Análise exploratória

Resumos numéricos - medidas de dispersão

Prof. Me. Lineu Alberto Cavazani de Freitas

Departamento de Estatística
Laboratório de Estatística e Geoinformação

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Análise exploratória

- Parte primordial de qualquer análise estatística é chamada **análise descritiva** ou **exploratória**.
- Consiste basicamente de **tabelas**, **resumos numéricos** e **análises gráficas** das variáveis disponíveis em um conjunto de dados.
- Trata-se de uma etapa de extrema importância e deve preceder qualquer análise mais sofisticada.
- As técnicas de análise exploratória visam **resumir** e **apresentar** as informações de um conjunto de dados brutos.

Análise exploratória

- A análise exploratória de dados é uma área relativamente nova.
- Nasceu do clássico livro **Exploratory Data Analysis** de **John Tukey** em 1977.
- Algo curioso é que Tukey tinha uma relação próxima com a Ciência da Computação e definiu os termos **bit** e **software**.

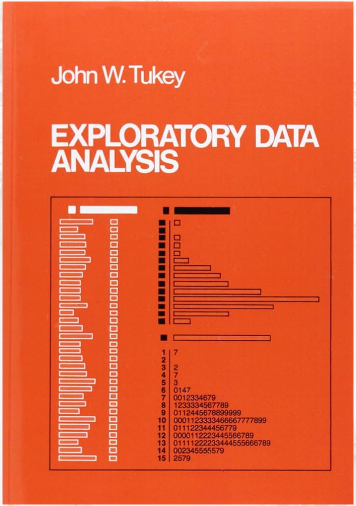


Figura 1. Capa do livro Exploratory Data Analysis de John Tukey.

Análise exploratória

- Como quase tudo em análise de dados, o **avanço computacional** permitiu com que a análise exploratória evoluísse substancialmente.
- Por exemplo: historicamente o processo de criação de um gráfico era reservado a pessoas qualificadas pois a produção de uma visualização era difícil.
- Hoje qualquer pessoa pode inserir dados em um aplicativo e gerar um gráfico.
- Este tipo de facilidade é importante para disseminação e democratização dos métodos, porém abre margem para certas práticas inadequadas.

Análise exploratória

- Tentar compreender um conjunto de dados sem algum método que permita resumir as informações é inviável.
- A análise exploratória é a primeira forma de tentarmos entender o que acontece nos nossos dados.
- Uma das tarefas é a etapa de consistência dos dados, isto é, verificar se os dados coletados são condizentes com a realidade.

Figura 2. Extraído de pixabay.com.

Análise exploratória

- O conjunto de técnicas aplicáveis está diretamente associado ao **tipo das variáveis de interesse** (quantitativas x qualitativas) e suas ramificações.
- Podemos conduzir análises focadas nas variáveis uma a uma (**análises univariadas**).
- Também podemos conduzir análises focadas em avaliar a relação entre as variáveis (**análises multivariadas**).

Figura 3. Extraído de pixabay.com.

Análise exploratória

Podemos fazer uso diversas técnicas, tais como

- Tabelas de frequência absolutas.
- Tabelas de frequência relativas.
- Tabelas de frequência acumuladas.
- Tabelas para múltiplas variáveis.
- Gráficos.

- Medidas de posição central.
- Medidas de posição relativa.
- Medidas de forma.
- Medidas de dispersão.
- Medidas de associação.

Resumos numéricos

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

- 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 idênticas, 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.

Medidas de dispersão

Considere os seguintes conjuntos de valores:

A	5	5	5	5	5	5	5	5	5	5
B	5	4	4	5	6	5	4	6	5	6
C	0	5	9	0	5	11	10	5	5	0

- Os conjuntos apresentam valores distintos, mas as medidas de posição central (média, moda e mediana), são idênticas.
- Precisamos de formas de mensurar o quanto os valores variam.

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.
- Estas medidas buscam mensurar o quanto os dados estão “compactados” ou “espalhados”.
- Uma medida de dispersão **não pode ser negativa**: ela será zero, indicando que todos os dados são iguais, ou ela é positiva, indicando algum grau de variabilidade nos dados.

Medidas de dispersão

- As medidas de dispersão mais usadas são baseadas nas diferenças entre cada observação e uma medida de posição central, esta diferença é chamada de **desvio**.
- Um jeito de medir a variabilidade como um todo é encontrar um **valor típico para os desvios**, como uma média.
- Fazer isso com os desvios simples não é muito inteligente. Desvios negativos se anulam com os positivos e a soma dos desvios com relação a média sempre será 0.
- Uma alternativa é calcular a média dos **desvios absolutos ou quadráticos** com relação a alguma medida de posição central.

Medidas de dispersão

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

Amplitude

- 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

Exemplo

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

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

$$Y: \text{Notas obtidas.}$$

$$Amp = 97 - 48 = 49$$

Desvio absoluto médio

- Tomamos todos os **desvios absolutos** com relação a alguma medida de posição central (média ou mediana).
- Calculamos a **média** destes desvios.
- Uma medida alternativa é o **desvio absoluto mediano** em que em vez de calcular a média dos desvios absolutos calculamos a mediana.

$$DAM_{MÉDIA} = \frac{1}{n}\sum_{i=1}^{n} |(y_i - \overline{y})| \qquad DAM_{MEDIANA} = \frac{1}{n}\sum_{i=1}^{n} |(y_i - md)|$$

Desvio absoluto médio

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 absoluto médio com relação à média e à mediana.

Desvio absoluto médio

Exemplo - desvio absoluto médio com relação à média

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

$$DAM = \frac{1}{10}\left(15{,}3 + 10{,}3 \ldots + 4{,}7 + 27{,}3\right) = 14{,}44$$

Desvio absoluto médio

Exemplo - desvio absoluto médio com relação à mediana

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

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

Variância

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

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

- A **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).
- A justificativa teórica para isso está relacionada com **estimadores não viciados** e com a **distribuição amostral da média**, tópicos que serão discutidos em inferência estatística.

Desvio padrão

- 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.

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

- A **variância** e o **desvio padrão** são **invariantes** com respeito a localização dos dados. Isso significa que, se somarmos ou subtrairmos uma constante em todos os valores, não alteramos a dispersão.

Lei de Chebyshev

- Independente da forma da distribuição dos dados e de sua variabilidade, conhecemos a **proporção mínima dos valores contidos em intervalos simétricos em relação à média**:
  - Pelo menos 3/4 (75%) dos valores estão no intervalo $(\bar{y} - 2s, \bar{y} + 2s)$.
  - Pelo menos 8/9 (89%) dos valores estão no intervalo $(\bar{y} - 3s, \bar{y} + 3s)$.
  - Pelos menos $(1 - 1/k^2)$ dos dados estará no intervalo $(\bar{y} - ks, \bar{y} + ks)$.

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:

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

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

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

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

$$\mathrm{CV} = 100 \cdot \frac{s}{\bar{y}}$$

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.

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

z-escore

- O z-escore pode ser visto como uma **medida de variabilidade individual** que nos diz quantos desvios padrões determinada observação está distante da média dos dados.
- O z-escore é dado por:

$$z = \frac{y_i - \bar{y}}{s}$$

z-escore

Exemplo

- No problema das notas de 10 alunos, em que as notas obtidas foram:

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

os z-escores para cada nota seriam:

$$-0{,}8794;\ -0{,}5920;\ 0{,}0977;\ 1{,}1323;\ -1{,}1093;\ 1{,}0749;\ 1{,}2473;\ 0{,}3276;\ 0{,}2702;\ -1{,}5692$$

Dispersão para variáveis qualitativas

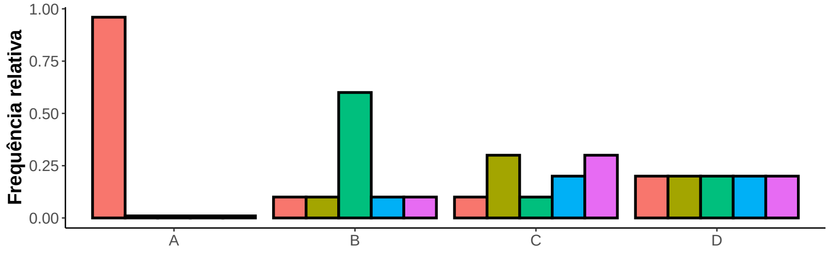
- Para variáveis qualitativas a **moda** é a única medida de posição que faz sentido.
- Como medida de dispersão, a ideia de **entropia** pode ser usada.
- Uma proposta, chamada de **índice de Shannon**, é dada por:

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

- Em que S representa o número de categorias da variável e f_r 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

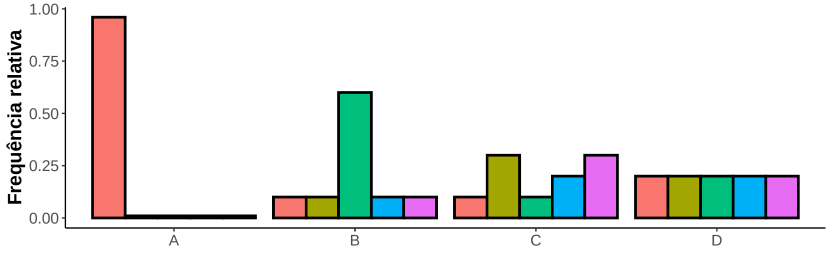
Qual é o mais homogêneo? Qual é o mais heterogêneo?



	f_{r1}	f_{r2}	f_{r3}	f_{r4}	f_{r5}
A	0.96	0.01	0.01	0.01	0.01
B	0.10	0.10	0.60	0.10	0.10
C	0.10	0.30	0.10	0.20	0.30
D	0.20	0.20	0.20	0.20	0.20

Dispersão para variáveis qualitativas

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



	f_{r1}	f_{r2}	f_{r3}	f_{r4}	f_{r5}	H
A	0.96	0.01	0.01	0.01	0.01	0.223
B	0.10	0.10	0.60	0.10	0.10	1.228
C	0.10	0.30	0.10	0.20	0.30	1.505
D	0.20	0.20	0.20	0.20	0.20	1.609

Desvio, variância, desvio padrão, coeficiente de variação, entropia

- Amplitude, desvio absoluto médio, variância e desvio padrão são **sensíveis a valores extremos**. Variância e desvio padrão ainda mais por serem baseados nos desvios quadráticos.
- **Variância** e **desvio padrão** tem **propriedades favoráveis**.
- O **desvio absoluto mediano da mediana** é uma medida que **não é influenciada**, assim como variâncias e desvios padrões aparados.
- Quando a distribuição dos dados é **simétrica** estas medidas tendem a convergir.
- O **coeficiente de variação** permite comparar a variabilidade de variáveis em **diferentes escalas**.
- O **z-escore** pode ser usado como uma medida de **variabilidade individual**.
- Para **variáveis qualitativas** existem medidas específicas, como o **índice de Shannon**.

O que foi visto:

- Medidas de dispersão.
  - Amplitude.
  - Desvio absoluto médio/mediano.
  - Variância.
  - Desvio padrão.
  - Coeficiente de variação.
  - z-escore.
  - Entropia.

Próximos assuntos:

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