{aligned}
s^2 &= \frac{(1.3-1.15)^2+(1.3-1.15)^2+(0.2-1.15)^2+(1.8-1.15)^2}{4-1} \\
&= \frac{(0.15)^2+(0.15)^2+(-0.95)^2+(0.65)^2}{3} \\
&= \frac{1.37}{3} \approx 0.46cm^2.
\end{aligned}
$$

Variância

Observações

1. Note que unidade de medida da variável em questão fica expressa ao quadrado na variância.
2. A variância é uma medida de variabilidade em torno da média.
   - Se todas as observações são iguais, não há variabilidade nos dados e $s^2 = 0$.
   - Se os dados são muito dispersos com respeito a média, então a variabilidade é grande e s^2 assume um valor alto.
   - Se os dados são pouco dispersos em torno da média, então a variabilidade é pequena e s^2 assume um valor baixo.

Variância

- Uma forma alternativa de calcularmos a variância é dada pela expressão[1].

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}.$$

[1] **Sua vez:** demonstre o resultado da expressão alternativa. Dica: desenvolva o quadrado da soma, reescreva o somatório e perceba que $\sum_{i=1}^{n} x_i = n\bar{x}$ e $n\bar{x}^2 = (\sum_{i=1}^{n} x_i)^2/n$

Variância

- Esta expressão nos auxilia no cálculo de s^2 quando estamos usando uma planilha eletrônica que comporta fórmulas, ou um software como R que realiza operações vetoriais.

-	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	x_i^2
	7	1	1	49
	5	-1	1	25
	6	0	0	36
	4	-2	4	16
	8	2	4	64
Soma	30	0	10	190

Variância

- Assim, temos
  - $\bar{x} = 30/5 = 6$;
  - $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{10}{4} = 2,5$.
- Ou, pela fórmula alternativa,
  - $s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{190 - \frac{30^2}{5}}{4} = 2,5$.

Desvio padrão

Desvio padrão

- Uma dificuldade com a variância, como medida descritiva da dispersão, é o fato de não poder ser apresentada com a **mesma unidade** com que a variável foi medida.
- A solução é extrair a raiz quadrada positiva da variância, já que, com isso, se volta à unidade original da variável.
- Essa nova medida de variabilidade é denominada **desvio padrão**.

Desvio padrão

- O desvio padrão populacional é definido como

$$\sigma = \sqrt{\sigma^2}.$$

- E o desvio padrão amostral[2] é definido como

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

- **Sua vez:** calcule o desvio padrão para os exemplos anteriores.

[2] Também é denotado por *DP* (de desvio padrão) ou *SD* (do inglês *standard deviation*).

Coeficiente de variação

Coeficiente de variação

- Quando se analisa a mesma variável em duas amostras, pode-se comparar os desvios padrão observados e verificar onde a variação (variabilidade) é maior.

Exemplo (espessura das sementes)

- Amostra 1 apresenta desvio padrão igual 1,29mm;
- Amostra 2 apresenta desvio padrão igual 0,51mm;
  - Logo a Amostra 2 apresenta uma menor variabilidade da espessura das sementes em relação a Amostra 1.

Coeficiente de variação

- No entanto, o mesmo não pode ser feito em se tratando de variáveis que foram mensuradas em diferentes unidades de medida.
- Se as sementes do exemplo anterior foram também pesadas, e o desvio padrão foi 0,009g para a Amostra 1, **não se pode afirmar** que o peso das sementes é uma característica com **menor dispersão** ("menos variável") do que a sua espessura.

Coeficiente de variação

- Para comparar variabilidades, neste caso, deve-se usar o **coeficiente de variação (CV)**, que é uma medida de dispersão **independente da unidade de mensuração** da variável.
- O coeficiente de variação é definido como a razão entre o desvio padrão e a média da variável

$$CV = \frac{s}{\bar{x}} \quad \text{ou} \quad CV = 100\% \times \frac{s}{\bar{x}}.$$

Coeficiente de variação

Exemplo (continuação)

- Suponha que a média da espessura das sementes tenha sido observada em 5 e a média do peso das sementes tenha sido 0,05.
- Para a espessura, temos que $CV_{esp} = \frac{1{,}29}{5} = 0,258$, e para o peso, temos que $CV_{peso} = \frac{0{,}009}{0{,}05} = 0,18$.
- Concluímos que o peso possui menor variabilidade que a espessura neste conjunto de dados.

Para casa

1. Resolver os exercícios 1 a 4, 6 a 8 do Capítulo 9.7 do livro **Fundamentos de Estatística**[3] (disponível no Sabi+).
2. Para o seu levantamento estatístico, calcule a amplitude de variação, a variância, o desvio padrão e coeficiente de variação, de acordo com a classificação das variáveis. Compartilhe no Fórum Geral do Moodle.

[3] Vieira, S. **Fundamentos de Estatística**, Atlas, 2019, p. 135-136.

Próxima aula

- Medidas de variabilidade (continuação):
  - Amplitude entre quartis;
    - Separatrizes: quantis, quartis, decis, percentis;
  - Gráfico de *boxplot*.
  - ComplementaR.

Por hoje é só!

Bons estudos!