Fundamentos de Análise Exploratória de Dados

Conceitos e Aplicações

Encontro 2

Resumos numéricos e análises bivariadas.

Prof. Me. Lineu Alberto Cavazani de Freitas

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**.

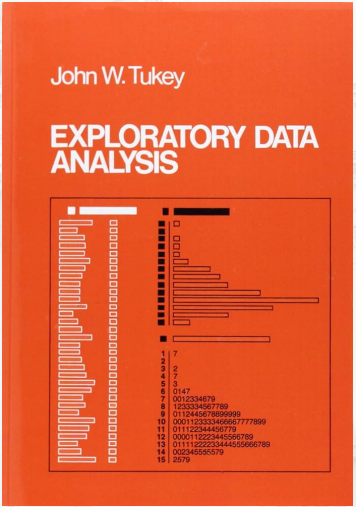


Figura 1. Capa do livro Exploratory Data Analysis de John Tukey.

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 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**: $média = mediana = moda$.

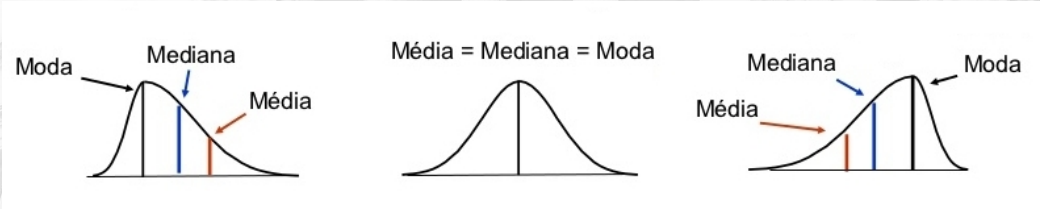


Figura 4. Relação medidas descritivas e assimetria

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).

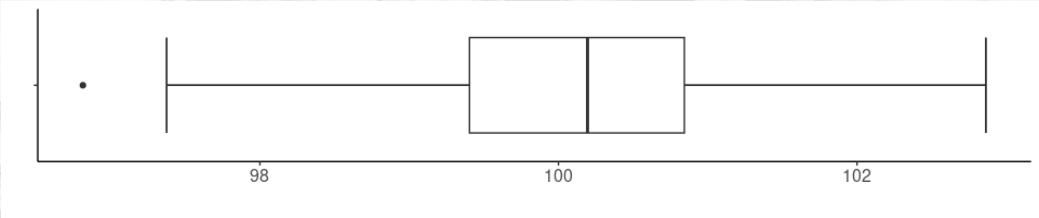


Figura 5. Ilustração box-plot completo.

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.

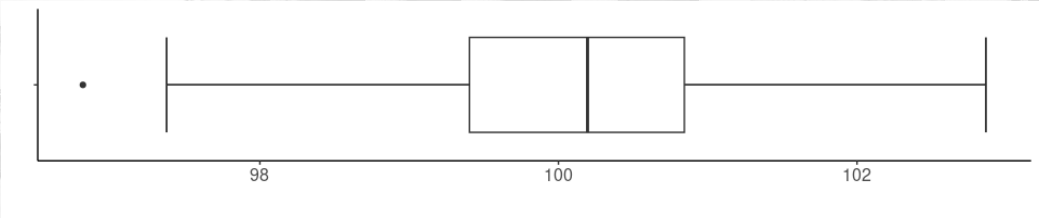


Figura 6. Ilustração box-plot completo.

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**.

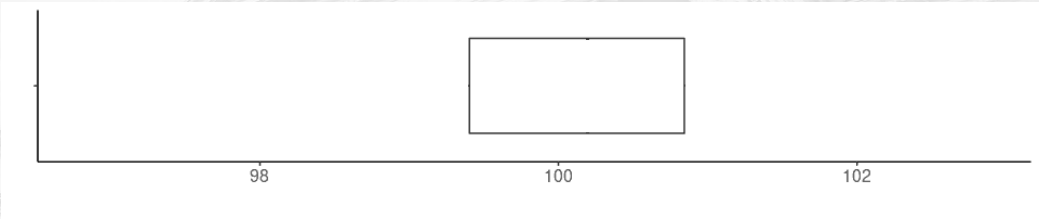


Figura 7. Arestas de um box-plot.

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.

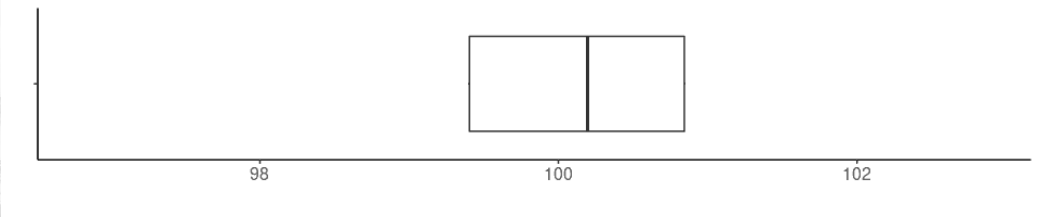


Figura 8. Arestas e mediana emum box-plot.

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.

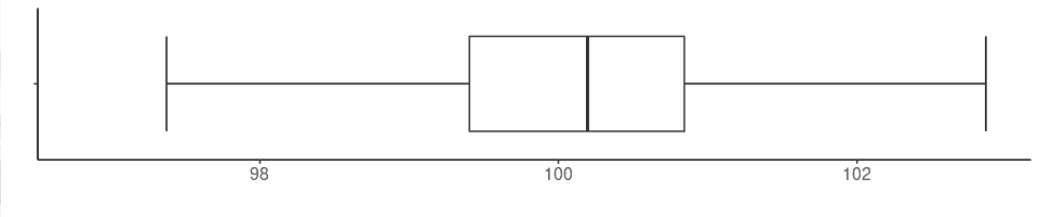


Figura 9. Inclusão dos limites de um box-plot.

Quartis e o Box-plot

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

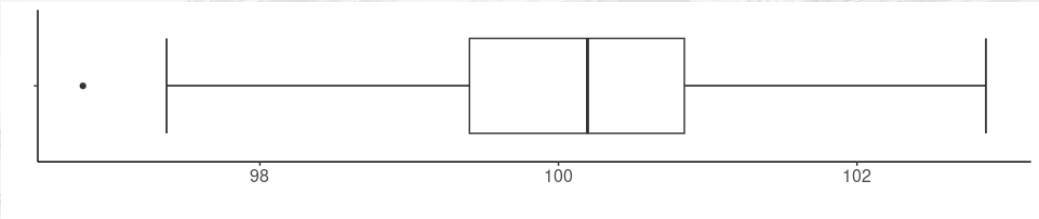


Figura 10. Box-plot completo.

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 frequências acumuladas**.

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 : 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.

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

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

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 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:

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

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

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

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.

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

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:

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

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

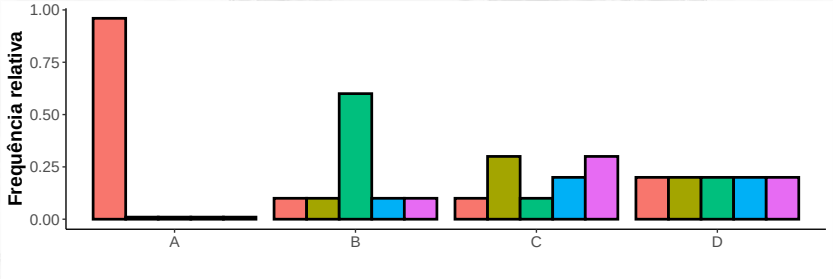
- Para variáveis qualitativas a **moda** é a única medida de posição que faz sentido.
- Como medida de dispersão, a ideia de **entropia** pode ser usada.
- Uma proposta, chamada de **índice de Shannon**, é dada por:

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

- 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

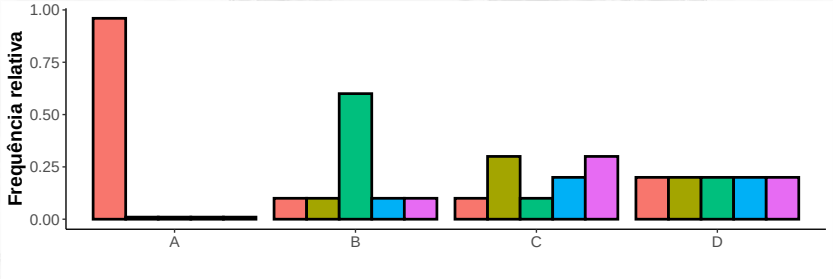
Qual é o mais homogêneo? Qual é o mais heterogêneo?



	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**.

Análise exploratória bivariada

Análise exploratória bivariada

- Em alguns casos podemos estar interessados na análise de **duas variáveis simultaneamente**.
- O objetivo é investigar a relação de **associação** entre as variáveis.
- **Tabelas**, **gráficos** e **coeficientes** específicos para relação entre variáveis podem ser usados.
- Tal como nas análises univariadas, as escolhas dependem dos tipos das variáveis.
- Considerando variáveis aos pares, as combinações podem ser:
  - Qualitativa x qualitativa.
  - Quantitativa x quantitativa.
  - Quantitativa x qualitativa.

Análise bivariada para variáveis qualitativas

Análise bivariada para variáveis qualitativas

- Neste tipo de situação avaliamos a **frequência** de observações para cada **combinação** de níveis das duas variáveis.
- Podem ser usadas **tabelas de frequências cruzadas**, também chamadas de **tabelas de dupla entrada**.
- Também é possível representar as frequências por meio de **recursos gráficos**.

Tabelas de frequências cruzadas

- As **linhas** dizem respeito aos **níveis** de uma variável.
- As **colunas** aos **níveis** da outra variável.
- As **células** mostram as **frequências** (absolutas ou relativas).
- As tabelas de dupla entrada também são chamadas de **distribuição conjunta**.
- As **margens** mostram as **frequências marginais** (de apenas uma das duas variáveis), também chamada de **distribuição marginal**.
- No caso de frequências relativas podem ser usados o **total geral** ou os totais **linha** e **coluna**.

Tabelas de frequências cruzadas

Tabela 11. Tabela de dupla entrada usando frequências absolutas.

	capital	interior	outro	Total
casado	7	8	5	20
solteiro	4	4	8	16
Total	11	12	13	36

Tabelas de frequências cruzadas

Tabela 12. Tabela de dupla entrada usando frequências relativas.

	capital	interior	outro	Total
casado	0.19	0.22	0.14	0.56
solteiro	0.11	0.11	0.22	0.44
Total	0.31	0.33	0.36	1.00

Tabelas de frequências cruzadas

Tabela 13. Tabela de dupla entrada usando frequências relativas aos totais linha.

	capital	interior	outro	Total
casado	0.35	0.40	0.25	1
solteiro	0.25	0.25	0.50	1
Total	0.31	0.33	0.36	1

Tabelas de frequências cruzadas

Tabela 14. Tabela de dupla entrada usando frequências relativas aos totais coluna.

	capital	interior	outro	Total
casado	0.64	0.67	0.38	0.56
solteiro	0.36	0.33	0.62	0.44
Total	1.00	1.00	1.00	1.00

Análise bivariada para variáveis qualitativas

- As **frequências cruzadas** podem ser representadas por meio de gráficos.
- Variações de **gráficos de barras** são as opções mais comuns.
- As possibilidades podem usar as **frequências absolutas**, **relativas** e permitem comparar a **composição** das variáveis.

- Gráficos para frequência para duas variáveis qualitativas:
  - Gráficos de barras lado a lado.
  - Gráfico de barras empilhadas.
  - Gráficos de barras empilhadas relativo.

Gráficos de barras lado a lado

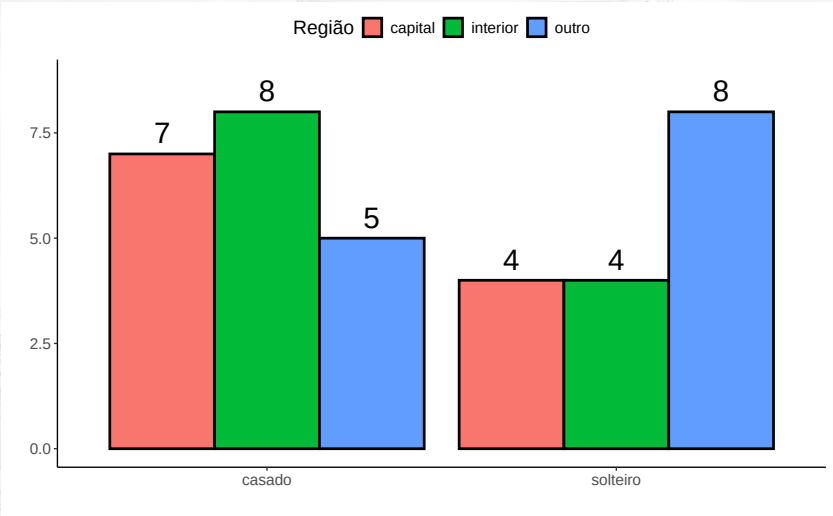


Figura 11. Gráfico de barras lado a lado.

Gráficos de barras lado a lado

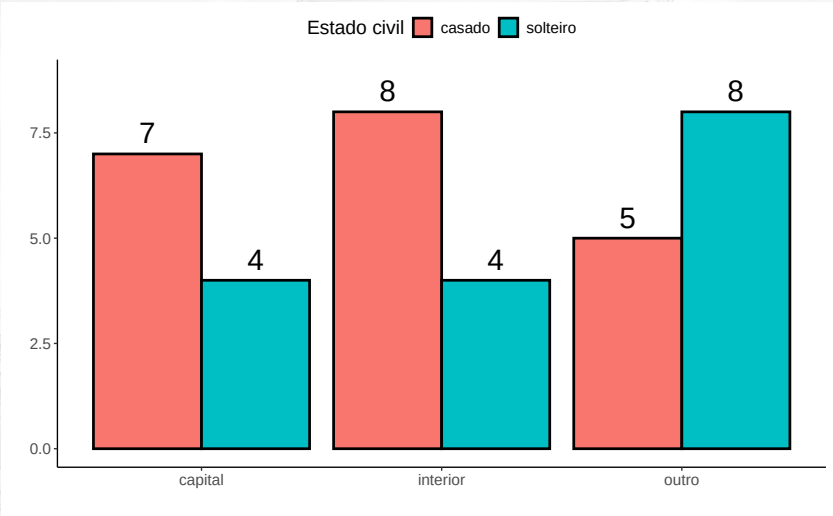


Figura 12. Gráfico de barras lado a lado.

Gráficos de barras empilhadas

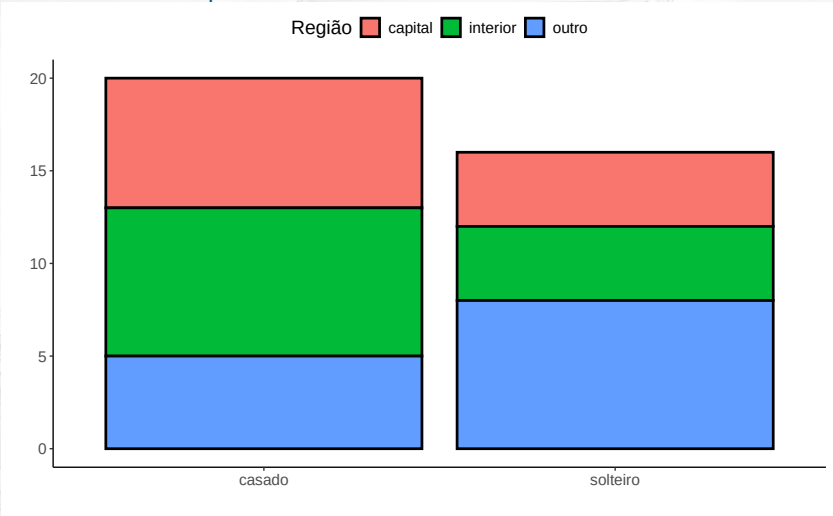


Figura 13. Gráfico de barras empilhadas.

Gráficos de barras empilhadas

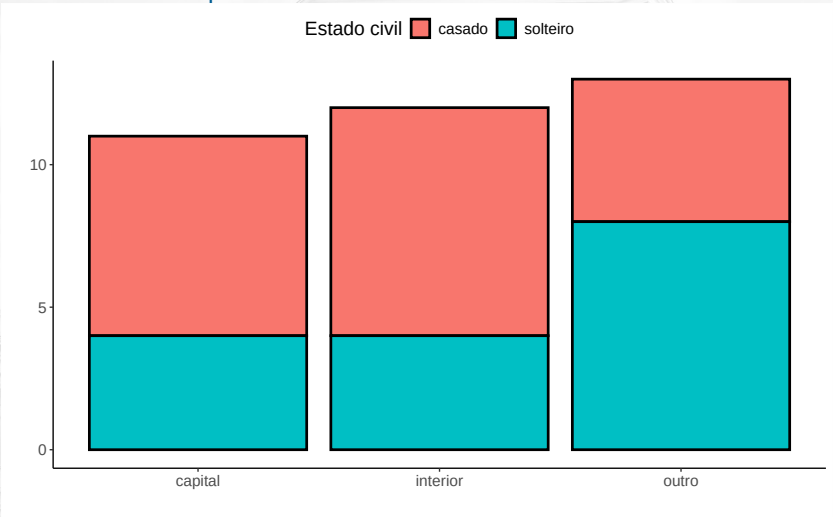


Figura 14. Gráfico de barras empilhadas.

Gráficos de barras empilhadas relativo

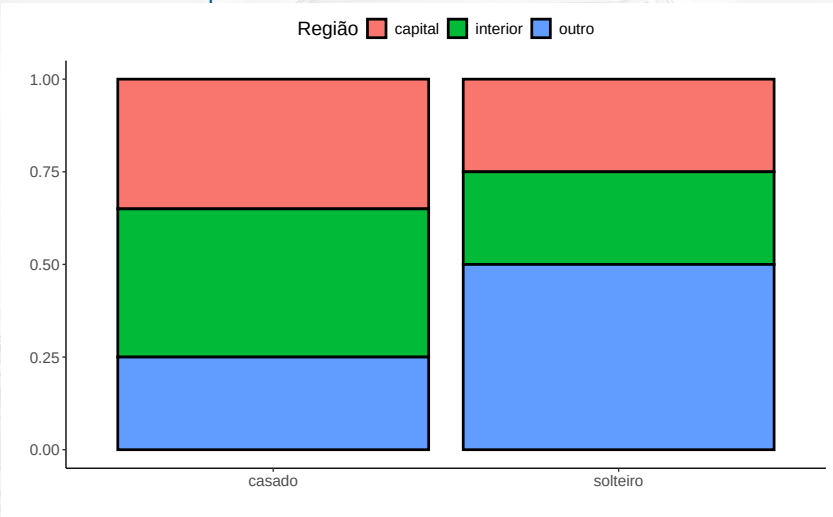


Figura 15. Gráfico de barras empilhadas relativo.

Gráficos de barras empilhadas relativo

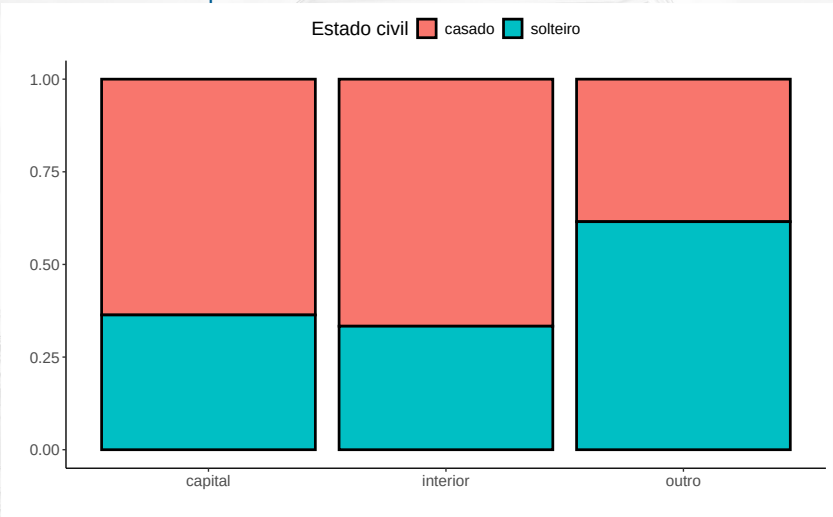


Figura 16. Gráfico de barras empilhadas relativo.

Medidas de associação para variáveis qualitativas

- Existem **medidas** que visam quantificar o **grau de associação** entre variáveis qualitativas.
- Uma dessas medidas é chamada de **Qui-quadrado**.
- Esta medida compara as **frequências observadas** em uma tabela de dupla entrada com as **frequências esperadas** caso não houvesse associação.
- Para obter a tabela de valores esperados basta, para cada casela, obter o produto entre o total da respectiva linha pelo total da respectiva coluna e dividir pelo total geral.

Medidas de associação para variáveis qualitativas

- O qui-quadrado é dado por:

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Quanto mais próximo de 0, menor a evidência de associação.
- Como o valor é irrestrito, existem variações desta quantidade que visam ter os limites definidos.

Medidas de associação para variáveis qualitativas

Tabela 15. Valores observados.

	capital	interior	outro	Total
casado	7	8	5	20
solteiro	4	4	8	16
Total	11	12	13	36

Tabela 16. Valores esperados.

	capital	interior	outro	Total
casado	6.11	6.67	7.22	20
solteiro	4.89	5.33	5.78	16
Total	11.00	12.00	13.00	36

Medidas de associação para variáveis qualitativas

Tabela 17. $\frac{(o-e)^2}{e}$.

	capital	interior	outro
casado	0.13	0.27	0.68
solteiro	0.16	0.33	0.85

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 2{,}42$$

Análise bivariada para variáveis quantitativas

Análise bivariada para variáveis quantitativas

- Buscamos identificar **padrões** e **tendências** na análise das duas variáveis.
  - A medida que os valores de uma variável aumentam, a outra reduz?
  - A medida que os valores de uma variável aumentam, a outra aumenta?
  - A medida que os valores de uma variável aumentam, a outra se mantém estável?

- As principais técnicas são o **coeficiente de correlação** e o **diagrama de dispersão**.
  - O coeficiente é uma métrica que avalia a associação linear entre um par de variáveis numéricas.
  - O diagrama é um gráfico de pares ordenados.

Coeficiente de correlação linear de Pearson

- Usado para determinar se existe **relação linear** entre variáveis quantitativas.
  - Assume valores entre -1 e 1.
  - Se o valor é maior 0, então existe uma associação linear **positiva**.
  - Se o valor é menor que 0, então existe uma associação linear **negativa**.
  - Se o valor é igual a 0, então **não existe** uma associação linear.

- **CORRELAÇÃO NÃO IMPLICA EM CAUSALIDADE**.
  - O fato de existir uma correlação linear, seja positiva ou negativa, não implica que uma variável possui real influência nos desfechos da outra.
  - Causalidade causa correlação, mas correlação não implica em causalidade.

Covariância e correlação

- A covariância entre duas variáveis Y_1 e Y_2 é dada por:

$$\text{Cov}(y_1, y_2) = \frac{1}{n-1} \sum_{i=1}^{n} (y_{1i} - \overline{y}_1) \cdot (y_{2i} - \overline{y}_2).$$

- A partir da covariância podemos obter a correlação, que padroniza a medida pelas variâncias, fazendo com que, independente das variáveis, sempre seja um valor entre -1 e 1.

$$r = \frac{\sum_{i=1}^{n} (y_{1i} - \overline{y}_1) \cdot (y_{2i} - \overline{y}_2)}{\sqrt{\sum_{i=1}^{n} (y_{1i} - \overline{y}_1)^2} \cdot \sqrt{\sum_{i=1}^{n} (y_{2i} - \overline{y}_2)^2}} = \frac{\text{Cov}(y_1, y_2)}{\sqrt{\text{V}(y_1) \cdot \text{V}(y_2)}}.$$

Outros tipos de correlação

- A correlação de Pearson não serve para descrever associações que não sejam lineares.
- Existem outros tipos de correlação que servem inclusive para variáveis de outros tipos.
- Alguns exemplos são:
  - Correlação de Spearman.
  - Correlação de Kendall.
  - Ponto-bisserial.

Diagrama de dispersão

- O **diagrama de dispersão** é a principal ferramenta para visualizar duas variáveis quantitativas.
- Em um eixo são representados os valores de uma variável.
- No outro eixo os valores de uma segunda variável.
- Os pares ordenados são representados por pontos.

Diagrama de dispersão

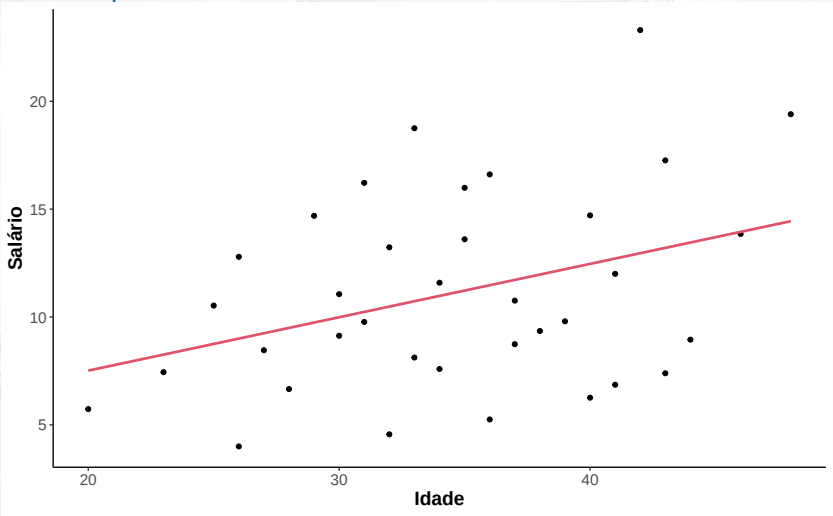


Figura 17. Diagrama de dispersão para o salário em função da idade.

Interpretação gráfica

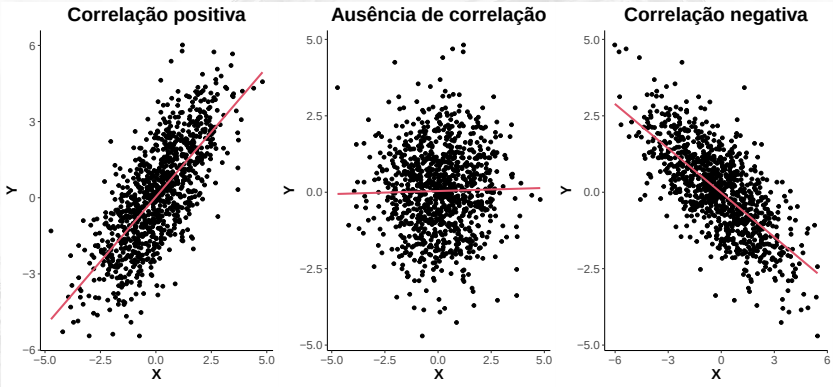


Figura 18. Avaliação de correlação usando diagramas de dispersão.

Covariância, correlação e diagrama de dispersão

Exemplo

- Considere as variáveis peso (Y_1) e altura (Y_2) de um conjunto de 10 indivíduos.

$$Y_1 : 60{,}09; 57{,}97; 54{,}12; 70{,}76; 59{,}74; 50{,}41; 58{,}19; 65{,}35; 71{,}18; 54{,}76$$

$$Y_2 : 1{,}54; 1{,}62; 1{,}52; 1{,}76; 1{,}63; 1{,}52; 1{,}65; 1{,}67; 1{,}66; 1{,}57$$

- $\overline{Y_1} = 60{,}26$; $\overline{Y_2} = 1{,}61$.
- $Var(Y_1) = 47{,}8$; $Var(Y_2) = 0{,}006$.
- Obtenha a covariância, coeficiente de correlação e o diagrama de dispersão.

Covariância, correlação e diagrama de dispersão

Exemplo

$$\text{Cov}(y_1, y_2) = \frac{1}{10-1}\{[(60{,}09 - 60{,}26) \cdot (1{,}54 - 1{,}61)] + \ldots + [(57{,}76 - 60{,}26) \cdot (1{,}57 - 1{,}61))]\}$$

$$\text{Cov}(y_1, y_2) = 0{,}44$$

$$r = \frac{0{,}44}{\sqrt{47{,}8 \cdot 0{,}006}} = 0{,}82$$

Covariância, correlação e diagrama de dispersão

Exemplo - digrama de dispersão

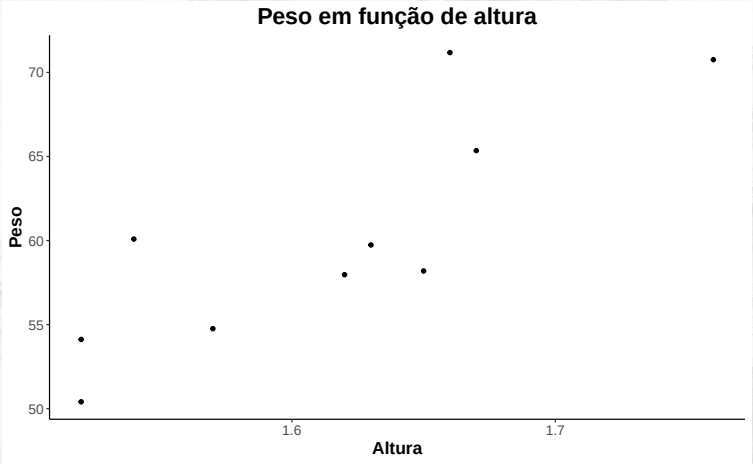


Figura 19. Diagrama de dispersão para peso e altura.

Covariância, correlação e diagrama de dispersão

Exemplo - digrama de dispersão

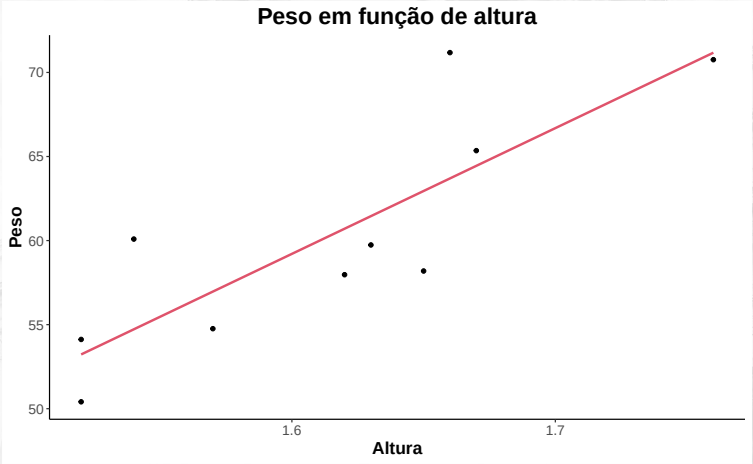


Figura 20. Diagrama de dispersão para peso e altura com linha de tendência linear.

Análise bivariada para uma variável qualitativa e uma quantitativa

Análise bivariada para uma variável qualitativa e uma quantitativa

- Neste caso estamos interessados em avaliar se os valores da variável numérica estão associados com os níveis da variável categórica.
- Podemos usar **medidas descritivas** para os valores dentro de cada um dos níveis da variável categórica.
- Para representar graficamente esta situação podemos criar um **box-plot** da variável numérica para cada nível do fator de interesse.

Tabela de medidas descritivas para níveis de um fator

Tabela 18. Medidas descritivas do salário em função da região.

Região	Média	Mediana	Desvio padrão
capital	11.46	9.77	5.48
interior	11.55	10.64	5.30
outro	10.45	9.80	3.15

Box-plot para níveis de um fator

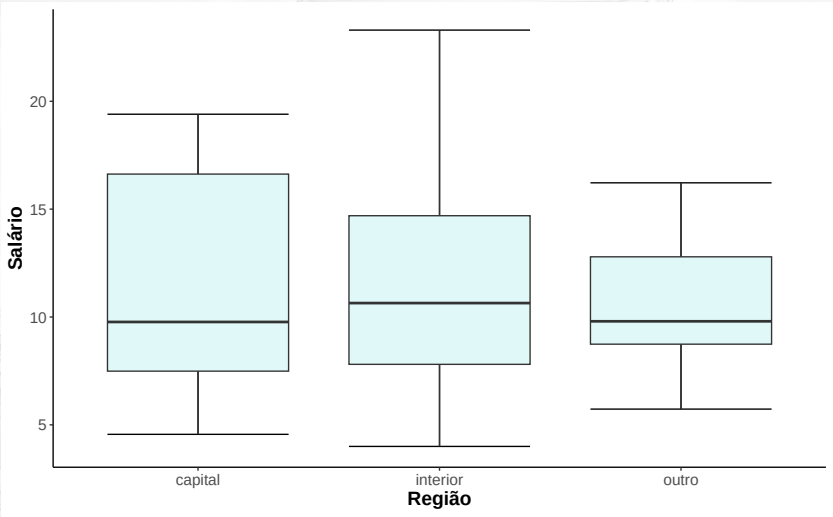


Figura 21. box-plots para o salário em função da região.

Outros tipos de gráficos e análises

Outros tipos de gráficos e análises

- Vimos as alternativas usuais para representação e análise de variáveis quantitativas e qualitativas.
- Contudo existem diversas situações particulares que exigem análises específicas.
- Algumas casos são: mapas, séries temporais, gráficos de perfil, nuvens de palavras.
- Também é possível trabalhar com gráficos que representam mais de duas variáveis ao mesmo tempo.
- Outra possibilidade é combinar gráficos.

O que foi visto:

- Resumos numéricos.
- Medidas de posição central.
- Medidas de posição relativa.
- Medidas de dispersão.
- Análises bivariadas.
  - Qualitativa x qualitativa.
  - Quantitativa x quantitativa.
  - Quantitativa x qualitativa.

Próximos assuntos:

- Revisando conceitos e extraindo informações de um conjunto de dados real com R.
- Considerações finais.