Análise exploratória

Resumos numéricos - medidas de posição

Prof. Me. Lineu Alberto Cavazani de Freitas

Departamento de Estatística
Laboratório de Estatística e Geoinformação

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**.

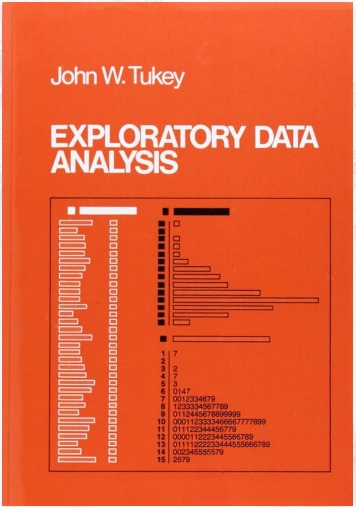

Figura 1. Capa do livro Exploratory Data Analysis de John Tukey.

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 dissemeninação 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 enteder 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 posição central

Medidas de posição central

- Um passo fundamental na exploração dos dados é definir um **valor típico** (uma estimativa onde a maior parte dos dados está localizada).
- Considerando um conjunto de valores qualquer, como definir um valor central? A resposta é: depende do critério.

- As medidas de posição central buscam expressar o **centro** de uma variável por meio de ideias como:
  - Centro de massa.
  - Valor que divide a amostra em partes iguais.
  - Valores de maior frequência ou densidade.

- Algumas possiblidades são
  - Média.
  - Mediana.
  - Moda.
  - Média geométrica.
  - Média harmônica.
  - Média aparada.

Média aritmética

- Soma de todos os valores dividida pela quantidade de elementos.
- Interpretação física de centro de gravidade.
- Medida influenciada por valores extremos.

Expressão

Sejam $y_1, y_2, \ldots, y_n$ os n valores de uma variável Y, a média é dada por:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

Média aritmética

Exemplo

- Considere que uma turma possui 10 alunos.
- Estes alunos realizaram uma avaliação.
- Considere que as notas obtidas foram:

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

- Qual foi a nota média da turma?

Y : Notas obtidas.

$$\overline{y} = \frac{60 + 65 + 77 + 95 + 56 + 94 + 97 + 81 + 80 + 48}{10} = \frac{753}{10} = 75{,}3$$

Média aritmética ponderada

- Indicada para **dados agrupados** em tabelas de frequência ou situações em que existe motivo para unidades receberem um **peso** maior.
- Obtêm-se os produtos entre frequências absolutas (ou pesos) e os valores que a variável assume.
- Somam-se os produtos e divide-se pela soma das frequências (quantidade de elementos).
- No caso de faixas de valores, usa-se o centro da faixa.

$$\overline{y} = \frac{\sum_{i=1}^{k} f_i \cdot y_i}{\sum_{i=1}^{k} f_i}.$$

- f_i representa a frequência da classe i.
- k representa o número de classes ($k \leq n$).

Média aritmética ponderada

Exemplo 1

- Considere que uma prova com 10 questões de múltipla escolha foi aplicada em uma turma com 100 alunos.
- Só temos acesso à uma tabela de frequências do número de questões corretas.
- Qual é o número médio de questões corretas?

Tabela 1. Tabela de frequências do número de questões acertadas.

Acertos	0	1	2	3	4	5	6	7	8	9	10
Frequência	1	0	0	5	2	30	21	29	8	3	1

Média aritmética ponderada

Exemplo 1

Y : Número de acertos.

$$\overline{y} = \frac{(0 \times 1) + (1 \times 0) + (2 \times 0) + (3 \times 5) + \ldots + (7 \times 29) + (8 \times 8) + (9 \times 3) + (10 \times 1)}{100}$$

$$\overline{y} = \frac{0 + 0 + 0 + 15 + 8 + 150 + 126 + 203 + 64 + 27 + 10}{100} = 6{,}03$$

Média aritmética ponderada

Exemplo 2

- Considere a seguinte tabela de frequências da idade dos funcionários de uma empresa.
- Qual é a idade média dos funcionários?

Tabela 2. Tabela de frequências das notas obtidas pelos alunos.

Faixas	[20,25]	(25,30]	(30,35]	(35,40]	(40,45]	(45,50]	(50,55]	(55,60]	(60,65]	(65,70]
Frequência	3	45	191	310	248	140	54	7	0	2

Média aritmética ponderada

Exemplo 2

Y : Idade do funcionário.

$$\overline{y} = \frac{(22{,}5 \times 3) + (27{,}5 \times 45) + (32{,}5 \times 191) \ldots + (57{,}5 \times 7) + (62{,}5 \times 0) + (67{,}5 \times 2)}{1000}$$

$$\overline{y} = \frac{67{,}5 + 1237{,}5 + 6207{,}5 + 11625 + \ldots + 2835 + 402{,}5 + 0 + 135}{1000} = 39{,}7$$

Outros tipos de média

- Média aritmética e ponderada são os tipos de média mais comuns.
- Contudo existem outras possibilidades como
  - Média geométrica.
  - Média harmônica.
  - Média aparada.

Mediana

- Valor que ocupa a **posição intermediária** dos valores ordenados.
- Divide o vetor de valores em 2 partes de mesmo tamanho.
- Metade dos valores é menor que a mediana e a outra metade maior que a mediana.
- Existem diferentes métodos para se obter a mediana, um deles é o chamado **método de Tukey**.
- No método de Tukey basta **ordenar o conjunto de valores** e verificar qual é o valor central.
- Se o número de observações for ímpar, a mediana é o valor central.
- Se o número de observações for par, a mediana é a média dos dois valores centrais.

Mediana (pelo método de Tukey)

- Passo 1: ordenar.

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n-1)} \leq y_{(n)}.$$

- Passo 2: obter a mediana de acordo com o número de elementos.

$$md = \begin{cases} y_{((n+1)/2)}, & \text{se } n \text{ for ímpar.} \\ (y_{(n/2)} + y_{(n/2+1)})/2, & \text{se } n \text{ for par.} \end{cases}$$

Mediana (pelo método de Tukey)

Exemplo

- Uma concessionária está fazendo o levantamento anual de vendas.
- Considere que as vendas por mês do ano anterior estão dadas na tabela.
- Qual é o número mediano de vendas?

Tabela 3. Tabela de frequências das vendas mensais.

Mês	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Vendas	93	113	112	104	84	104	107	105	96	92	93	97

Mediana (pelo método de Tukey)

Exemplo

- Passo 1: ordenar os valores.

Tabela 4. Vendas ordenadas.

(i)	1	2	3	4	5	6	7	8	9	10	11	12
Vendas	84	92	93	93	96	97	104	104	105	107	112	113

- Passo 2: obter a mediana de acordo com o número de elementos.
  - O número de elementos é par, portanto a mediana será a média dos dois valores centrais.
  - Mediana: $(97 + 104)/2 = 100{,}5$

Moda

- Valor ou classe que apresenta **maior frequência ou densidade**.
- Valor mais **típico**, aquele que mais se repete.
- Quando todos os valores são distintos, não existe moda.
- Quando a maior frequência está associada a mais de um valor, existe mais de uma moda.

Exemplo

- Considere que os valores a seguir dizem respeito ao número de filhos por pessoa em um grupo.

2; 3; 6; 1; 3;

4; 1; 2; 0; 1;

1; 0; 1; 4; 1

- Qual é a moda?
  - O valor mais frequente é 1, que aparece 6 vezes.

Média, mediana e moda

- Na prática, estas medidas possuem **vantagens** e **desvantagens**.
- Caso haja **valores discrepantes** a **média** é uma medida **altamente influenciada**, o que não acontece com a moda e a mediana.
- Já a **mediana** é difícil de ser obtida quando existem muitos dados, dado que o **processo de ordenação é custoso**.
- A dificuldade com a **moda** surge quando trabalha-se com **distribuições multimodais**, isto é diversos valores tem a mesma frequência de ocorrência.

Média, mediana e moda

- A **média** tende a ser uma boa alternativa quando a distribuição é **unimodal, simétrica e sem valores extremos**.
- A **mediana** tende a ser uma boa alternativa para **distribuições assimétricas** ou com presença de **valores extremos**.
- A **moda** tende a ser uma boa alternativa quando **valores se repetem**, estão **agrupados em classes** ou trata-se de uma **variável qualitativa**.
- Média, moda e mediana aproximam-se em distribuições **unimodais simétricas**.

Média, mediana, moda e assimetria

- Vimos anteriormente como avaliar assimetria por meio de recursos gráficos.
- Podemos utilizar as medidas de posição central
  - **Assimetria à direita**: $moda < mediana < média$.
  - **Assimetria à esquerda**: $média < mediana < moda$.
  - **Simetria**: $média = mediana = moda$.

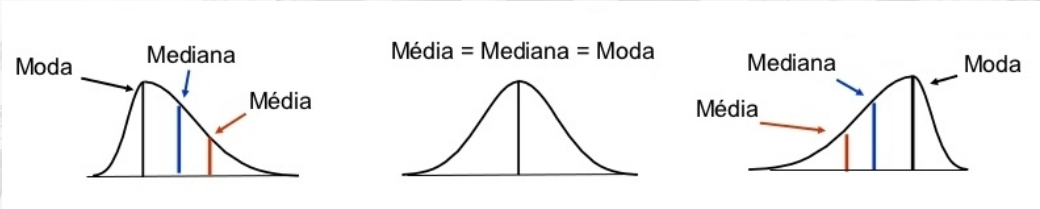

Figura 4. Relação medidas descritivas e assimetria

Medidas de posição relativa

Medidas de posição relativa

- As medidas de posição relativa ou separatrizes buscam representar **pontos do domínio** em que a variável apresenta porções com frequências conhecidas.
- Visam encontrar valores que representam alguma parcela dos dados.

- Algumas possiblidades são
  - Quartis.
  - Decis.
  - Percentis.
  - Máximo.
  - Mínimo.

Quartis

- Dividem a amostra em 4 **partes de mesmo tamanho**.
- A ideia para obtenção é similar à da **mediana**.
- Na verdade, a mediana é um dos quartis: o segundo.
- O primeiro e terceiro quartil são as **medianas** das duas partes divididas pela mediana (método de Tukey).

Quartis

- O **primeiro quartil** (Q_1) é o valor que marca $1/4$ das observações, isto é, 25%.
- O **segundo quartil** (Q_2) é o valor que marca $2/4 = 1/2$ das observações, isto é, 50% (a mediana).
- O **terceiro quartil** (Q_3) é o valor que marca $3/4$ das observações, isto é, 75%.
- A diferença entre primeiro e terceiro quartil é chamada de **amplitude interquartílica** ($AIQ = Q_3 - Q_1$).
- Estas quantidades são usadas para criação de um poderoso gráfico: o **box-plot**.

Quartis

Exemplo

- Considere os seguintes valores:

$$6; 12; 14; 7; 11; 7; 6; 12; 4; 11; 3; 4; 3; 4; 2$$

- Obtenha os quartis e a amplitude interquartílica.
- Passo 1: **ordenar**.

Tabela 5. Valores ordenados.

Posição	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Valor	2	3	3	4	4	4	6	6	7	7	11	11	12	12	14

Quartis

Exemplo

- Passo 2: **obter o segundo quartil (mediana)**.
  - Número de elementos: 15.
  - Posição do segundo quartil: 8.
  - Valor do segundo quartil: 6.
- Passo 3: **obter a mediana dos valores da primeira parcela.**
  - Número de elementos: 8 (da posição 1 até 8).
  - Posição da mediana da primeira parcela: 4,5.
  - Valor do segundo quartil: $(4 + 4)/2 = 4$.
- Passo 4: **obter a mediana dos valores da segunda parcela.**
  - Número de elementos: 8 (da posição 8 até 15).
  - Posição da mediana da segunda parcela: 4,5.
  - Valor do segundo quartil: $(11 + 11)/2 = 11$.
- $Q_1 = 4$, $Q_2 = 6$, $Q_3 = 11$.
- Amplitude interquartílica.

$$AIQ = Q_3 - Q_1 = 11 - 4 = 7$$

Quartis e o Box-plot

- O box-plot faz uso dos **quartis** para obtenção de um **gráfico**.
- Com ele é possível analisar a distribuição dos dados: **posição**, **variabilidade**, **assimetria**, **valores atípicos** (outliers).

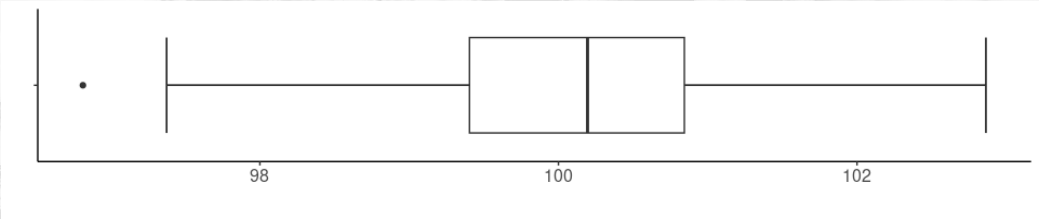

Figura 5. Ilustração box-plot completo.

Quartis e o Box-plot

- O Box-plot é construído a partir de **5 pontos** que resumem a distribuição dos dados observados: o **limite inferior**, o **1º quartil**, a **mediana**, o **3º quartil** e o **limite superior**.
- Os **limites inferior** e **superior** são utilizados para detectar observações que estão longe da massa central localizada entre o primeiro e o terceiro quartis.
- Entre o primeiro e terceiro quartil está a **mediana**. Não necessariamente a mediana estará no centro da caixa.

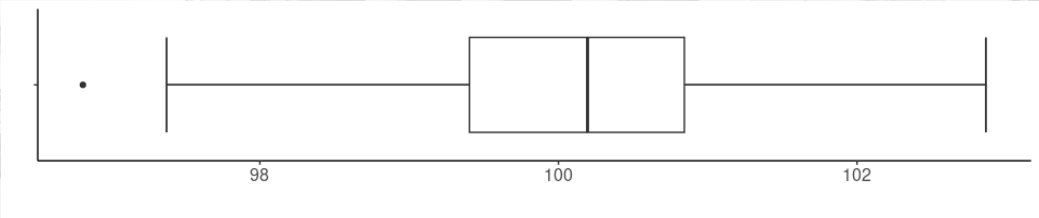

Figura 6. Ilustração box-plot completo.

Quartis e o Box-plot

- A construção de um box-plot inicia-se com um retângulo em que a aresta inferior coincide com o **primeiro quartil** e a superior com o **terceiro quartil**.

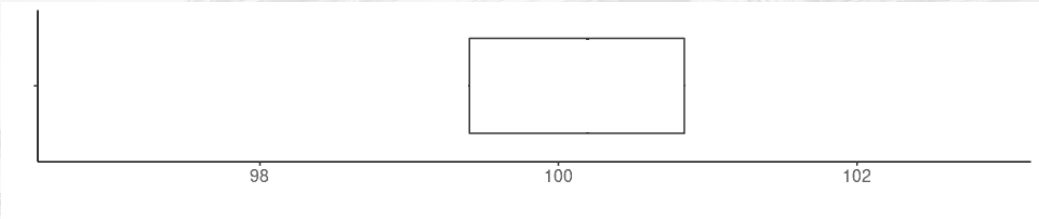

Figura 7. Arestas de um box-plot.

Quartis e o Box-plot

- A **mediana** é representada por um traço entre as duas arestas.
- De Q_1 até Q_3 estão 50% das observações centrais, o que dá uma ideia a respeito de quão dispersos são os valores.

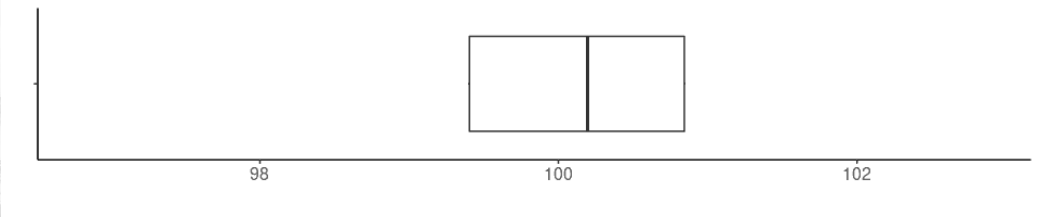

Figura 8. Arestas e mediana emum box-plot.

Quartis e o Box-plot

- Para obtenção da **amplitude do box-plot** além do retângulo faz-se $[Q1 - 1{,}5AIQ; Q3 + 1{,}5AIQ]$.
- Desenha-se então uma linha até estes valores.
  - Se estes valores excedem o mínimo e o máximo da variável, então a linha para no mínimo e no máximo da variável.

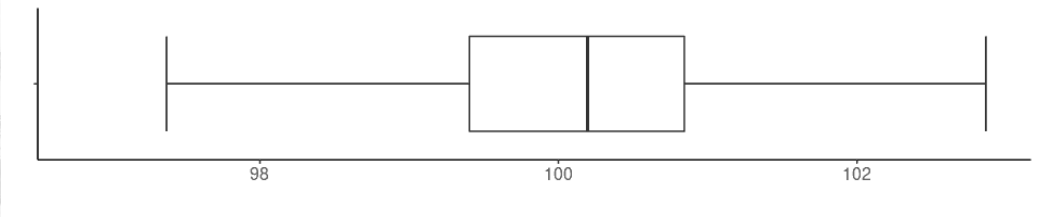

Figura 9. Inclusão dos limites de um box-plot.

Quartis e o Box-plot

- Valores além destes extremos são marcados como um ponto ou asterisco e são os candidatos a **valores atípicos**.

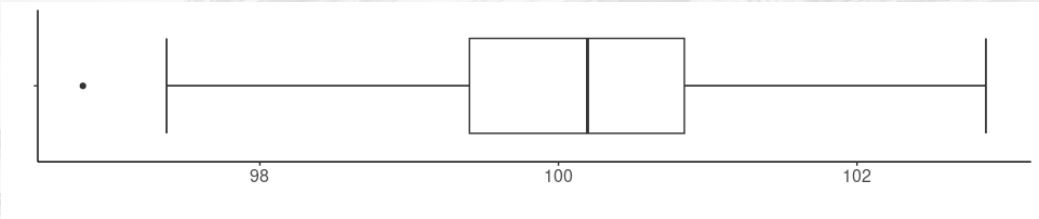

Figura 10. Box-plot completo.

Quartis e o Box-plot

- Os limitantes inferior e superior de um box-plot também são conhecidos como **valores adjacentes** ou também como **mínimo e máximo típicos**.
- Existem outras formas de obtenção de um box-plot, como por exemplo o box-plot em que não são calculados o mínimo e máximo típicos.
- Podem-se usar também outros quantis e outros pontos de corte, ou seja, existem outras formas para detectar pontos distantes da massa de dados.
- A interpretação do gráfico vai depender de como ele foi construído.
- Quanto mais observações, mais confiável será o box-plot.
- Contudo, quanto mais observações é natural que surjam mais pontos além dos limites do gráfico.

Quartis e o Box-plot

- Os pontos fora dos limites do box-plot costumam ser chamados de **valores atípicos ou outliers**.
- A definição exata de outlier é bastante **subjetiva** e vai além dos box-plots.
- Qualquer valor que seja muito distante dos outros valores em um conjunto de dados pode ser considerado outlier. Podemos usar o z-escore para verificar quais são os candidatos a outliers.
- Ser um outlier não torna um valor inválido ou errado, mas é um indicativo de um comportamento atípico (que pode ser causado por um erro de medida por exemplo).

Quartis para dados agrupados

Para calcular os quartis quando os dados estão agrupados, considere:

- n é o número total de observações;
- $Q_i (i = 1,2,3)$ é o quartil que desejamos obter;
- $(i \cdot n/4)$ é a posição na qual se encontra o quartil Q_i;
- l é o limite inferior da classe que contem Q_i;
- f é a frequência na classe que contem Q_i;
- h é a amplitude na classe que contem Q_i;
- F_{ant} é a frequência acumulada até a classe anterior à que contem Q_i.

O quartil Q_i é obtido aplicando-se a seguinte fórmula:

$$Q_i = l + \frac{(i \cdot n/4 - F_{ant})}{f} \cdot h$$

Outras medidas

- O **mínimo** e o **máximo** também são medidas de posição relativa e fornecem informação quanto ao domínio da variável.
- **Quartis** são a forma mais famosa de particionamento dos dados, porém qualquer outro percentual pode ser obtido.
- Se temos um conjunto de n valores, organizados de forma crescente, o P-ésimo percentil é um número tal que P% dos valores estejam à sua esquerda e $(100 - P)$% à sua direita.
- Por exemplo, se obtivermos os valores que separam a amostra em 10 partes com frequência 1/10, temos os decis.
- Estas **separatrizes** podem ser obtidas por meio do **gráfico de densidade empírica**.

O que foi visto:

- Resumos numéricos.
- Medidas de posição central.
- Medidas de posição relativa.

Próximos assuntos:

- Medidas de dispersão.