Aula 01 - Introdução à Estatística Descritiva

Probabilidade e Estatística - CRT 0018

Prof. Marciel Barros Pereira
Campus de Crateús (Engenharias)

2025.2

ESTATÍSTICA DESCRITIVA

1. Introdução
2. Medidas de Descrição Numérica de Dados
3. Distribuições de Frequências e Histogramas
4. Gráficos de Espalhamento
5. Métodos Gráficos de Descrição de Dados
6. Diagrama de Caule-e-Folha
7. Diagrama de Caixa (box-plot)

Estatística descritiva acompanha a necessidade de:

- Calcular medidas numéricas de localização (média, mediana) e dispersão (desvio-padrão, amplitude, quartis, percentis) de uma amostra de dados
- Interpretar os resultados das medidas calculadas em problemas de descrição de dados
- Determinar e interpretar a distribuição de frequências de um conjunto de dados

Estatística descritiva acompanha a necessidade de:

- Construir, interpretar gráficos de descrição de dados
- Comparar conjuntos de dados com base e ferramentas gráficas
- Determinar se a distribuição de uma população está próxima de ser Normal com base em um conjunto de dados amostrados

Características importantes de qualquer conjunto de dados

- Centro
- Variação
- Distribuição
- Valores atípicos

TIPOS DE DESCRIÇÕES DE DADOS

- NUMÉRICA
  - **Localização** de estatísticas;
  - **Dispersão** (agrupamento);
  - **Proporção**
- GRÁFICA
  - Distribuição dos dados;

MEDIDAS DE DESCRIÇÃO NUMÉRICA

MEDIDAS DE LOCALIZAÇÃO

MÉDIA AMOSTRAL

- Símbolo: $\bar{x}$
- Medida estatística de localização de uma grandeza **quantitativa** e **numérica;**
- Representada pela **soma aritmética** da amostra **dividida pelo seu tamanho**;

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

MÉDIA AMOSTRAL

MÉDIA AMOSTRAL - EXEMPLOS DE MEDIDAS

- Média de acidentes de trânsito graves em rodovias por mês
  - Quantidade: número de ocorrências (número) em cada mês
  - Fórmula: **soma de ocorrências em um ano / 12**
- Média de tempo de viagem de casa para o trabalho:
  - Quantidade: tempo em minutos em cada dia
  - Fórmula: **soma dos tempos de viagem em um mês / 30**

MÉDIA AMOSTRAL

MÉDIA AMOSTRAL - NÃO SE APLICA EM:

- Grandezas qualitativas:
  - “Qual o tipo sanguíneo médio da turma”
- Nesses casos, a medida numérica descritiva adotada é a **proporção**.

Por sua vez, a média é afetada por **valores extremos.**

MEDIDAS DE LOCALIZAÇÃO

MEDIANA $\tilde{x}$

- Símbolo:
- Medida estatística de localização de uma grandeza **quantitativa** e **numérica;**
- **Os dados devem possuir nível ordinal:** serem ordenados crescente.
- Representa um valor numérico intermediário na amostra;
- A mediana divide a amostra em duas partes iguais: **inferior e superior**.

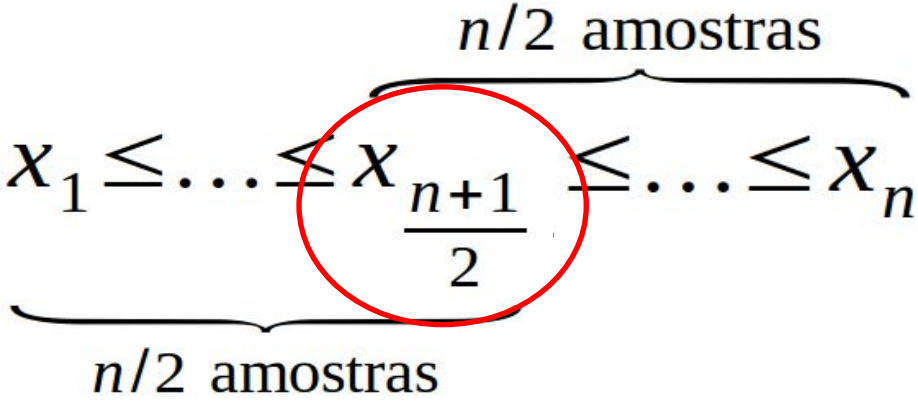
MEDIANA

FÓRMULA

- Ou seja, tomando os dados em ordem crescente:
  - Metade das n amostras está abaixo de $\tilde{x}$
  - Metade das n amostras está acima de $\tilde{x}$

MEDIANA

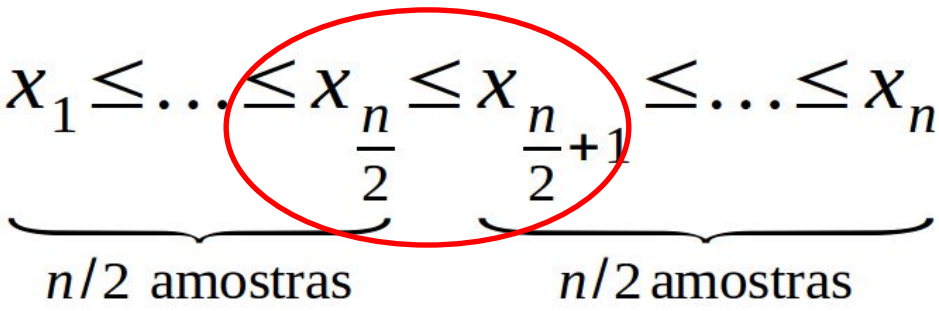
SE O TAMANHO DE AMOSTRA FOR ÍMPAR



$$\tilde{x} = x_{\frac{n+1}{2}}$$

MEDIANA

SE O TAMANHO DE AMOSTRA FOR PAR



$$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

MEDIDAS DE LOCALIZAÇÃO

MODA

- A moda é o valor que ocorre com maior frequência em um conjunto de dados;
- Dependendo do conjunto de dados, ele pode ser
  - **Sem moda** quando nenhum valor se repete
  - **Unimodal** quando existe apenas um valor repetido com maior frequência
  - **Bimodal** quando existem dois valores com a mesma maior frequência
  - **Multimodal** quando mais de dois valores se repetem com a mesma frequência

MEDIDAS DE LOCALIZAÇÃO

MODA

Vantagens

- Resistente à valores extremos
- É a única medida de centro que pode ser usada para dados
- qualitativos

Desvantagens

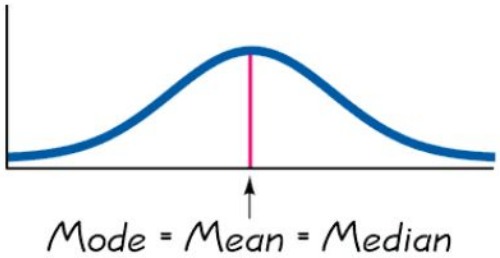
- É uma medida viesada

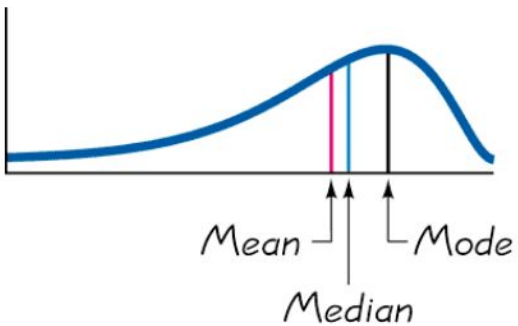
Skewness e Curtose

Skewness: Medida de assimetria de uma massa de dados

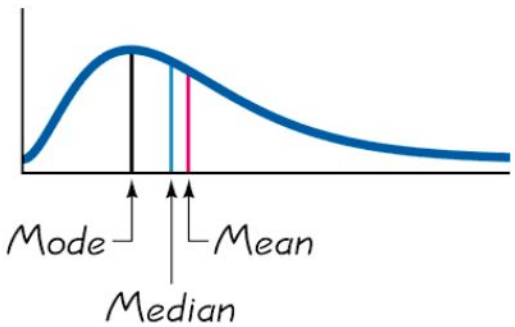
Curtose: relacionada com a dispersão dos dados (espalhamento) em relação a uma distribuição normal

MEDIDAS DE LOCALIZAÇÃO



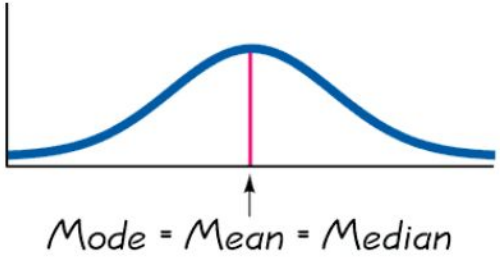


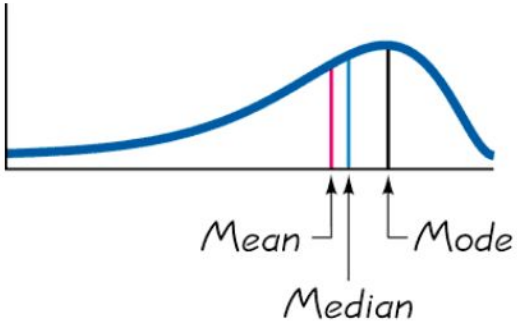
(a) Skewed to the Left (Negatively)



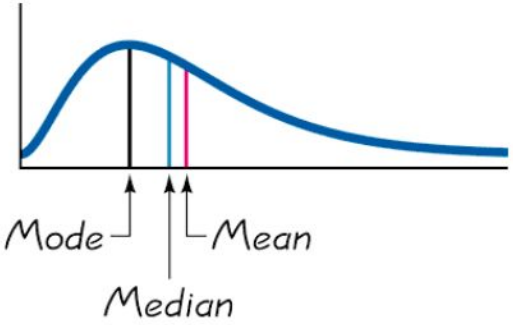
(c) Skewed to the Right (Positively)

MEDIDAS DE LOCALIZAÇÃO

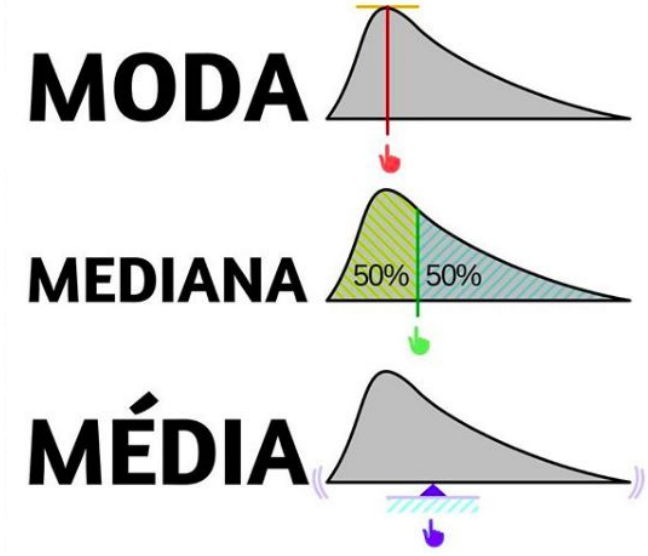




(a) Skewed to the Left (Negatively)



(c) Skewed to the Right (Positively)



MEDIDAS DE LOCALIZAÇÃO

Exemplo: Os dados abaixo se referem ao percentual de cobertura de vegetação em duas áreas de uma floresta.

Área A: 43 47 48 51 51 55 55 57 59

Área B: 20 22 45 46 53 54 56 57

Calcule a média, a mediana e a moda para as áreas A e B

MEDIDAS DE DISPERSÃO

AMPLITUDE

- Símbolo (A)
- Distância entre o maior e o menor valor observado da amostra;

FÓRMULA:

$$A = x_{\max} - x_{\min}$$

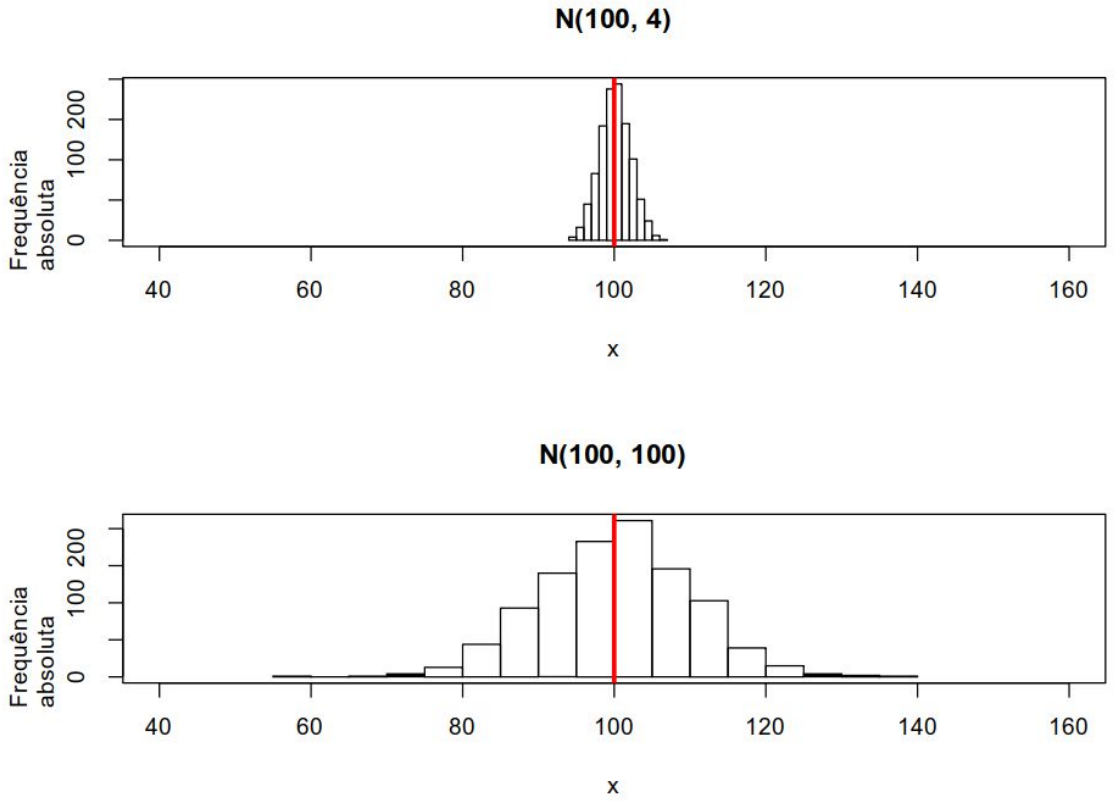
MEDIDAS DE DISPERSÃO

O resumo de um conjunto de dados exclusivamente por uma medida de centro, **esconde** toda a informação sobre a variabilidade do conjunto de observações;

Não é possível analisar um conjunto de dados apenas através de uma medida de tendência central;

Por isso precisamos de medidas que resumem a **variabilidade** dos dados em relação à um valor central;

MEDIDAS DE DISPERSÃO



MEDIDAS DE DISPERSÃO

VARIÂNCIA (s^2) E DESVIO PADRÃO (s)

- Representam o **desvio** (diferença média) entre uma amostra escolhida ao acaso e a média amostral.
- Indicam o quão longe estará uma amostra observada da média amostral;

FÓRMULAS: Variância:

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

Desvio Padrão:

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

MEDIDAS DE DISPERSÃO

VARIÂNCIA (s^2) E DESVIO PADRÃO (s)

- A variância amostral s^2 é considerada um estimador não viesado da variância populacional σ^2;
- É utilizada em diversos métodos estatísticos e caracteriza todas as distribuições de probabilidade
- No entanto, as unidades da **variância são diferentes das unidades dos dados originais** (são medidas ao quadrado, como notas ao quadrado ou cm^2);

MEDIDAS DE DISPERSÃO

Propriedades do desvio-padrão

- É uma medida de variação de todos os dados em relação à média;
- É sempre positivo ou nulo;
- Valores mais distantes da média tem desvio-padrão maior Valores mais próximos da média tem desvio-padrão menor;
- A unidade do desvio-padrão é a mesma dos dados originais (por exemplo notas ou cm);
- A inclusão de valores extremos pode afetar drasticamente o valor do desvio-padrão;

MEDIDAS DE PROPORÇÃO

PROPORÇÃO (*p*)

- Utilizada para representar numericamente grandezas qualitativas, como categorias ou rótulos;
- Representa o percentual de ocorrências de uma determinada categoria considerando a quantidade total de itens;

FÓRMULA:

Nº de elementos da categoria/ Nº total de elementos

MEDIDAS DE DESCRIÇÃO NUMÉRICA

REVISÃO DE CONCEITOS

- Localização: Média e Mediana;
- Dispersão: Amplitude e Variância;
- Proporção;

EXEMPLO: IDADE DOS PARTICIPANTES DA TURMA

MEDIDAS DE DESCRIÇÃO GRÁFICA

DADOS TABULADOS

Dados tabulados são representações de informações em formato de tabela;

- Em geral, formatos ***.xls*** (Excel) e ***.csv*** (Maioria dos programas de análise de dados), mas podem ser encontrados em formatos .dat, .ods, .tab etc.
- Formato ***.csv***: valores separados por vírgulas;

DADOS TABULADOS

- Tais dados em geral são disponibilizados na forma de DATASETS - conjuntos de dados - públicos ou privados, de interesse acadêmico, econômico ou social.
- Alguns datasets:
  - http://dados.gov.br/dataset
  - http://dados.fortaleza.ce.gov.br/catalogo/dataset
  - https://datapedia.info/
  - http://www.inmet.gov.br/portal/index.php?r=estacoes/estacoesAutomaticas

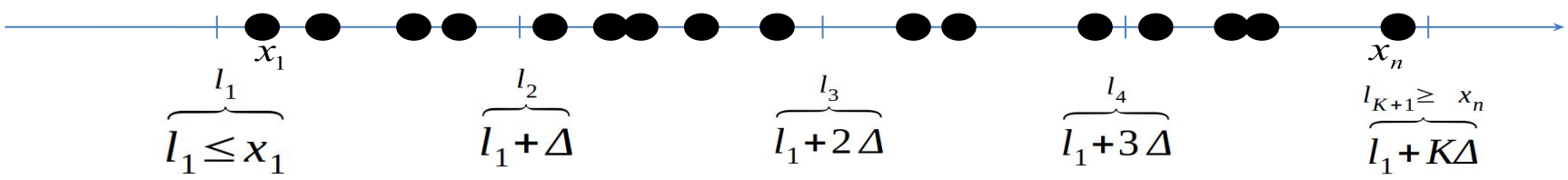
DISTRIBUIÇÃO DE FREQUÊNCIAS

Forma de quantificar ocorrência das amostras em determinadas faixas de valores;

- Intervalos de Classe: divide-se a amplitude ***A*** da amostra em um certo número ***K*** de intervalos (classes)
- Quantos? Sugere-se:
- Dessa forma, cada intervalo tem largura de:

$$K \cong \sqrt{n} \ \text{(inteiro)}$$

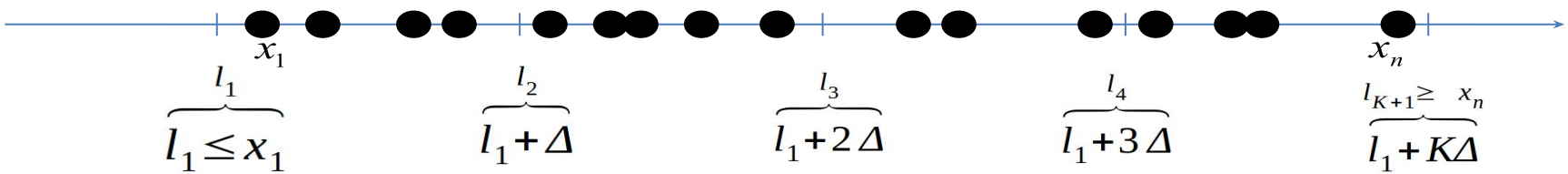
$$\Delta \geq \frac{A}{K}$$



DISTRIBUIÇÃO DE FREQUÊNCIAS

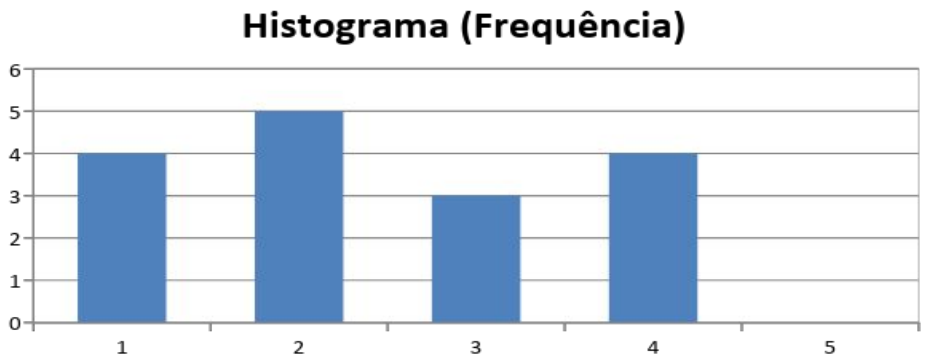
- Frequência: Conta-se o número de amostras dentro de cada intervalo
- Frequência Relativa: Determina-se a porcentagem de amostras dentro de cada intervalo (dividindo a frequência pelo tamanho *n* da amostra)
- Frequência Acumulada: Somam-se as frequências relativas de todas as classes anteriores a atual (incluindo-se a atual)

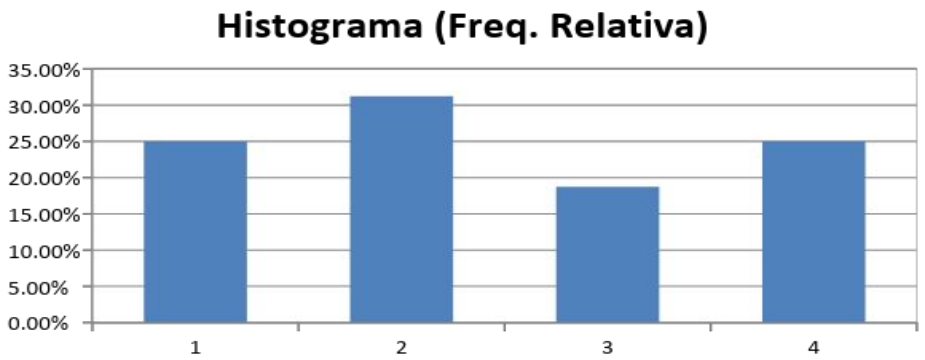
Classes	Frequência	Frequência Relativa (%)	Frequência Acumulada
$l_1 \vdash l_2$	4	4/16=25,00%	25,00%
$l_2 \vdash l_3$	5	5/16=31,25%	56,25%
$l_3 \vdash l_4$	3	3/16=18,75%	75,00%
$l_4 \vdash l_5$	4	2/16=25,00%	100,00%



HISTOGRAMA

É um gráfico de barras que representa a distribuição de frequências (ou frequências relativas)

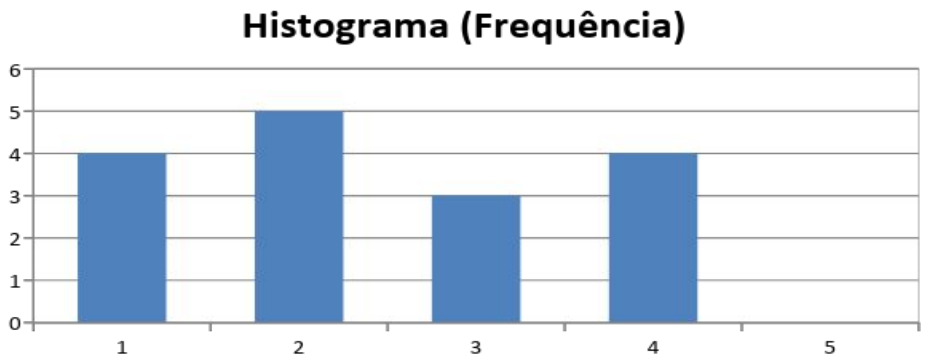


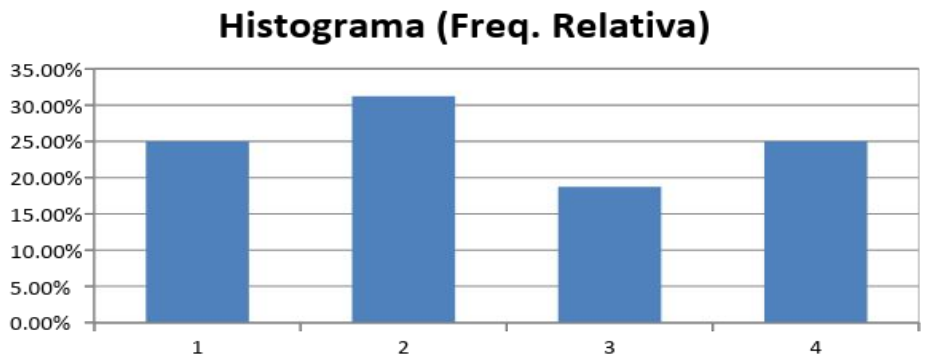


HISTOGRAMA

Como construir um histograma?

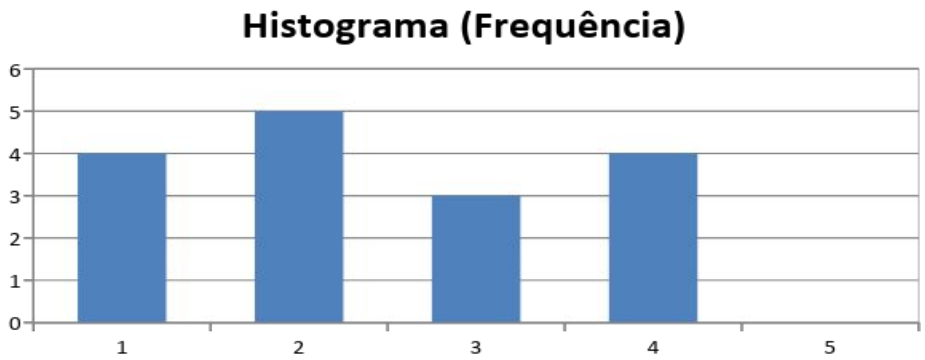
- Implementar uma tabela de frequência relativa;
- Fazer um gráfico de barras cujo eixo x representam os intervalos e y representa a quantidade ou percentual de elementos.

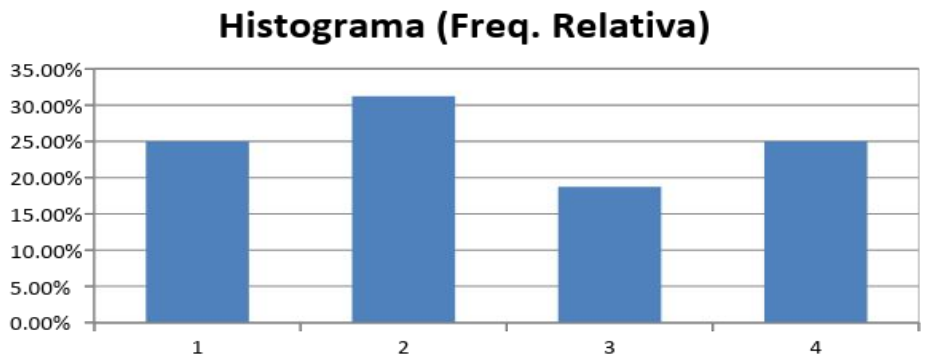




HISTOGRAMA

- OBS: Quando o eixo x representa classes em vez de intervalos, o gráfico é denominado **gráfico de barras**, e não histograma!





HISTOGRAMA - EXEMPLO

EXEMPLO: Observe as amostras de altura de uma espécie de árvore coletadas em quatro áreas diferentes:

Área A: 9.2 10.8 10.6 11.1 12.1 9.6 11.2 8.4 12.9 12.1 14.4 11.1 11.1 9.7 8.4 12.3 10.7 12.9 9.1 12.8
Área B: 12.5 18.5 21.3 14.3 18.5 19.0 10.8 23.1 17.4 10.7 14.3 16.3 18.0 7.1 12.8 14.7 11.3 8.2 13.8
Área C: 21.3 28.7 15.8 24.0 13.7 18.1 12.6 14.6 6.1 19.8 22.3 15.7 16.3 18.2 15.7 6.6 9.3 1.3 19.0
Área D: 13.7 8.6 14.9 10.2 14.0 10.5 15.0 5.2 10.0 11.7 18.7 9.3 7.9 6.5 11.5 12.0 8.3 8.3 9.8 4.7

- Fazer uma distribuição de frequência para cada área. Posteriormente, elaborar um histograma.
- É possível condensar os dados obtidos em um único gráfico?

Quartis e Percentis

Quartis e Percentis
Os **quartis** dividem as observações em quatro partes com (aproximadamente) igual número de amostras em cada cada parte
Os **percentis** dividem as observações em cem partes com (aproximadamente) igual número de amostras em cada cada parte

Determinando os Quartis:
Primeiro quartil (q1): ¼ das amostras estão abaixo de q1 (¾ acima)
Segunda quartil (q2): ½ das amostras estão abaixo de q2 (corresponde à mediana)
Terceiro quartil (q3): ¾ das amostras estão abaixo de q3 (¼ acima)

Quartis e Percentis

$$p = k \times \left(\frac{n+1}{4} \right) = i + f \quad \Rightarrow \quad q_k = x_i + f \times (x_{i+1} - x_i) = \begin{cases} x_i, & \text{se } f = 0 \\ \frac{3x_i + x_{i+1}}{4}, & \text{se } f = 0{,}25 \\ \frac{x_i + x_{i+1}}{2}, & \text{se } f = 0{,}5 \\ \frac{x_i + 3x_{i+1}}{4}, & \text{se } f = 0{,}75 \end{cases}$$

k é o número do quartil (1,2,3)

Parte inteira

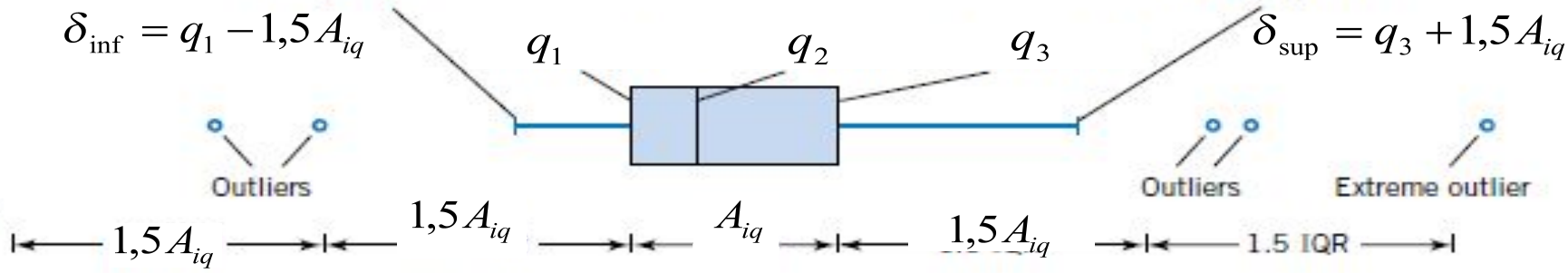
Parte fracionária

BOX PLOT

Também denominado gráfico de caixa, nele podemos visualizar diagramaticamente as seguintes informações:

- Valores máximo e mínimo: $x_{\max}, x_{\min}$
- Quartis (e mediana): $q_1, q_2, q_3 (\tilde{x} = q_2)$
- Amplitude interquartil: $A_{iq} = q_3 - q_1$
- Pontos discrepantes (‘*outliers*’):valores que se distanciam mais de $1{,}5A_{iq}$ abaixo de q_1 ou acima de q_3

BOX PLOT



Filamento se estende até a menor amostra acima do limiar

Filamento se estende até a maior amostra abaixo do limiar

MÉTODOS GRÁFICOS

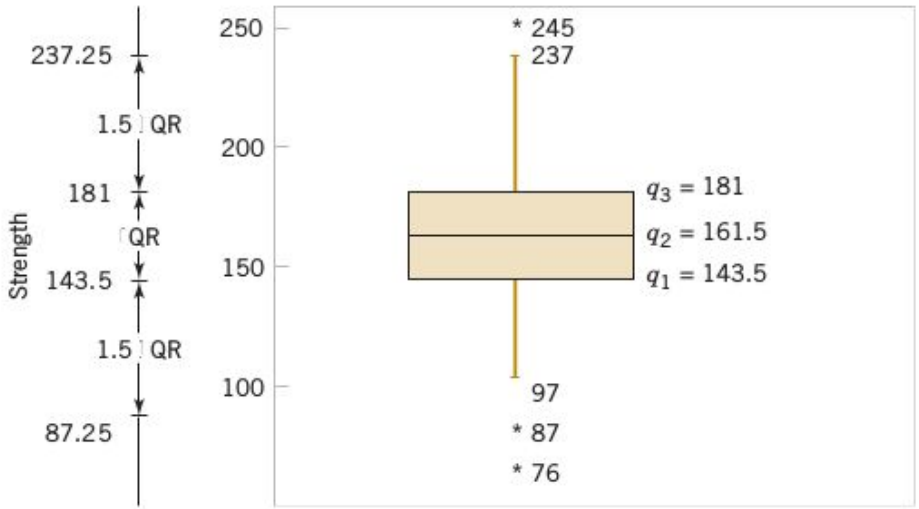


FIGURE 6-14 Box plot for compressive strength data in Table 6-2.

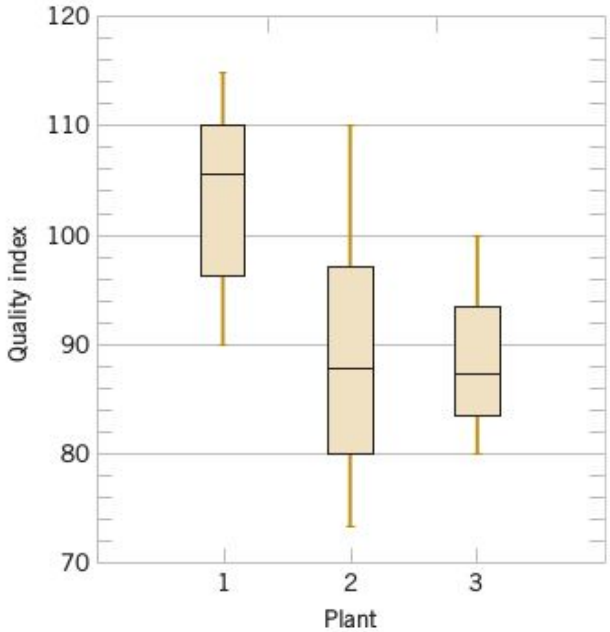


FIGURE 6-15 Comparative box plots of a quality index at three plants.

BOX PLOT - EXEMPLO

EXERCÍCIO: Criar um gráfico contendo box plots das quatro classes de áreas para a variável “altura de árvore”:

Área A:

9.2 10.8 10.6 11.1 12.1 9.6 11.2 8.4 12.9 12.1 14.4 11.1 11.1 9.7 8.4 12.3 10.7 12.9 9.1 12.8

Área B:

12.5 18.5 21.3 14.3 18.5 19.0 10.8 23.1 17.4 10.7 14.3 16.3 18.0 7.1 12.8 14.7 11.3 8.2 13.8

Área C:

21.3 28.7 15.8 24.0 13.7 18.1 12.6 14.6 6.1 19.8 22.3 15.7 16.3 18.2 15.7 6.6 9.3 1.3 19.0

Área D:

13.7 8.6 14.9 10.2 14.0 10.5 15.0 5.2 10.0 11.7 18.7 9.3 7.9 6.5 11.5 12.0 8.3 8.3 9.8 4.7

RESUMO

- Exibição de métodos de descrição estatística - numérica e gráfica
- Importância do uso de representações descritivas para comparação e análise de dados;

Aula 01 - Introdução à Estatística Descritiva

Probabilidade e Estatística - CRT 0018

Prof. Marciel Barros Pereira
Campus de Crateús (Engenharias)

2025.2