Introdução à Análise de Dados

Juliana Cobre e Katiane Silva Conceição

Departamento de Matemática Aplicada e Estatística - SME

Instituto de Ciências Matemáticas e de Computação - ICMC

Universidade de São Paulo - USP

Análise descritiva

O que é? Normalmente é a primeira análise feita depois de coletados os dados.

Análise descritiva

O que é? Normalmente é a primeira análise feita depois de coletados os dados.

Para que serve?

- conhecer as variáveis do conjunto de dados;
- conhecer o conjunto de dados como um todo;
- validar o conjunto de dados;
- verificar a existência de *outliers*;
- identificar possível associação entre variáveis;
- formular novas hipóteses.

Análise descritiva

O que é? Normalmente é a primeira análise feita depois de coletados os dados.

Para que serve?

- conhecer as variáveis do conjunto de dados;
- conhecer o conjunto de dados como um todo;
- validar o conjunto de dados;
- verificar a existência de *outliers*;
- identificar possível associação entre variáveis;
- formular novas hipóteses.

Como? Análise gráfica e medidas descritivas.

População

O que é? É uma coleção completa de todos os elementos a serem estudados e que possuem certa característica em comum.

Exemplos

- Brasileiros entre 16 e 25 anos;
- Peças produzidas em uma linha de produção de uma fábrica;
- Usuários do sistema XYZ.

Classificação da população

- **População finita:** indústrias situadas no Estado de São Paulo.
- **População infinita:** pressão atmosféricas ocorridas em diversos pontos do continente em determinado momento.

Qual a característica que mais interessa de uma população?

Definições

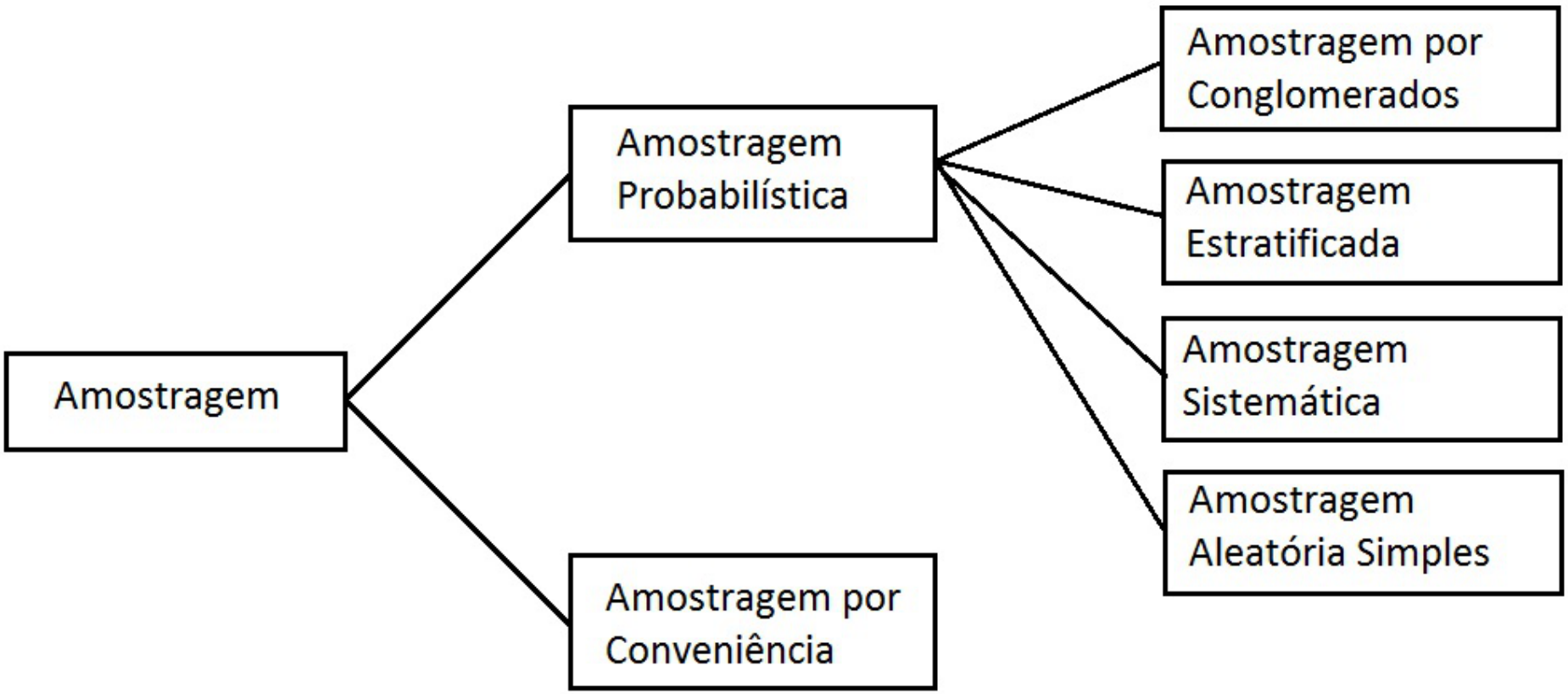
- **Variável:** Qualquer característica de interesse associada aos elementos de uma população.
- **Censo:** É um conjunto de dados obtidos de **todos** os membros da população.
- **Amostra:** É um subconjunto finito de membros selecionados de uma população.

Coleta de dados

- Os dados amostrais devem ser coletados de modo apropriado.
- Se os dados não são coletados de modo apropriado podem ser inúteis para análise estatística.
- A coleta apropriada dos dados está relacionada diretamente aos objetivos do experimento e as conclusões que se deseja chegar.
- Existem diferentes técnicas de amostragem para coleta de dados.

Nota: A ética é um ponto importante na elaboração da coleta de dados.

Técnicas de amostragem



Técnicas de amostragem

- **Amostragem probabilística:** Cada elemento da população tem uma chance conhecida de ser selecionado.
  - **Amostragem por conglomerado:** Divide-se a população em blocos (homogêneos) e amostram-se os blocos aleatoriamente e todos os indivíduos dentro do bloco são entrevistados.
  - **Amostragem estratificada:** Divide-se a população em grupo segundo algum estrato (pelo menos dois), amostram-se aleatoriamente indivíduos dentro de cada grupo.
  - **Amostragem sistemática:** Selecionar um elemento da população a cada k.
  - **Amostragem aleatória simples:** Cada indivíduo tem a mesma chance de ser amostrado.
- **Amostragem por conveniência:** Coletam-se elementos de fácil acesso ou de interesse para o estudo.

Classificação de variáveis

Classificação segundo à natureza

- **Qualitativa**
  - **Nominal**
  - **Ordinal**
- **Quantitativa**
  - **Discreta**
  - **Contínua**

Classificação de variáveis

- **Variáveis qualitativas:** Quando o resultado da observação é apresentado na forma de qualidade ou atributo.
  Exemplos: Sexo; religião; estado civil; setor de atividade econômica; porte de empresa; grau de escolaridade; etc.
  - **Variável qualitativa nominal:** Quando não existe qualquer ordenação para os resultados obtidos do processo de observação. Ex.: Sexo (feminino, masculino); setor de atividade econômica (industrial, comercial, serviços, etc).
  - **Variável qualitativa ordinal:** Quando existe uma certa ordenação (hierarquia) nos possíveis resultados das observações efetuadas. Ex.: porte de empresa (micro, pequena, média e grande); classe social (alta, média e baixa); grau de escolaridade (EF, EM, ES, PG).

Classificação de variáveis

- **Variáveis quantitativas:** Quando o resultado da observação é um número, decorrente de um processo de mensuração ou contagem. Exemplos: Número de empregados; salário mensal; faturamento anual; idade; peso; tamanho da família; etc.
  - **Variável quantitativa discreta:** Quando os resultados possíveis da observação formam um conjunto finito ou enumerável de números e que resultam, frequentemente, de uma contagem. Ex.: número de empregados $(1, 2, 3, \ldots)$; tamanho da família $(1, 2, 3, \ldots)$.
  - **Variável quantitativa contínua:** Quando os possíveis valores formam um intervalo ou uma união de intervalos de números reais e que resultam, normalmente, de uma mensuração. Ex.: Salário mensal; faturamento anual; altura; peso.

Classificação de variáveis

Observações

- Variável CEP?
- Variável idade?
- Variáveis codificadas (rotuladas)?

Classificação de variáveis

Classificação segundo a forma de mensuração

- Escala nominal ou classificadora;
- Escala ordinal ou por posto;
- Escala intervalar;
- Escala de razão.

Classificação de variáveis

- **Escala nominal ou classificadora:** Quando os números ou outros símbolos são usados para identificar os grupos a que vários objetos pertencem, esses números ou símbolos constituem uma escala nominal ou classificadora. Ex.: Sexo (0- Masculino; 1- Feminino).
- **Escala ordinal ou por posto:** Como na escala nominal, a escala ordinal permite verificar semelhanças e diferenças entre grupos. Porém, pode ocorrer que os grupos de classificação não sejam apenas diferentes, mas também apresentem uma certa relação entre eles do tipo: Mais alto do que; mais preferível a; mais difícil do que; etc. Ex.: Classe social (alta, média, baixa).

Classificação de variáveis

- **Escala intervalar:** Quando a escala tem todas as características de uma escala ordinal, e quando, além disso, se conhecem as distâncias entre dois números quaisquer da escala, então consegue-se uma mensuração consideravelmente mais forte que a ordinal. Em uma escala intervalar, o ponto zero e a unidade de medida são arbitrários. O ponto zero arbitrário significa que não existe o zero absoluto. Ex.: Temperatura (0°C).

  Observação: 0°C não significa ausência de calor; 40°C não é duas vezes mais quente que 20°C.

Classificação de variáveis

- **Escala de razão:** Quando uma escala tem todas as características de uma escala de intervalos e, além disso, tem um verdadeiro ponto zero como origem, é chamada de escalas de razão. Em uma escala de razões, a razão de dois pontos quaisquer da escala é independente da unidade de mensuração. Ex.: Distância de translado; Medição da estatura de um indivíduo.

  Observação: 6 km é o dobro de 3 km.

Medidas de posição

- Média aritmética.
- Média ponderada.
- Média geométrica.
- Média harmônica.
- Mediana.
- Moda.
- Ponto médio.
- Separatrizes (ou quantis: quartis; decis; centis ou percentis).

Medidas de posição: Medidas de tendência central

Medidas de tendência central ou de centro: indicam de alguma forma o centro ou o valor típico do conjunto de dados.

- Média
- Mediana.
- Moda.
- Ponto médio.

Medidas de posição: Média

Média aritmética (ou média): A **média aritmética** ou simplesmente **média** ($\bar{x}$) é a medida de centro encontrada pela soma de todos os valores do conjunto de dados dividido pelo número que representa a quantidade total de valores (ou tamanho do conjunto de dados).

Assim, se temos n observações da variável X, representados por $x_1, x_2, \ldots, x_n$, então a **média** é calculada por

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Nota: Tem a desvantagem de sofrer a influência de valores extremos.

No R: mean()

Medidas de posição: Média geométrica

Média geométrica: A **média geométrica** ($\bar{x}_g$) de um conjunto de números positivos é a medida de centro que é obtida com o produto de todos os membros do conjunto elevado ao inverso do número de valores. Assim

$$\bar{x}_g = \left\{ \prod_{i=1}^{n} x_i \right\}^{1/n} = \sqrt[n]{\prod_{i=1}^{n} x_i},$$

Nota: Adequada quando trabalhamos com taxas.

No R: geometric.mean(), pacote: `psych`

Medidas de posição: Média geométrica

Exemplo (Crescimento proporcional) Suponha que um investimento de US$ 100 acarreta em um lucro de 180, 210 e 300 seguindo os anos, então o crescimento é 80%, 16,67% e 42,86% para cada ano respectivamente.

1. Usando a **média aritmética**, calculamos a média do crescimento:

$$\frac{1,80 + 1,1667 + 1,4286}{3} = 1,4651.$$

   Se começarmos o investimento com US$ 100 e tivermos um crescimento de 46,51% cada ano, o resultado é US$ 314,48 e não 300. Ou seja, a **média aritmética** não corresponde aos valores de crescimento em cada ano.

Medidas de posição: Média geométrica

1. Em vez disso, podemos usar a **média geométrica**. Uma taxa de crescimento de 80% corresponde a multiplicar por 1,80 e assim por diante. Então, pegamos a média geométrica de 1,80, 1,1667 e 1,4286. Dessa forma, usando a **média geométrica** calculamos a média do crescimento:

$$\sqrt[3]{1,80 \times 1,1667 \times 1,4286} = 1,4423.$$

A "média" do crescimento por ano é 44,23%. Se nós começarmos com US$ 100 e pegarmos o número acrescido de 44,23% cada ano, o resultado é US$ 300.

Medidas de posição: Mediana

Mediana: A **mediana** (Md) é a medida de centro que divide o conjunto de dados ordenados (o rol, crescente ou decrescente) em duas partes iguais, ou seja, 50% dos dados antecedem o seu valor e 50% dos dados sucedem o seu valor.

Nota: É uma medida de tendência central resistente, porque não se altera muito devido à presença de valores extremos.

No R: median()

Medidas de posição: Mediana

Mediana: A **mediana** (Md) é a medida de centro que divide o conjunto de dados ordenados (o rol, crescente ou decrescente) em duas partes iguais, ou seja, 50% dos dados antecedem o seu valor e 50% dos dados sucedem o seu valor. Assim

$$Md = \begin{cases} \frac{x_{(n/2)} + x_{(n/2+1)}}{2}, & \text{se } n \text{ é par} \\ x_{([n/2]+1)}, & \text{se } n \text{ é ímpar} \end{cases}.$$

Nota: É uma medida de tendência central resistente, porque não se altera muito devido à presença de valores extremos.

No R: median()

Medidas de posição: Moda

Moda: A **moda** (*Mo*) é o valor mais frequente no conjunto de dados.

Nota: Um conjunto de dados pode ter uma moda (**unimodal**), duas modas (**bimodal**), mais de duas modas (**mutimodal**) ou nenhuma moda (**amodal**).

Medidas de posição: Quantil

Quantil: O **quantil de ordem** p ou o ***p*-quantil** ($q(p)$), $0 < p < 1$, é uma medida tal que $100p\%$ das observações são menores do que $q(p)$. Assim

$$q(p) = \begin{cases} \frac{x_{(np)} + x_{(np+1)}}{2}, & \text{se } np \text{ é inteiro} \\ x_{([np]+1)}, & \text{se } np \text{ não é inteiro} \end{cases}.$$

Nota: Alguns quantis têm nomes especiais: quartis ($p = 25\%, 50\%, 75\%$); decis ($p = 0,1; 0,2; \ldots, 0,9$); percentis ($p = 0,01; 0,02; \ldots, 0,99$).

No R: quantile()

Medidas de dispersão: Amplitude

Amplitude total ou amplitude: A amplitude total (A) é a diferença entre o maior e o menor valor presentes em um conjunto de dados, ou seja,

$$\begin{aligned} A &= \text{Valor máximo} - \text{Valor mínimo} \\ &= x_{(n)} - x_{(1)}. \end{aligned}$$

Nota: É muito sensível a valores extremos, pois se baseia apenas no **intervalo de variação**, $(x_{\text{min}}, x_{\text{max}})$, do conjunto de dados.

No R: range() e diff() ou max() e min()

Medidas de dispersão: Variância

Variância (amostral): A variância amostral (s^2) é a soma dos quadrados dos desvios em relação à média amostral ($Di = X_i - \hat{x}$) dividido por uma unidade a menos do número total de elementos da amostra (estimador não viciado). Assim

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

Nota1: Pode também ser dividido pelo número total de elementos (estimador viciado, superestimado ou subestimado).

Nota2: Difícil interpretação, pois sua unidade é a unidade dos dados ao quadrado. Ex.: cm^2, h^2, $R\$^2$.

No R: var()

Medidas de dispersão: Desvio padrão

Desvio padrão (amostral): É a raiz quadrada da variância (s). Assim

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}},$$

Nota1: É preferido à variância pois tem a mesma unidade dos dados.

Nota2: Também pode ser viciado ou não viciado dependendo do denominador.

No R: sd()

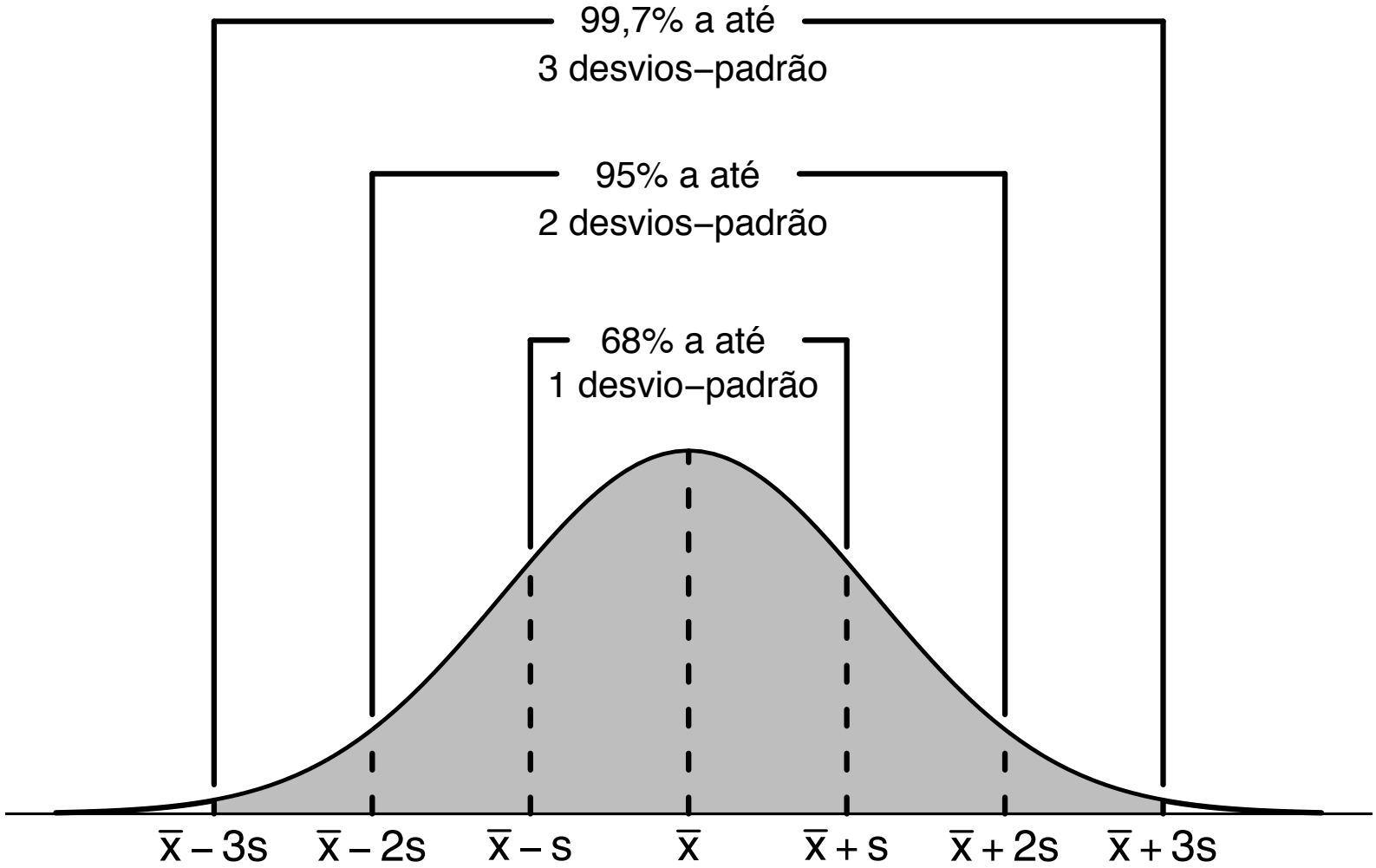
Medidas de dispersão: Observações

Variância e desvio padrão

- Só são iguais a zero se todos os valores observados são iguais.
- Maiores valores indicam maior variabilidade, dizemos que mais heterogêneo é o conjunto de dados.
- Sofrem influência de valores extremos.

Medidas de dispersão: Observações

Distribuições em forma de sino: Regra empírica



Medidas de dispersão: Coeficiente de variação

Coeficiente de variação de Pearson (ou dispersão relativa): o coeficiente de variação (CV) descreve o desvio padrão relativo à média (se esta for diferente de zero). Assim

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

Interpretação: Quanto maior o coeficiente de variação, maior é a variação (ou variabilidade) dos dados.

Nota: O CV é uma medida adimensional (sem unidade de medida).

No R: sd()/mean()

Representação gráfica

- **Variáveis qualitativas**
  - Gráfico de barras.
  - Gráfico de setores.

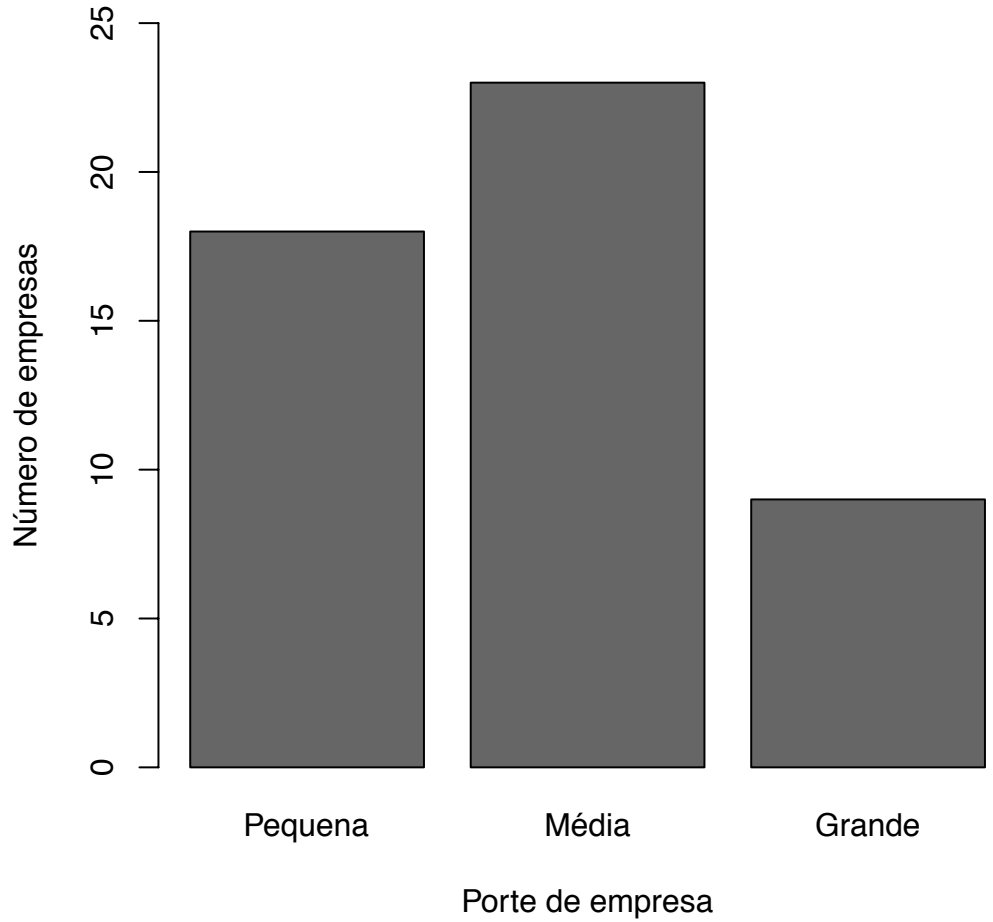
- **Variáveis quantitativas**
  - Gráfico de barras.
  - Gráfico de pontos.
  - Gráfico de setores.
  - Histograma.
  - Gráfico de caixa (box plot).
  - Gráfico de séries temporais (ou gráfico de linhas).
  - Diagrama de dispersão.

Representação gráfica: Gráfico de barras

- A escala vertical representa as frequências absolutas ou relativas.
- A escala horizontal identifica as diferentes categorias, classes ou observações dos dados.

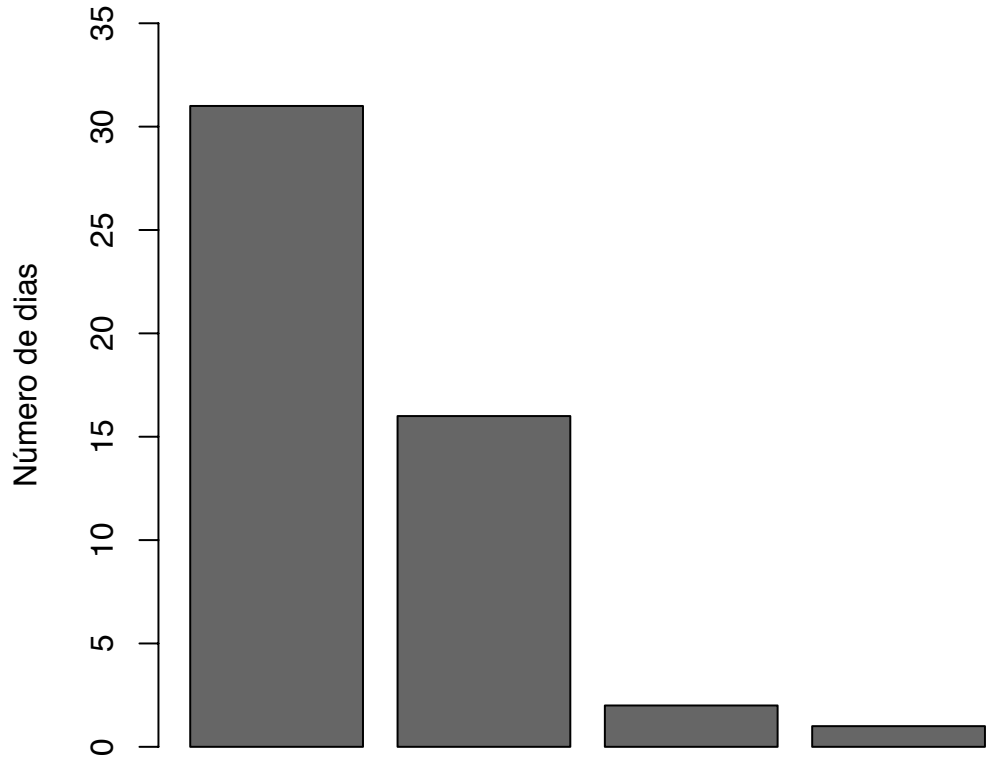
Nota: Podemos inverter os eixos (frequências na horizontal e variável na vertical).

Representação gráfica: Gráfico de barras



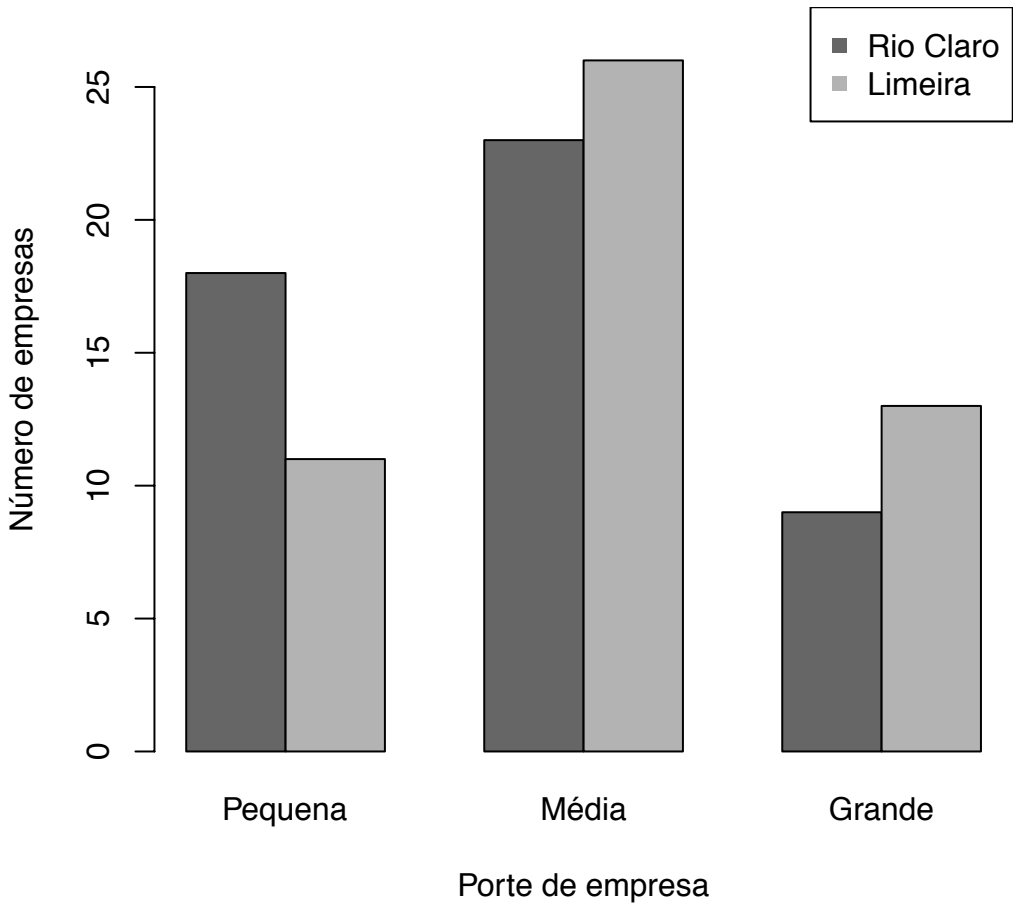
Representação gráfica: Gráfico de barras

Exemplo: Número de acidentes diários em um determinado cruzamento na cidade de São Carlos em 2013.



Representação gráfica: Gráfico de barras

Exemplo: Número de empresas segundo o seu porte nas cidades de Rio Claro e Limeira em 2015.

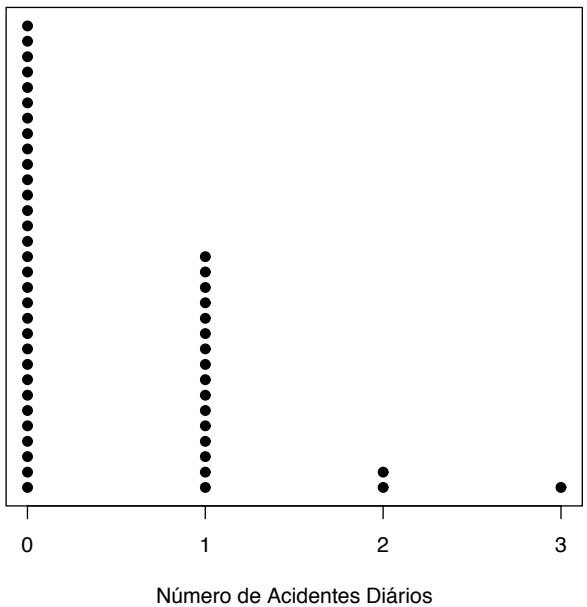


Representação gráfica: Gráfico de pontos

Gráfico de pontos

- É um gráfico no qual cada valor é plotado como um ponto ao longo de uma escala de valores.
- Os pontos que representam valores iguais são empilhados.

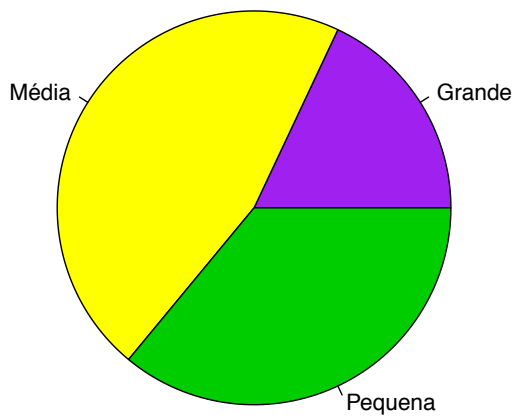
Exemplo: Número de acidentes diários em um determinado cruzamento na cidade de São Carlos em 2013.

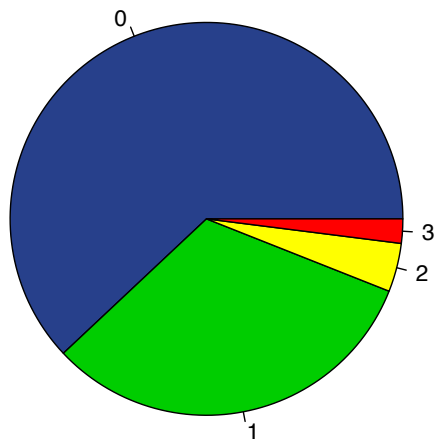


Representação gráfica: Gráficos de setores

Gráficos de setores (ou pizza)

- É um gráfico que retrata dados qualitativos ou quantitativos discretos como setores de um círculo.
- Cada setor é proporcional à contagem de frequência para a categoria ou observação.





Representação gráfica: Histograma

Histograma

- É uma versão gráfica da distribuição de frequência por classes.
- É um gráfico de barras adjacentes com bases iguais às amplitudes das classes e alturas iguais às densidades, sendo a densidade de cada classe dada por

$$f_{d_i} = \frac{f_i^*}{A_{c_i}}, i = 1, \ldots, k$$

  em que f_i^* é a frequência absoluta ou relativa da classe i e A_{c_i} é a amplitude da classe i.=
- Se as classes tiverem amplitude constante, as alturas das barras correspondem aos valores das frequências absolutas ou relativa.

No R: hist()

Representação gráfica: Histograma

Ilustração

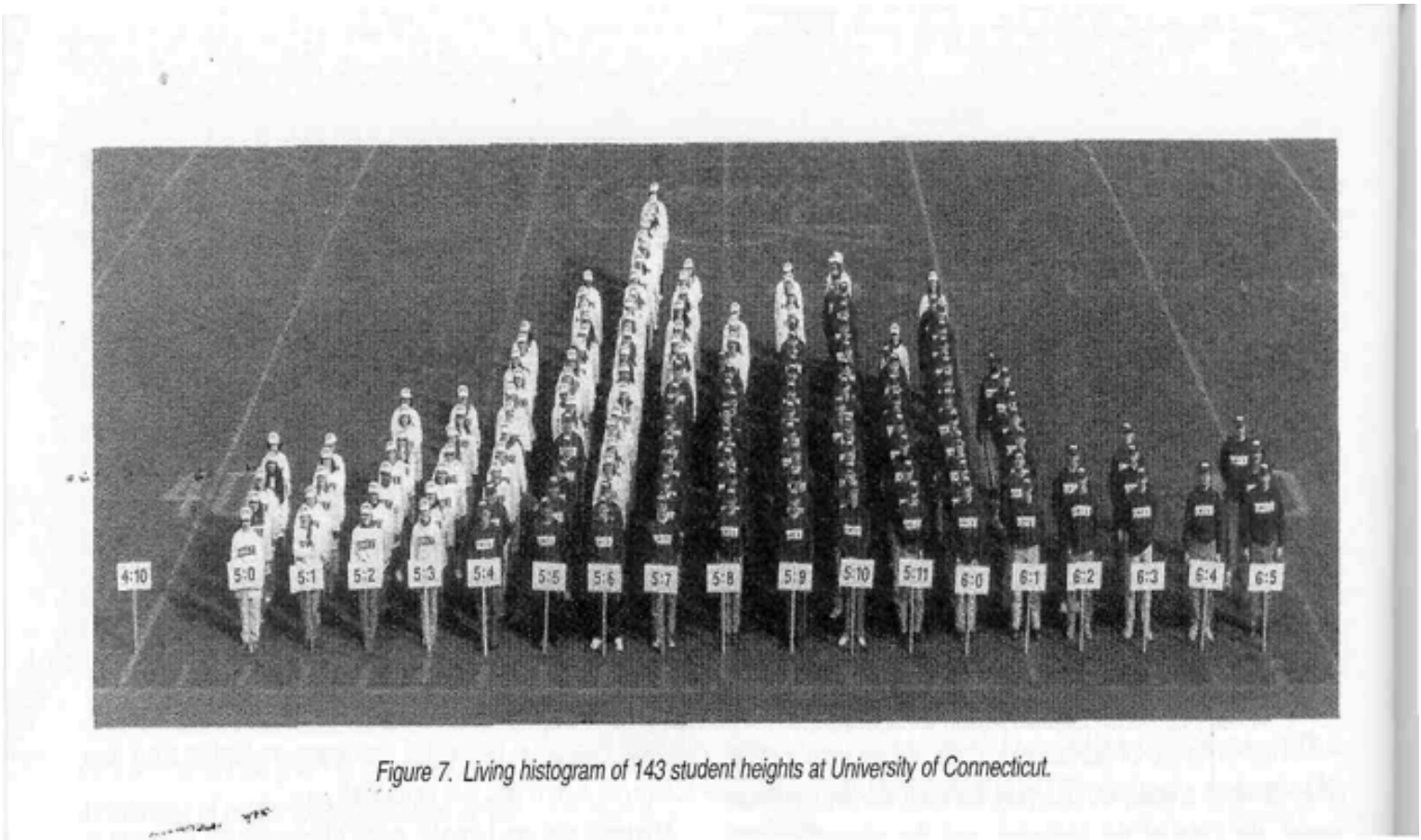
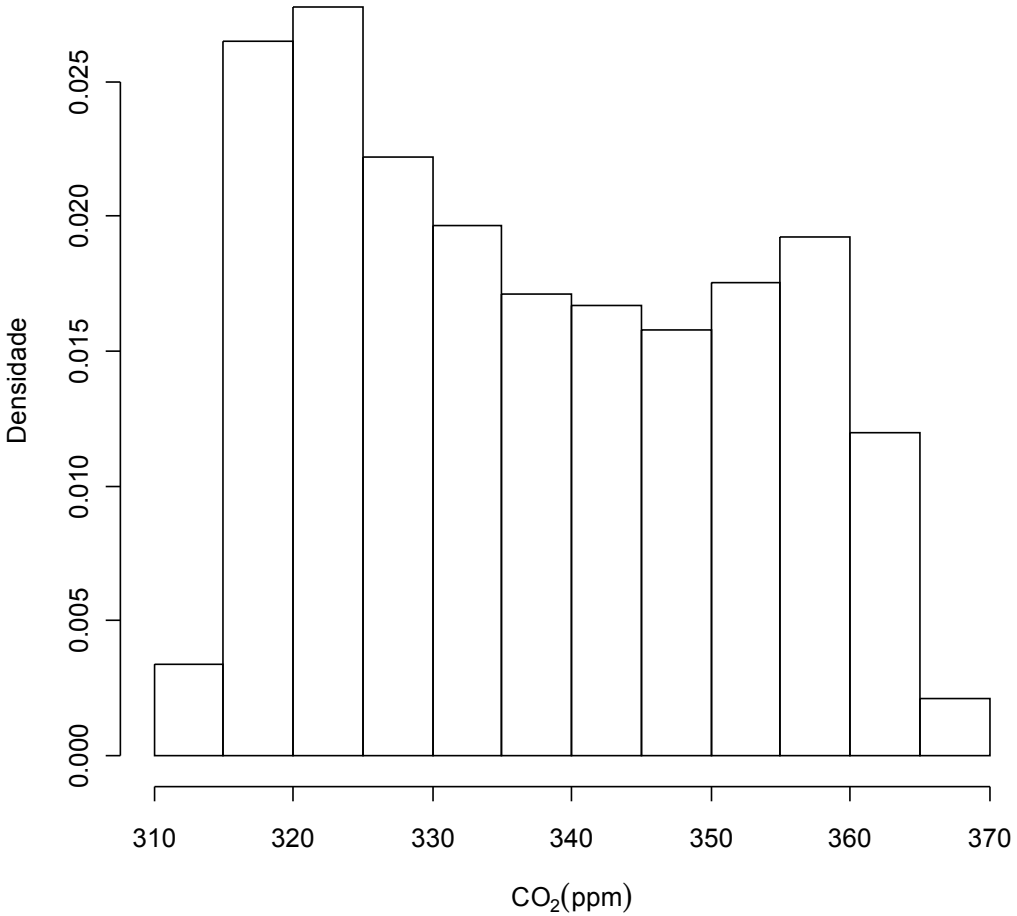


Figure 7. Living histogram of 143 student heights at University of Connecticut.

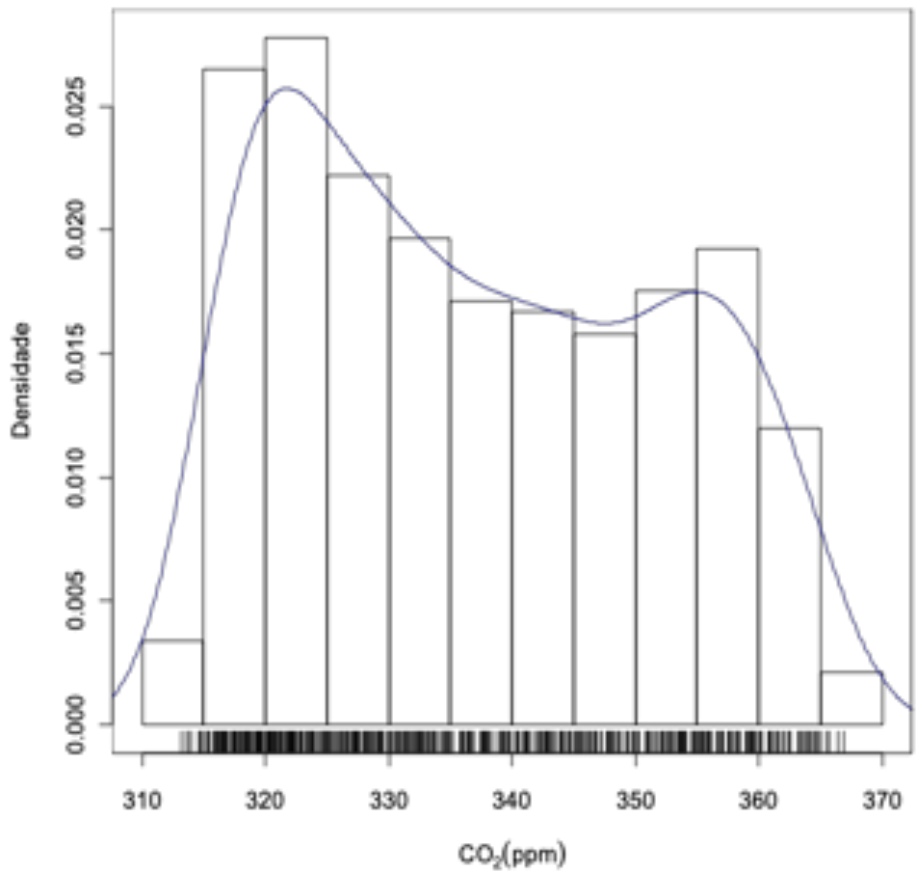
Representação gráfica: Histograma

Exemplo: Histograma da quantidade de CO_2.



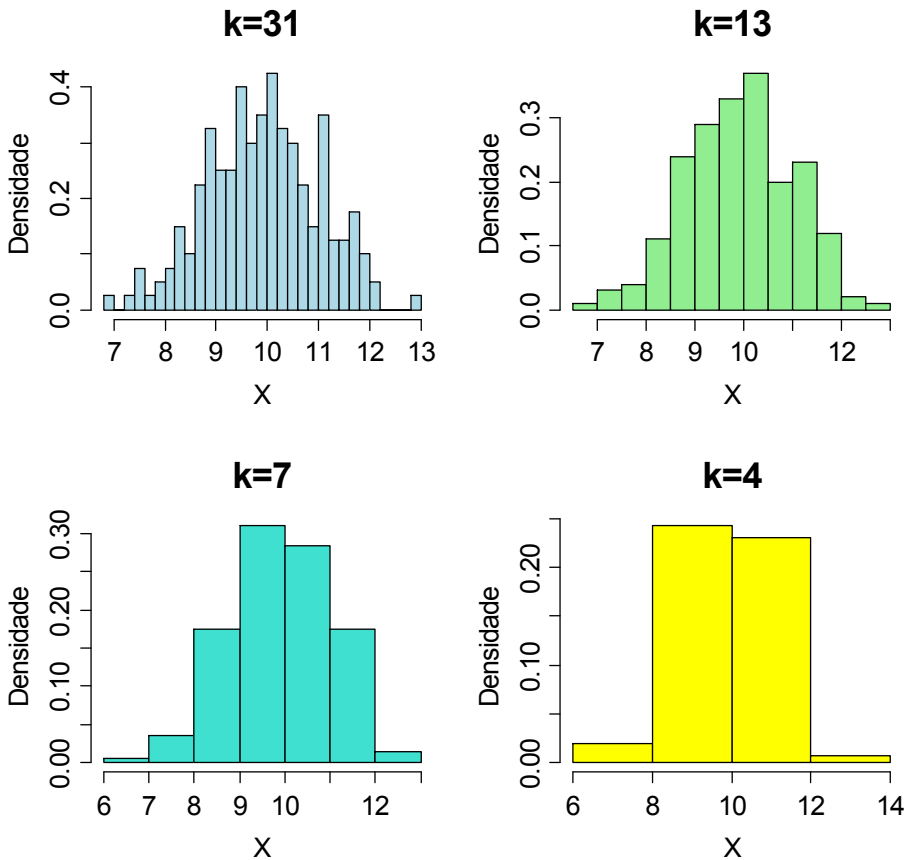
Representação gráfica: Histograma

Exemplo: Histograma da quantidade de CO_2 com a densidade aproximada.



Representação gráfica: Histograma

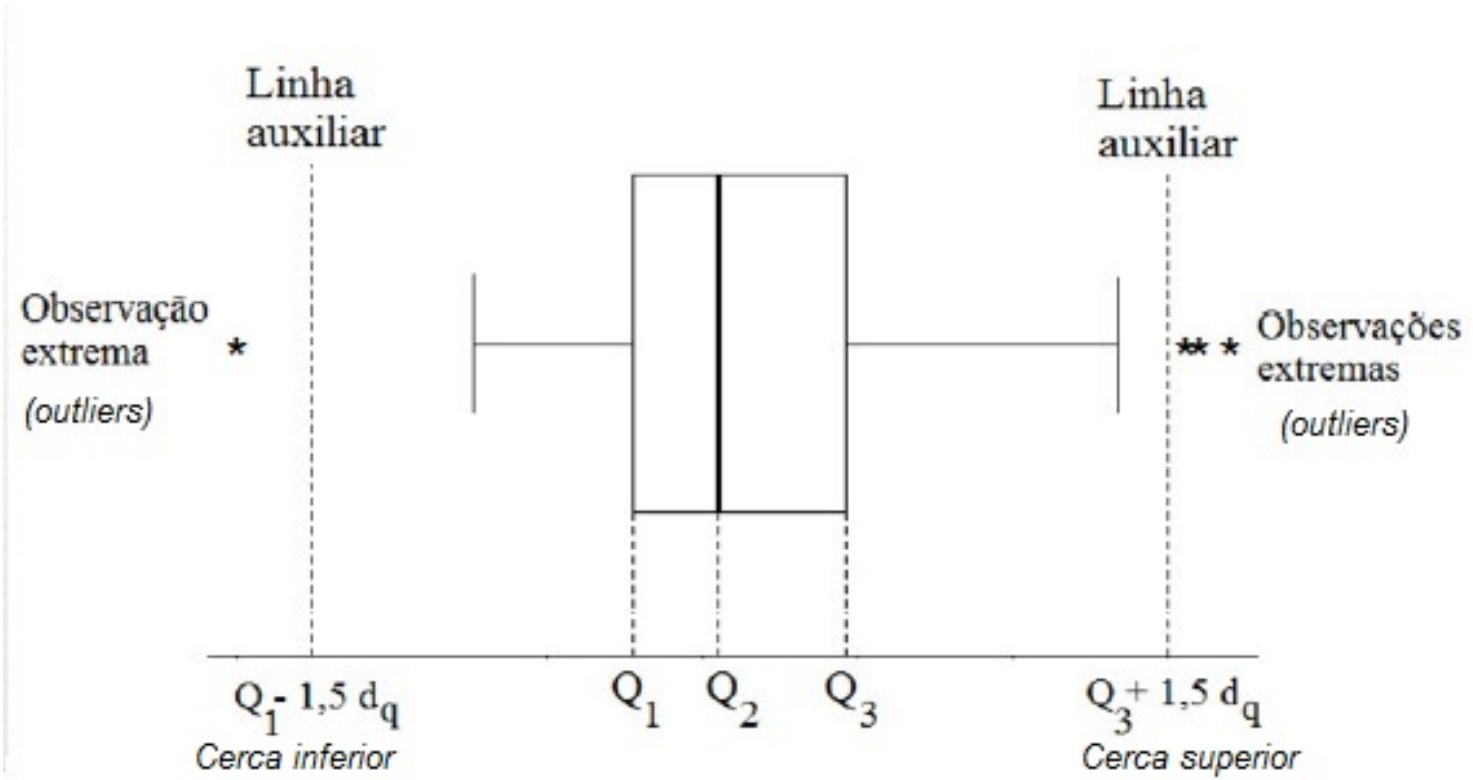
Exemplo: Diferentes números de classes (k).



Observação: Na construção de um histograma, quanto maior for n (tamanho amostral), melhor.

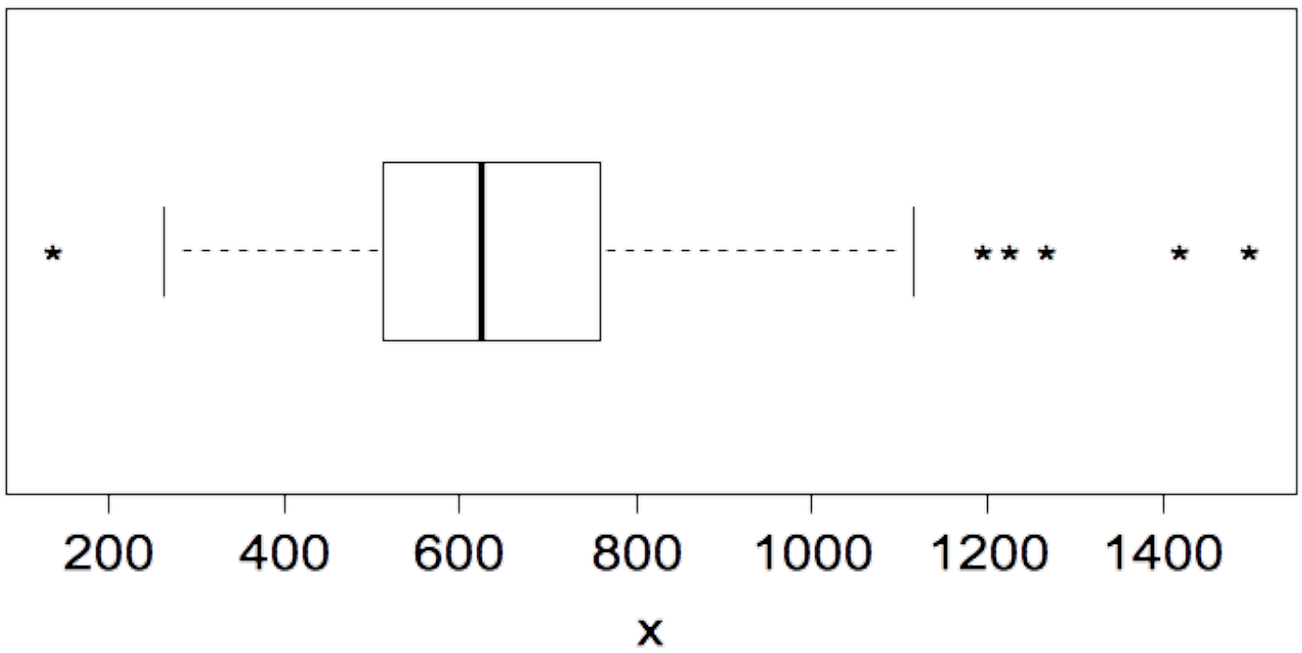
Representação gráfica: *Box plot*

Ilustração: $d_q = Q_3 - Q_1$ é a distância interquartílica.



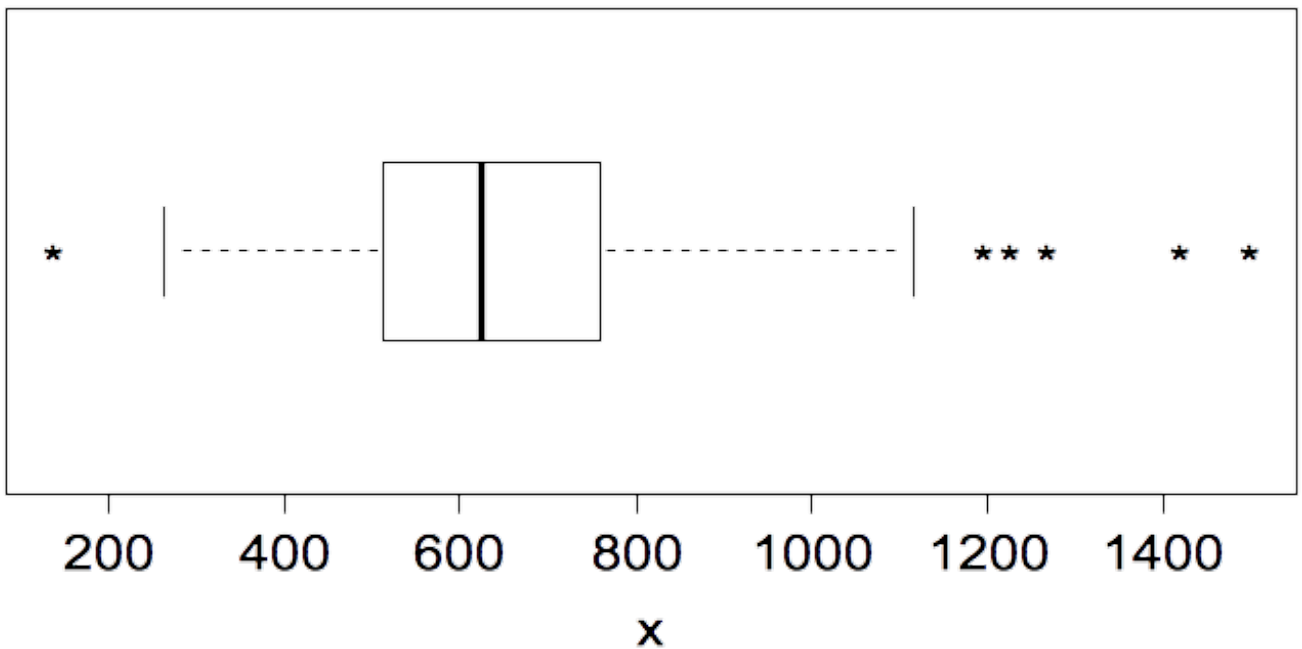
Representação gráfica: *Box plot*

O que é possível observar em um gráfico de caixa?



Representação gráfica: *Box plot*

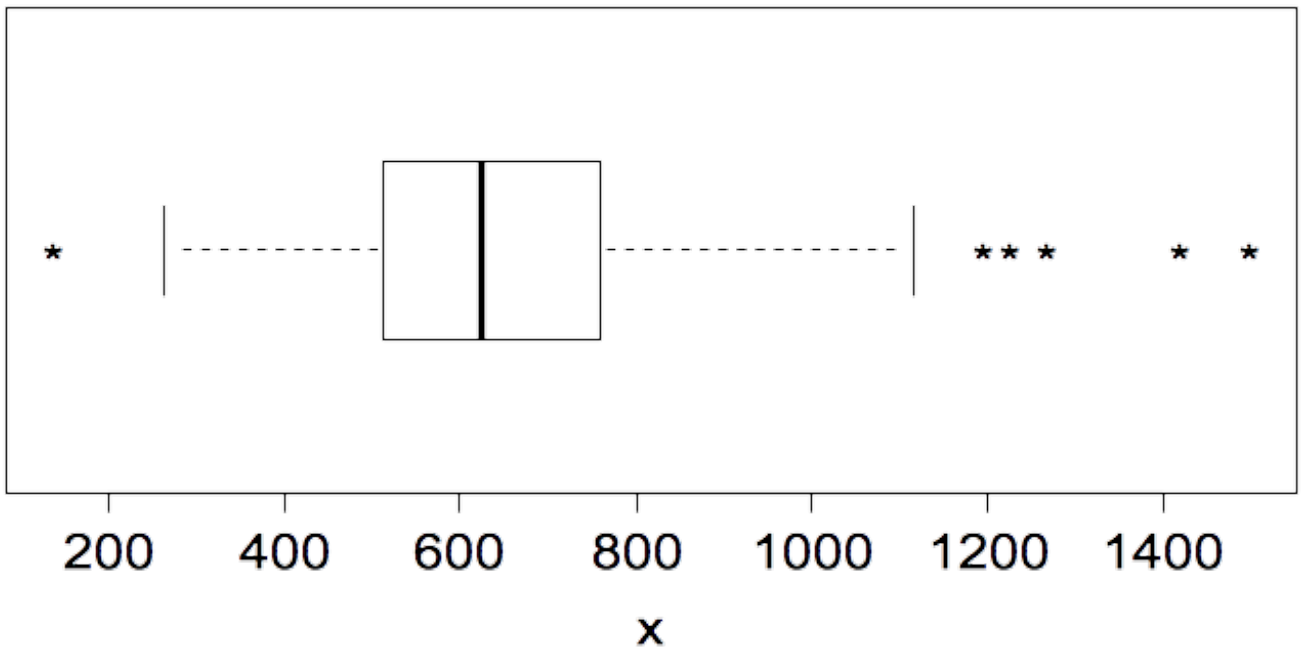
O que é possível observar em um gráfico de caixa?



Respostas: Medida de posição. Medida de dispersão. Simetria. Valores extremos.

Representação gráfica: *Box plot*

O que é possível observar em um gráfico de caixa?

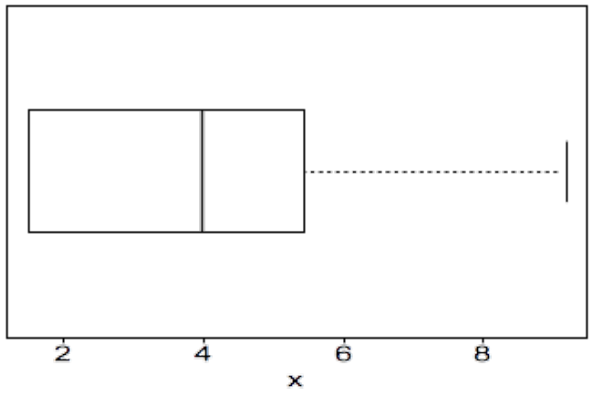


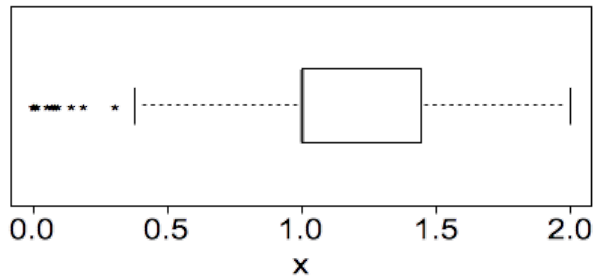
Respostas: Medida de posição. Medida de dispersão. Simetria. Valores extremos.

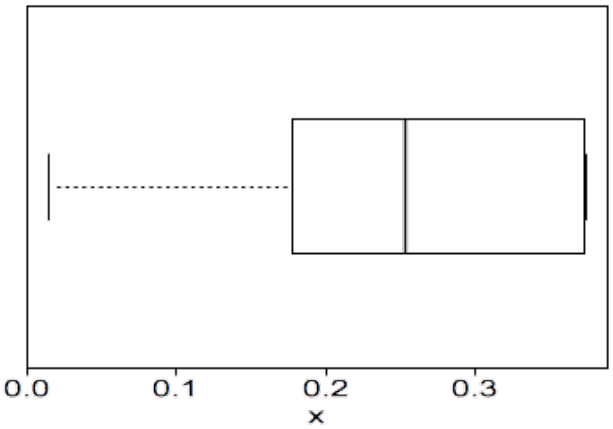
No R: boxplot()

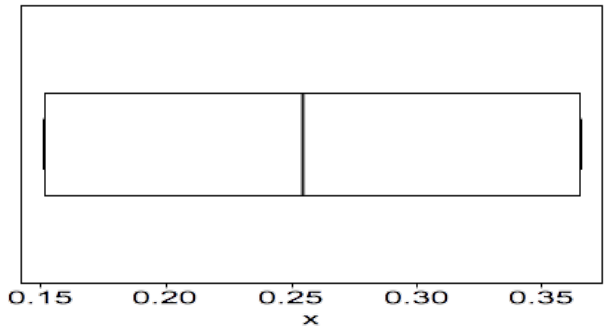
Representação gráfica: *Box plot*

Exercício: Descreva conjuntos de dados correspondentes a cada um dos gráficos.





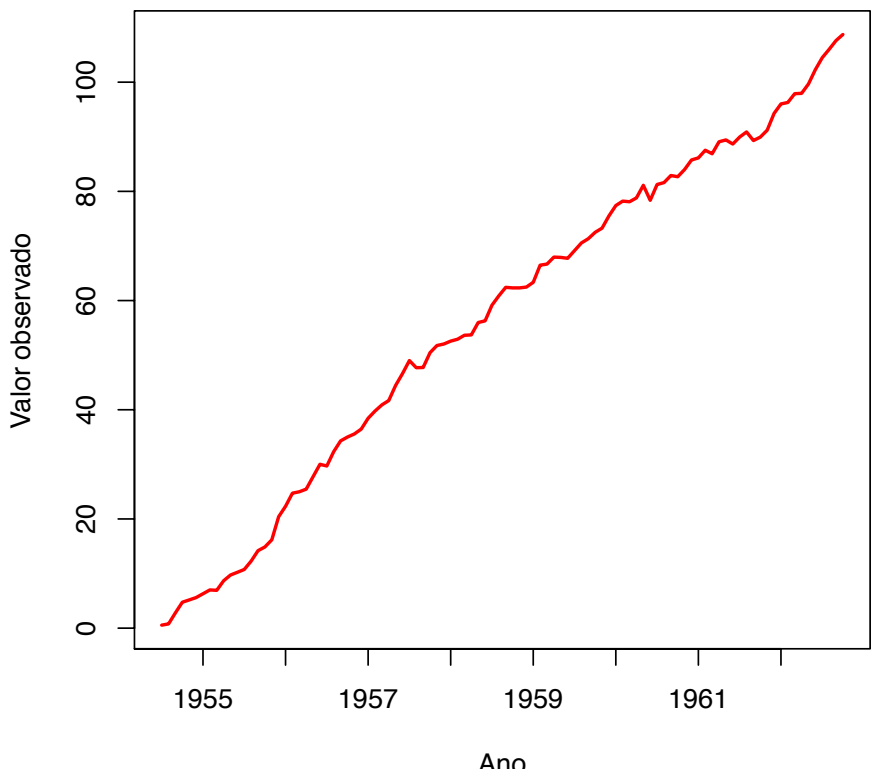




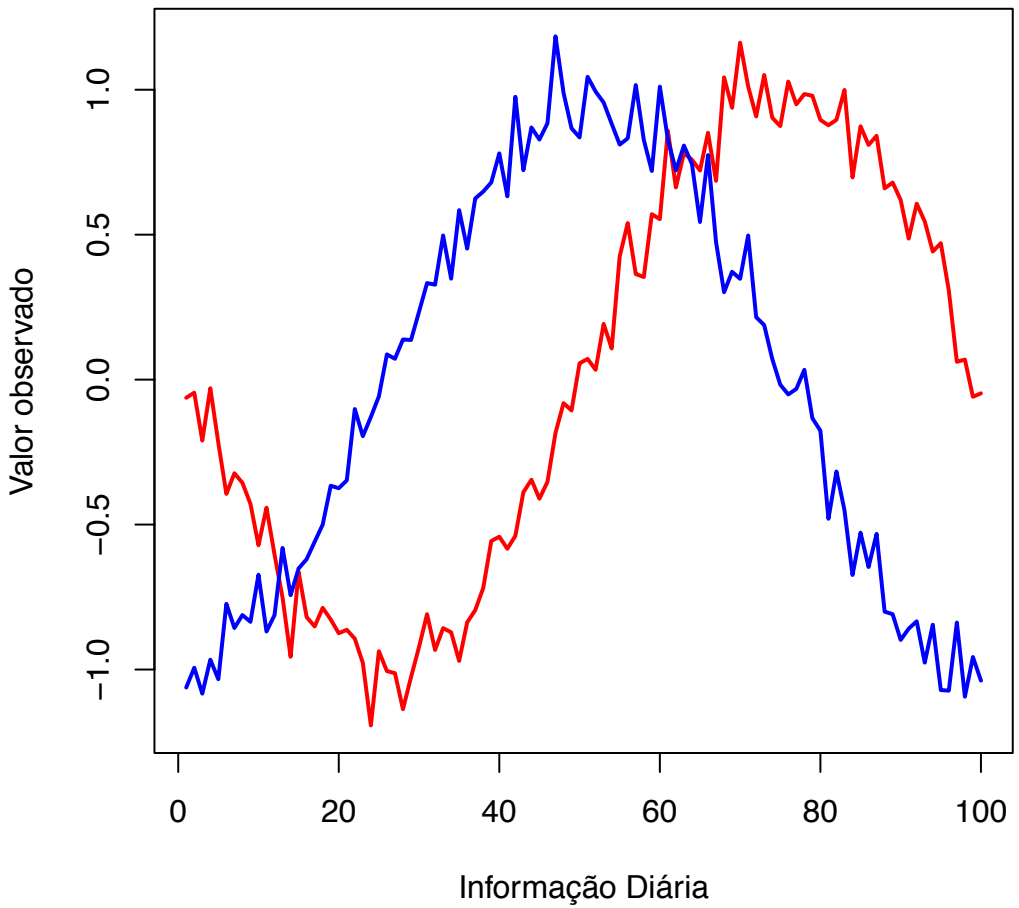
Representação gráfica: Gráfico de séries temporais

Gráfico de séries temporais

- É um gráfico de dados de **série temporal**.
- São dados quantitativos que foram coletados em pontos diferentes no tempo.
- Um ponto é ligado a outro por um segmento de reta.



Representação gráfica: Gráfico de séries temporais



Testes de hipóteses

Objetivo: Verificar se os dados dão evidências que apóiem ou não uma hipótese (estatística).

Testes de hipóteses

Objetivo: Verificar se os dados dão evidências que apóiem ou não uma hipótese (estatística).

Exemplo: Um fabricante A afirma que para seu produto a resistência média à tração é de 145 kg, com desvio padrão de 12 kg. Um fabricante B produz um similar com 155 kg e desvio padrão 20 kg. Se um lote desse produto é selecionado e não conhecemos o fabricante, como decidir qual é o fabricante?

Testes de hipóteses

Objetivo: Verificar se os dados dão evidências que apóiem ou não uma hipótese (estatística).

Exemplo: Um fabricante A afirma que para seu produto a resistência média à tração é de 145 kg, com desvio padrão de 12 kg. Um fabricante B produz um similar com 155 kg e desvio padrão 20 kg. Se um lote desse produto é selecionado e não conhecemos o fabricante, como decidir qual é o fabricante?

- Suponha que uma amostra de 25 produtos é selecionada e testada, sendo sua média igual a $\overline{x}$.

Testes de hipóteses

Regra de decisão: Se $\overline{x} \leq 150$ decidimos que A é o produtor, caso contrário B.

.

Testes de hipóteses

Regra de decisão: Se $\overline{x} \leq 150$ decidimos que A é o produtor, caso contrário B.

- Suponha que $\overline{x} = 148$. Então escolhemos .

Testes de hipóteses

Regra de decisão: Se $\overline{x} \leq 150$ decidimos que A é o produtor, caso contrário B.

- Suponha que $\overline{x} = 148$. Então escolhemos A.

Testes de hipóteses

Regra de decisão: Se $\overline{x} \leq 150$ decidimos que A é o produtor, caso contrário B.

- Suponha que $\overline{x} = 148$. Então escolhemos A.

Perguntas: Podemos estar errados? Uma amostra de B pode ter média 148?

Testes de hipóteses

Hipóteses

H_0 : Produtor é B, isto é, $X \sim \mathcal{N}(155, 20^2)$.

H_1 : Produtor é A, isto é, $X \sim \mathcal{N}(145, 12^2)$.

Testes de hipóteses

Hipóteses

H_0 : Produtor é B, isto é, $X \sim \mathcal{N}(155, 20^2)$.
H_1 : Produtor é A, isto é, $X \sim \mathcal{N}(145, 12^2)$.

Tipos de erros

Testes de hipóteses

Hipóteses

H_0 : Produtor é B, isto é, $X \sim \mathcal{N}(155, 20^2)$.

H_1 : Produtor é A, isto é, $X \sim \mathcal{N}(145, 12^2)$.

Tipos de erros

- **Erro tipo I:** dizer que os produtos são de A quando na realidade são de B.

Testes de hipóteses

Hipóteses

H_0 : Produtor é B, isto é, $X \sim \mathcal{N}(155, 20^2)$.

H_1 : Produtor é A, isto é, $X \sim \mathcal{N}(145, 12^2)$.

Tipos de erros

- **Erro tipo I:** dizer que os produtos são de A quando na realidade são de B.
- **Erro tipo II:** dizer que os produtos são de B quando na realidade são de A.

Testes de hipóteses

Hipóteses

H_0 : Produtor é B, isto é, $X \sim \mathcal{N}(155, 20^2)$.

H_1 : Produtor é A, isto é, $X \sim \mathcal{N}(145, 12^2)$.

Tipos de erros

- **Erro tipo I:** dizer que os produtos são de A quando na realidade são de B.
- **Erro tipo II:** dizer que os produtos são de B quando na realidade são de A.

De forma geral

- **Erro tipo I:** rejeitar H_0 quando H_0 é verdadeira.
- **Erro tipo II:** não rejeitar H_0 quando H_0 é falsa.

Testes de hipóteses

Região crítica: região de rejeição de H_0,

$$RC = \{y \in \mathbb{R} | y \leq 150\}.$$

Testes de hipóteses

Região crítica: região de rejeição de H_0,

$$RC = \{y \in \mathbb{R} | y \leq 150\}.$$

- $P(\text{Erro tipo I}) = P(\overline{X} \in RC | H_0 \text{ é verdadeira})$

$$P(\overline{X} \leq 150 | \overline{X} \sim \mathcal{N}(155, (20/\sqrt{25})^2)) = 10,56\%.$$

Testes de hipóteses

Região crítica: região de rejeição de H_0,

$$RC = \{y \in \mathbb{R} | y \leq 150\}.$$

- $P(\text{Erro tipo I}) = P(\overline{X} \in RC | H_0 \text{ é verdadeira})$

$$P(\overline{X} \leq 150 | \overline{X} \sim \mathcal{N}(155, (20/\sqrt{25})^2)) = 10,56\%.$$

- $P(\text{Erro tipo II}) = P(\overline{X} \notin RC | H_0 \text{ é falsa})$

$$P(\overline{X} > 150 | \overline{X} \sim \mathcal{N}(145, (12/\sqrt{25})^2)) = 1,88\%.$$

Testes de hipóteses

Região crítica: região de rejeição de H_0,

$$RC = \{y \in \mathbb{R} | y \leq 150\}.$$

- $P(\text{Erro tipo I}) = P(\overline{X} \in RC | H_0 \text{ é verdadeira})$

$$P(\overline{X} \leq 150 | \overline{X} \sim \mathcal{N}(155, (20/\sqrt{25})^2)) = 10,56\%.$$

- $P(\text{Erro tipo II}) = P(\overline{X} \notin RC | H_0 \text{ é falsa})$

$$P(\overline{X} > 150 | \overline{X} \sim \mathcal{N}(145, (12/\sqrt{25})^2)) = 1,88\%.$$

Nota: P(Erro tipo I) $>$ P(Erro tipo II).

Testes de hipóteses

Pergunta: E se mudarmos a RC?

Testes de hipóteses

Pergunta: E se mudarmos a *RC*?

Resposta: Mudamos os valores de α e β.

Testes de hipóteses

Pergunta: E se mudarmos a *RC*?

Resposta: Mudamos os valores de α e β.

Procedimento: Fixar um dos erros, por exemplo α, e encontrar a regra de decisão que irá corresponder à $P(\text{Erro tipo I}) = \alpha$.

Testes de hipóteses

Pergunta: E se mudarmos a *RC*?

Resposta: Mudamos os valores de α e β.

Procedimento: Fixar um dos erros, por exemplo α, e encontrar a regra de decisão que irá corresponder à $P(\text{Erro tipo I}) = \alpha$.

Exemplo: No exemplo anterior considere $\alpha = 5\%$ e obtenha *RC*.

Testes de hipóteses

Outra situação: Se os produtos não são fabricados por B, mas sabemos que os outro fabricantes produzem com resistência média menor do que 155.

Testes de hipóteses

Outra situação: Se os produtos não são fabricados por B, mas sabemos que os outro fabricantes produzem com resistência média menor do que 155.

H_0: os produtos são de B ($\mu = 155, \sigma = 20$).

Testes de hipóteses

Outra situação: Se os produtos não são fabricados por B, mas sabemos que os outro fabricantes produzem com resistência média menor do que 155.

H_0: os produtos são de B ($\mu = 155, \sigma = 20$).
H_1: os produtos não são de B ($\mu < 155, \sigma$ desconhecido).

Testes de hipóteses

Função característica de operação (função CO) é definida por

$$\beta(\mu) = P(\text{ não rejeitar } |\mu)$$

Testes de hipóteses

Função característica de operação (função CO) é definida por

$$\beta(\mu) = P(\text{ não rejeitar } |\mu)$$

Função poder do teste é dada por

$$\pi(\mu) = 1 - \beta(\mu),$$

ou seja, é a probabilidade de rejeitar H_0 como função de μ.

Testes de hipóteses

Função característica de operação (função CO) é definida por

$$\beta(\mu) = P(\text{ não rejeitar } |\mu)$$

Função poder do teste é dada por

$$\pi(\mu) = 1 - \beta(\mu),$$

ou seja, é a probabilidade de rejeitar H_0 como função de μ.

Outra situação: Se não são fabricados por B, não sabemos por quem são fabricados.

Testes de hipóteses

Função característica de operação (função CO) é definida por

$$\beta(\mu) = P(\text{ não rejeitar } |\mu)$$

Função poder do teste é dada por

$$\pi(\mu) = 1 - \beta(\mu),$$

ou seja, é a probabilidade de rejeitar H_0 como função de μ.

Outra situação: Se não são fabricados por B, não sabemos por quem são fabricados.

H_0: os produtos são de B ($\mu = 155, \sigma = 20$).

Testes de hipóteses

Função característica de operação (função CO) é definida por

$$\beta(\mu) = P(\text{ não rejeitar } |\mu)$$

Função poder do teste é dada por

$$\pi(\mu) = 1 - \beta(\mu),$$

ou seja, é a probabilidade de rejeitar H_0 como função de μ.

Outra situação: Se não são fabricados por B, não sabemos por quem são fabricados.

H_0: os produtos são de B ($\mu = 155, \sigma = 20$).
H_1: os produtos não são de B (μ e σ desconhecidos).

Passos para a construção de um teste de hipóteses

Testes de hipóteses

Passos para a construção de um teste de hipóteses

Passo 1: Fixe qual a hipótese H_0 a ser testada e qual a hipótese alternativa.

Testes de hipóteses

Passos para a construção de um teste de hipóteses

Passo 1: Fixe qual a hipótese H_0 a ser testada e qual a hipótese alternativa.

H_1 : representa a questão a ser respondida, a teoria a ser testada (nova ideia, conjectura).

Testes de hipóteses

Passos para a construção de um teste de hipóteses

Passo 1: Fixe qual a hipótese H_0 a ser testada e qual a hipótese alternativa.

H_1 : representa a questão a ser respondida, a teoria a ser testada (nova ideia, conjectura).

H_0 : anula ou se opõe a H_1 (status quo) (hipótese que ser rejeitada conduza a um erro tipo I mais importante de ser evitado.

Testes de hipóteses

Passos para a construção de um teste de hipóteses

Passo 1: Fixe qual a hipótese H_0 a ser testada e qual a hipótese alternativa.

H_1 : representa a questão a ser respondida, a teoria a ser testada (nova ideia, conjectura).

H_0 : anula ou se opõe a H_1 (status quo) (hipótese que ser rejeitada conduza a um erro tipo I mais importante de ser evitado.

Conclusões possíveis:

Testes de hipóteses

Passos para a construção de um teste de hipóteses

Passo 1: Fixe qual a hipótese H_0 a ser testada e qual a hipótese alternativa.

H_1 : representa a questão a ser respondida, a teoria a ser testada (nova ideia, conjectura).

H_0 : anula ou se opõe a H_1 (status quo) (hipótese que ser rejeitada conduza a um erro tipo I mais importante de ser evitado.

Conclusões possíveis:

- rejeitar H_0 a favor de H_1, pois há evidências suficientes nos dados.

Testes de hipóteses

Passos para a construção de um teste de hipóteses

Passo 1: Fixe qual a hipótese H_0 a ser testada e qual a hipótese alternativa.

H_1 : representa a questão a ser respondida, a teoria a ser testada (nova ideia, conjectura).

H_0 : anula ou se opõe a H_1 (status quo) (hipótese que ser rejeitada conduza a um erro tipo I mais importante de ser evitado.

Conclusões possíveis:

- rejeitar H_0 a favor de H_1, pois há evidências suficientes nos dados.
- não rejeitar H_0, pois não há evidências suficientes.

Testes de hipóteses

Passos para a construção de um teste de hipóteses

Passo 1: Fixe qual a hipótese H_0 a ser testada e qual a hipótese alternativa.

H_1 : representa a questão a ser respondida, a teoria a ser testada (nova ideia, conjectura).

H_0 : anula ou se opõe a H_1 (status quo) (hipótese que ser rejeitada conduza a um erro tipo I mais importante de ser evitado.

Conclusões possíveis:

- rejeitar H_0 a favor de H_1, pois há evidências suficientes nos dados.
- não rejeitar H_0, pois não há evidências suficientes.

Exemplo:

Testes de hipóteses

Passos para a construção de um teste de hipóteses

Passo 1: Fixe qual a hipótese H_0 a ser testada e qual a hipótese alternativa.

H_1 : representa a questão a ser respondida, a teoria a ser testada (nova ideia, conjectura).

H_0 : anula ou se opõe a H_1 (status quo) (hipótese que ser rejeitada conduza a um erro tipo I mais importante de ser evitado.

Conclusões possíveis:

- rejeitar H_0 a favor de H_1, pois há evidências suficientes nos dados.
- não rejeitar H_0, pois não há evidências suficientes.

Exemplo:

H_0 o réu é inocente.

Testes de hipóteses

Passos para a construção de um teste de hipóteses

Passo 1: Fixe qual a hipótese H_0 a ser testada e qual a hipótese alternativa.

H_1 : representa a questão a ser respondida, a teoria a ser testada (nova ideia, conjectura).

H_0 : anula ou se opõe a H_1 (status quo) (hipótese que ser rejeitada conduza a um erro tipo I mais importante de ser evitado.

Conclusões possíveis:

- rejeitar H_0 a favor de H_1, pois há evidências suficientes nos dados.
- não rejeitar H_0, pois não há evidências suficientes.

Exemplo:

H_0 o réu é inocente.

H_1 o réu é culpado.

Testes de hipóteses

Passo 2: Usar teoria e informações para decidir qual estatística (estimador) usar para testar H_0.

Testes de hipóteses

Passo 2: Usar teoria e informações para decidir qual estatística (estimador) usar para testar H_0.

- Teste paramétrico: assume distribuições para as variáveis aleatórias das hipóteses.
- Teste não paramétrico: não assume distribuições para as variáveis aleatórias das hipóteses.

Testes de hipóteses

Passo 2: Usar teoria e informações para decidir qual estatística (estimador) usar para testar H_0.

- Teste paramétrico: assume distribuições para as variáveis aleatórias das hipóteses.
- Teste não paramétrico: não assume distribuições para as variáveis aleatórias das hipóteses.

Passo 3: Fixar α e construir RC.

Testes de hipóteses

Passo 2: Usar teoria e informações para decidir qual estatística (estimador) usar para testar H_0.

- Teste paramétrico: assume distribuições para as variáveis aleatórias das hipóteses.
- Teste não paramétrico: não assume distribuições para as variáveis aleatórias das hipóteses.

Passo 3: Fixar α e construir RC.

Passo 4: Usar a amostra para calcular a estatística do teste (estimativa).

Testes de hipóteses

Passo 2: Usar teoria e informações para decidir qual estatística (estimador) usar para testar H_0.

- Teste paramétrico: assume distribuições para as variáveis aleatórias das hipóteses.
- Teste não paramétrico: não assume distribuições para as variáveis aleatórias das hipóteses.

Passo 3: Fixar α e construir RC.

Passo 4: Usar a amostra para calcular a estatística do teste (estimativa).

Passo 5: Rejeitar H_0 se a estatística pertencer a RC, e não rejeitar, caso contrário.

Testes de hipóteses

Valor *p*: é o nível de significância mais baixo para o qual o valor observado (com base na amostra) é significante (valor mais baixo para o qual não rejeitamos H_0).

Testes de hipóteses

Valor *p*: é o nível de significância mais baixo para o qual o valor observado (com base na amostra) é significante (valor mais baixo para o qual não rejeitamos H_0).

Escala de significância de Fisher

Valor p	0,10	0,05	0,025	0,01	0,005	0,001
Natureza da evidência	marginal	moderada	substancial	forte	muito forte	fortíssima

Testes de hipóteses

Valor *p*: é o nível de significância mais baixo para o qual o valor observado (com base na amostra) é significante (valor mais baixo para o qual não rejeitamos H_0).

Escala de significância de Fisher

Valor p	0,10	0,05	0,025	0,01	0,005	0,001
Natureza da evidência	marginal	moderada	substancial	forte	muito forte	fortíssima

ATENÇÃO: Grande número de amostras pode levar à rejeição da hipótese H_0.

Testes de hipóteses

Teste paramétrico:

- Assume distribuições para as variáveis aleatórias das hipóteses.
- Requer um número menor de observações para atingir o poder de teste necessário.
- Se a distribuição assumida não corresponder à realidade, os resultados/conclusões ficam comprometidos.

Teste não paramétrico

- Não assume distribuições para as variáveis aleatórias das hipóteses.
- Requer um tamanho suficientemente grande de amostras.
- Não considera mais de uma covariável (outra informação observada) simultaneamente.

Testes de hipóteses

Testes paramétricos

- Teste normal
- Teste t
- Teste t pareado
- Teste F
- Teste χ^2
- ANOVA

Testes não paramétricos

- Teste de Wilcoxon
- Teste de Wilcoxon pareado
- Teste de Mann-Whitney
- Teste de Kruskal-Wallis
- Teste dos sinais
- Teste de aderência