Análise exploratória IV

Resumos numéricos - medidas de dispersão

Prof. Me. Lineu Alberto Cavazani de Freitas

CE003 – Estatística II

Departamento de Estatística
Laboratório de Estatística e Geoinformação

Resumos numéricos

- Uma forma de resumir a informação contida em um conjunto de dados é por meio dos **resumos numéricos**.
- Resumos numéricos são basicamente **números que resumem números**.
- Os dois principais grupos são as medidas de **posição** (central e relativa) e **dispersão**.
- Existem outros conjuntos de medidas, como as medidas de **forma** e também as de **relação/associação**.

Medidas de dispersão

Medidas de dispersão

- As medidas de dispersão são utilizadas para expressar informações como o **domínio** da variável, grau de **dispersão** ao redor do centro (**variabilidade**), e também **distanciamento** dos valores com relação ao centro.

- Algumas medidas possíveis são
  - Amplitude.
  - Desvio absluto (médio ou mediano).
  - Variância.
  - Desvio padrão.
  - Coeficiente de variação.

Medidas de dispersão

- Em geral usamos uma **medida de posição central**, que nos dá uma ideia de centro dos dados.
- Mas conjuntos de dados com diferentes valores podem gerar as mesmas medidas de posição.
- E mesmo com medidas de posição idêncitas, um pode ser **mais disperso** que o outro.
- Portanto **complementamos a informação** a respeito do centro **com uma medida de dispersão**, que nos dá uma noção de quão dispersos são os dados.
- Outra utilidade das medidas de dispersão é expressar o domínio da variável.

Amplitude total

- Diferença entre o **maior** e o **menor** valor da variável.
- Sensível a valores extremos.
- Usa apenas duas medidas.

$$Amp = max(y) - min(y) = y(n) - y(1)$$

Amplitude total

Exemplo

- Retomando o problema das notas de 10 alunos, em que as notas obtidas foram:

$$60;\ 65;\ 77;\ 95;\ 56;\ 94;\ 97;\ 81;\ 80;\ 48$$

- A amplitude é dada pelo maior menos o menor valor:

$$Amp = 97 - 48 = 49$$

Desvio absoluto médio

- Um **desvio absoluto médio** é uma medida de distância da observação para uma medida de posição central.
- Podemos usar como referência a **média** ou a **mediana**.
- Tomamos todos os desvios absolutos.
- Calculamos a média.

$$\text{desvio médio} = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \overline{y})|$$

$$\text{desvio mediano} = \frac{1}{n}\sum_{i=1}^{n}|(y_i - md)|$$

Desvio

Exemplo

- Retomando o problema das notas de 10 alunos, em que as notas obtidas foram:

$$60;\ 65;\ 77;\ 95;\ 56;\ 94;\ 97;\ 81;\ 80;\ 48$$

- A média é $\overline{y} = 75{,}3$ e a mediana é $md = 78{,}5$.
- Obtenha o desvio médio e mediano.

Desvio

Exemplo - desvio médio

$$\text{desvio médio} = \frac{1}{10}\left(|(60-75{,}3)| + |(65-75{,}3)| \ldots + |(80-75{,}3)| + |(48-75{,}3)|\right)$$

$$\text{desvio médio} = \frac{1}{10}(15{,}3 + 10{,}3 \ldots + 4{,}7 + 27{,}3) = 14{,}44$$

Desvio

Exemplo - desvio mediano

$$\text{desvio mediano} = \frac{1}{10}\left(|(60 - 78{,}5)| + |(65 - 78{,}5)| \ldots + |(80 - 78{,}5)| + |(48 - 78{,}5)|\right)$$

$$\text{desvio mediano} = \frac{1}{10}\left(18{,}5 + 13{,}5 \ldots + 1{,}5 + 30{,}5\right) = 14{,}1$$

Variância e Desvio padrão

- Em vez dos desvios, usa a **soma dos quadrados dos desvios** em relação à média.

$$s^2 = \text{Var}(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right)$$

- Variância populacional (σ^2): usa apenas n no demominador e é usada quando temos todos os elementos da população. Caso contrário, calculamos sempre a estimativa amostral (s^2).
- Para ter uma medida de dispersão com a mesma unidade de medida dos dados originais definiu-se o desvio-padrão como a raiz quadrada da variância.

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

Regra empírica variância e desvio padrão

Quando a distribuição dos dados é simétrica sabemos que:

- Aproximadamente 68% das observações estão entre mais ou menos um desvio padrão.
- Aproximadamente 95% das observações estão entre mais ou menos dois desvios padrões.
- Aproximadamente 100% das observações estão entre mais ou menos três desvios padrões.

Variância e desvio padrão

Exemplo

- Retomando o problema das notas de 10 alunos, em que as notas obtidas foram:

$$60;\ 65;\ 77;\ 95;\ 56;\ 94;\ 97;\ 81;\ 80;\ 48$$

- A média é $\overline{y} = 75{,}3$.
- Obtenha o variância e desvio padrão.

Variância e desvio padrão

Exemplo

- Primeira maneira:

$$s^2 = \text{Var}(y) = \frac{1}{10-1}\left((60-75{,}3)^2 + (65-75{,}3)^2 + \ldots + (80-75{,}3)^2 + (48-75{,}3)^2\right)$$

$$s^2 = \text{Var}(y) = \frac{1}{9}\left((-15{,}3)^2 + (-10{,}3)^2 + \ldots + (4{,}7)^2 + (-27{,}3)^2\right)$$

$$s^2 = \text{Var}(y) = \frac{1}{9}(234{,}09 + 106{,}09 + \ldots + 22{,}09 + 745{,}29) = 302{,}68$$

$$s = \sqrt{s^2} = \sqrt{302{,}68} = 17{,}4$$

Variância e desvio padrão

Exemplo

- Segunda maneira:

$$s^2 = \text{Var}(y) = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right)$$

$$s^2 = \text{Var}(y) = \frac{1}{9}\left(59425 - \frac{753^2}{10}\right) = \frac{1}{9}(59425 - 56700.9) = 302{,}68$$

$$s = \sqrt{s^2} = \sqrt{302{,}68} = 17{,}4$$

Coeficiente de variação

- Medida de variabilidade relativa à média.
- Quociente do desvio-padrão pela média.
- **Medida adimensional**, geralmente apresentada na forma de porcentagem.
- Permite comparar a variabilidade de variáveis de diferentes naturezas

$$\text{CV} = 100 \cdot \frac{s}{\overline{y}}$$

Coeficiente de variação

Exemplo

- Retomando o problema das notas de 10 alunos, em que as notas obtidas foram:

$$60;\ 65;\ 77;\ 95;\ 56;\ 94;\ 97;\ 81;\ 80;\ 48$$

- A média é $\overline{y} = 75{,}3$ e o desvio padrão é $s = 17{,}4$.
- Obtenha o coeficiente de variação.

$$\text{CV} = 100 \cdot \frac{17{,}4}{75{,}3} = 23{,}11$$

Dispersão para variáveis qualitativas

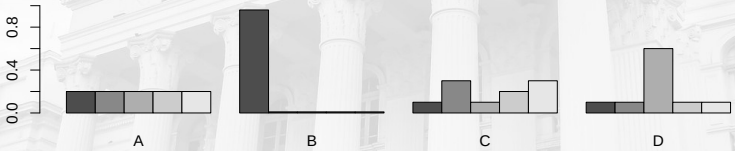
- Para variáveis qualitativas a moda é a única medida de posição que faz sentido.
- Como medida de dispersão, a ideia de entropia pode ser usada.
- Uma proposta, chamada de índice de Shannon é dada por:

$$H = \sum_{i=1}^{S} p_i ln(p_i)$$

- Em que S representa o número de categorias da variável e p_i representa a frequência relativa associada à categoria i.
- Quanto mais distante de 0 for o valor de H, mais heterogênea é a variável.

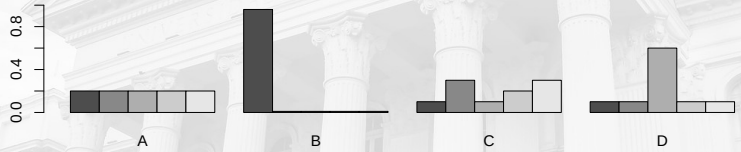
Dispersão para variáveis qualitativas

Qual é o mais homogêneo? Qual é o mais heterogêneo?



Dispersão para variáveis qualitativas

Qual é o mais homogêneo? Qual é o mais heterogêneo?



	A	B	C	D
H	1.609	0.223	1.505	1.228

Desvio, variância, desvio padrão, coeficiente de variação

- A amplitude é simples de calcular, mas é influenciada por valores extremos.
- Os desvios absolutos (médio ou mediano) são menos influenciados por valores extremos.
- Variância e desvio padrão são influenciados por valores extremos mas tem propriedades favoráveis.
- O coeficiente de variação permite comparar a variabilidade de variáveis em diferentes escalas.
- Para variáveis qualitativas existem medidas específicas, como o índice de Shannon.

O que foi visto:

- Medidas de dispersão.

Próximos assuntos:

- Análises bivariadas.
  - Qualitativa x qualitativa.
  - Quantitativa x quantitativa.
  - Quantitativa x qualitativa.