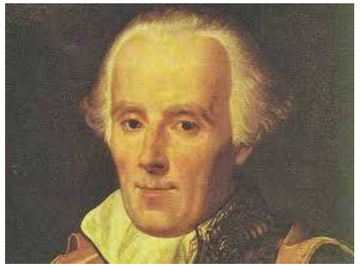
Estatística Descritiva I

O que é Estatística

- A Estatística tem sua origem relacionada com a coleta e construção de tabelas de dados para o governo.
- No século XIX, o avanço na Teoria da Probabilidade e outras metodologias matemáticas, tais como a técnica de Mínimos Quadrados (Legendre, 1805), a Distribuição Normal (Gauss, 1809) e o Teorema do Limite Central (Laplace, 1810) foram fundamentais para o desenvolvimento da Estatística.

Adrien-Marie Legendre (1752-1833)

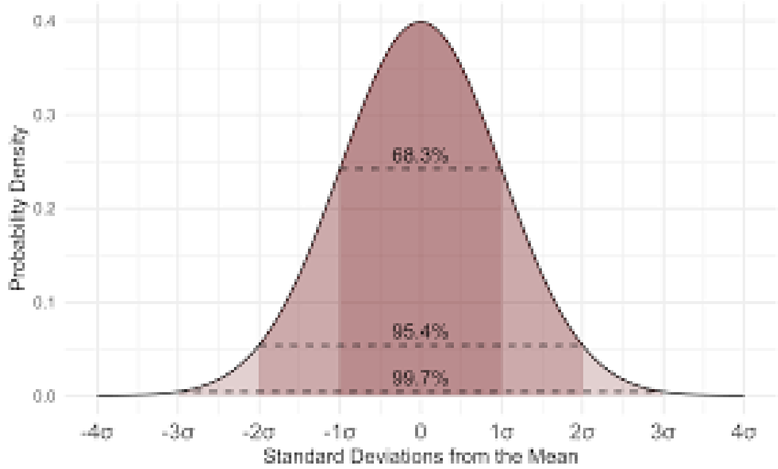
Johann Carl Friedrich Gauss (1777-1855)

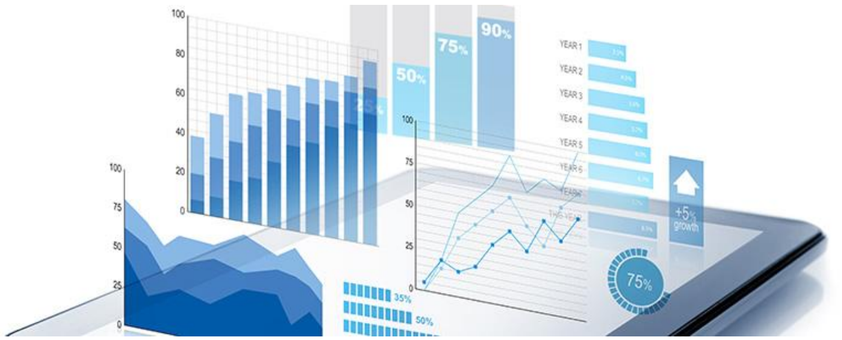


Pierre-Simon Laplace (1749-1827)

O que é Estatística

- No século XX a Estatística desenvolve-se como uma área específica do conhecimento a partir da introdução da Inferência Estatística, metodologia que faz uso da Teoria das Probabilidades e com ampla aplicação em ciências experimentais.
- A Estatística hoje consiste em uma metodologia científica para a obtenção, organização e análise de dados oriundos das mais variadas áreas das ciências experimentais, cujo objetivo principal é auxiliar a tomada de decisões em situações de incerteza.



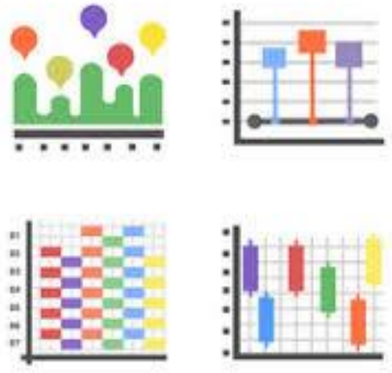


Etapas da análise estatística

População

Amostra

Análise descritiva



Inferência estatística

Conclusões sobre características da população

Amostragem

- Associada à coleta de dados, a tecnologia da amostragem desenvolve técnicas para obtenção de amostras da população de interesse de forma conveniente e informativa.
- As amostras que consideraremos neste curso são do tipo **probabilístico**, em que cada indivíduo da população pode ou não fazer parte da amostra com a mesma probabilidade que os demais.
- Exemplos de uso:
  - Pesquisas de mercado
  - Pesquisas de opinião pública
  - Ensaios clínicos
  - Controle de qualidade

Estatística descritiva

- Etapa inicial da análise utilizada para descrever, organizar e resumir os dados coletados.
- A disponibilidade de uma grande quantidade de dados e de métodos computacionais eficientes tem revigorado esta área da Estatística.

Probabilidade

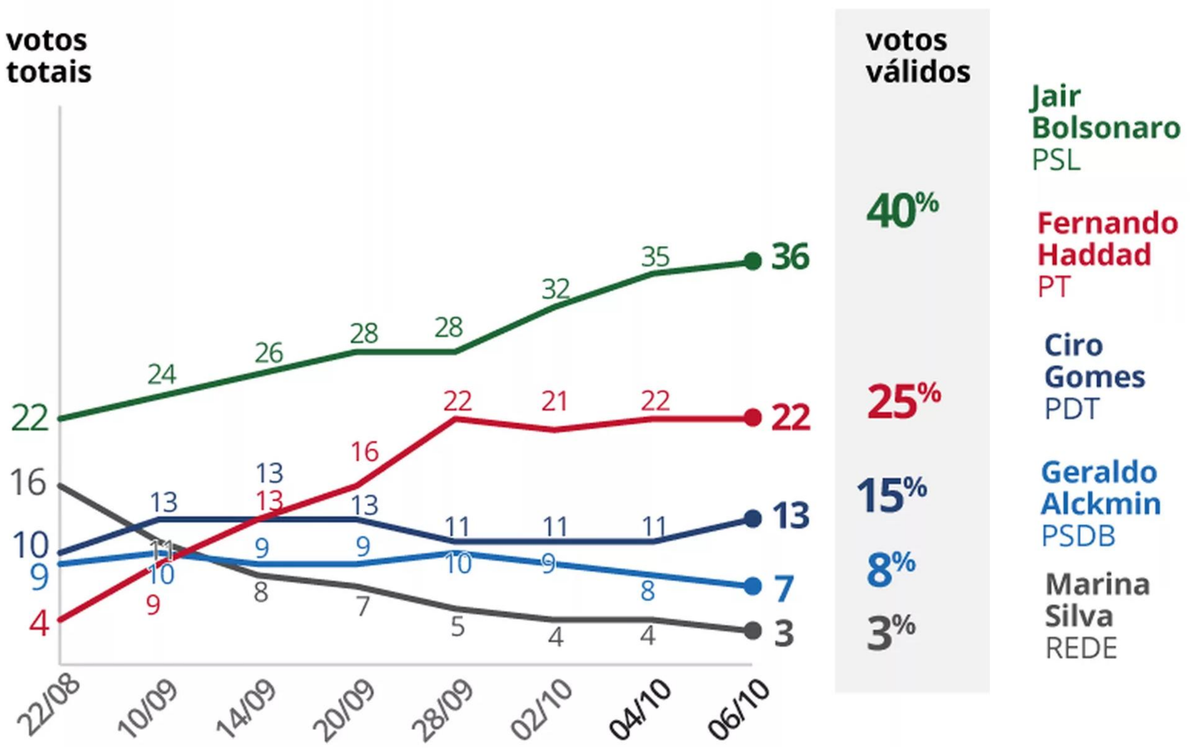
- A Teoria das Probabilidades propicia a modelagem matemática de fenômenos aleatórios, em que está presente a incerteza. É uma ferramenta fundamental para a Inferência Estatística.

Inferência estatística

- Conjunto de técnicas que permite, a partir de dados amostrais, tirar conclusões sobre a população de interesse, com um controle probabilístico sobre a confiabilidade de tais inferências.

Exemplo

Pesquisa eleitoral



Pesquisa DataFolha, com estimativas da intenção de voto para presidente do Brasil para o primeiro turno das eleições de 2018.

Amostra de 19.552 eleitores em 382 municípios.

Margem de erro de 2% com 95% de confiança.

Estatística descritiva

O que fazer com as observações que coletamos?

Primeira etapa da análise estatística:

- Resumos numéricos dos dados (medidas de posição e dispersão)
- Gráficos

Variável: qualquer característica associada a uma população.

Classificação das variáveis:

- **QUALITATIVA**
  - NOMINAL — **sexo, cor dos olhos, estado civil**
  - ORDINAL — **classe social, grau de instrução**
- **QUANTITATIVA**
  - CONTÍNUA — **peso, altura, salário, idade**
  - DISCRETA — **número de filhos, número de carros**

Variáveis Quantitativas

Medidas resumo

MEDIDA DE POSIÇÃO: valor que resume a posição, numa escala, de um conjunto de dados.

Exemplos: Mínimo, Máximo, Moda, Média, Mediana, Percentis

MEDIDAS DE DISPERSÃO: valor que resume a variabilidade de um conjunto de dados.

Exemplos: Amplitude, Distância Interquartil, Variância, Desvio Padrão, Coeficiente de Variação.

Medidas de Posição

- Máximo (***max***): é o maior valor observado
- Mínimo (***min***): é o menor valor observado
- Moda (***mo***): é o valor (ou atributo) que ocorre com a maior frequência.

Exemplo: para o conjunto de observações 4, 5, 4, 6, 5, 8, 4, temos que

$$min = 4$$

$$mo = 4$$

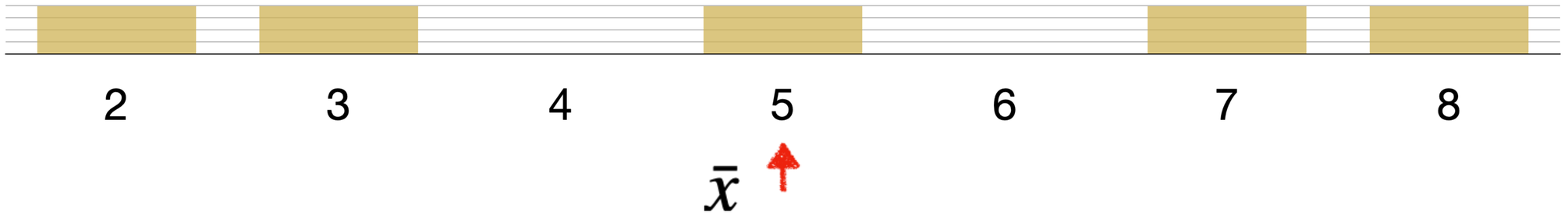
$$max = 8$$

Medidas de Posição

Média: $\bar{x} = \dfrac{x_1 + x_2 + \ldots x_n}{n} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Exemplo: para o conjunto de observações 2, 5, 3, 7, 8, temos que

$$\bar{x} = \frac{2+5+3+7+8}{5} = 5$$



Medidas de Posição

- **Mediana**: é o valor da variável que ocupa a posição central de um conjunto de n **dados ordenados**.

posição da mediana para n dados ordenados → $\frac{n+1}{2}$

Dados: 2, 6, 3, 7, 8 ($n = 5$)

Dados ordenados: 2 3 6 7 8

Posição da mediana: $\frac{5+1}{2} = 3$ → $md = 6$

Medidas de Posição

- **Mediana**: é o valor da variável que ocupa a posição central de um conjunto de n **dados ordenados**.

posição da mediana para n dados ordenados → $\frac{n+1}{2}$

Dados: 4, 8, 2, 1, 9, 6 (n = 6)

Dados ordenados: 1 2 4 6 8 9

Posição da mediana: $\frac{6+1}{2} = 3{,}5$ → $md = \frac{4+6}{2} = 5$

Percentil

O **percentil de ordem p×100** ($0 < p < 1$) em um conjunto de dados de tamanho n é o valor da variável que ocupa a posição $p \times (n + 1)$ do conjunto de dados ordenados.

Casos particulares:

Percentil 50 = mediana ou segundo quartil (md)

Percentil 25 = primeiro quartil (Q_1)

Percentil 75 = terceiro quartil (Q_3)

Percentil 10 = primeiro decil (D_1)

Percentil

Dados: 0,9 1,0 1,7 2,9 3,1 5,3 5,5 12,2 12,9 14,0 33,6 (n = 11)

Posição da mediana: $\frac{11+1}{2} = 6$ → $md = 5{,}3$

Posição de Q_1: $0{,}25 \times (11+1) = 3$ → $Q_1 = 1{,}7$

Posição de Q_3: $0{,}75 \times (11+1) = 9$ → $Q_3 = 12{,}9$

Qual é a interpretação desses valores?

Percentil

Dados: **1**,9 2,0 2,1 2,5 3,0 3,1 3,3 3,7 6,1 7,7 (n = 10)

Posição da mediana: $\frac{10+1}{2} = 5{,}5$ → md = 3,05

Posição de Q_1: 0,25×(10+1) = 2,75 → $Q_1 = \frac{2+2{,}1}{2} = 2{,}05$

Posição de Q_3: 0,75×(10+1) = 8,25 → $Q_3 = \frac{3{,}7+6{,}1}{2} = 4{,}9$

Observação: os softwares utilizam fórmulas de cálculo de percentis um pouco diferentes e, assim, podem fornecer valores um pouco diferentes.

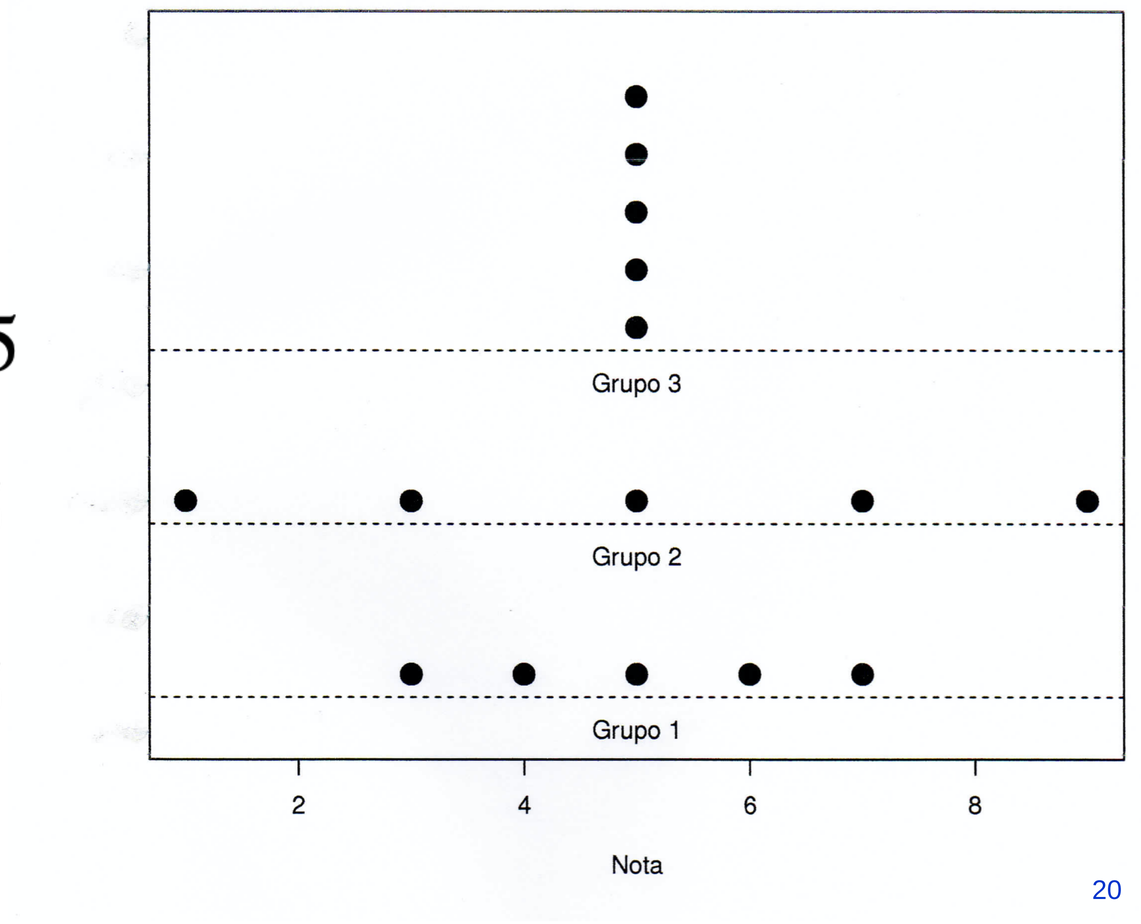
Exemplo

Considere as notas de um teste de 3 grupos de alunos.

Grupo 3: 5, 5, 5, 5, 5 $\quad \bar{x}_3 = 5, md_3 = 5$

Grupo 2: 1, 3, 5, 7, 9 $\quad \bar{x}_2 = 5, md_2 = 5$

Grupo 1: 3, 4, 5, 6, 7 $\quad \bar{x}_1 = 5, md_1 = 5$



Medidas de dispersão

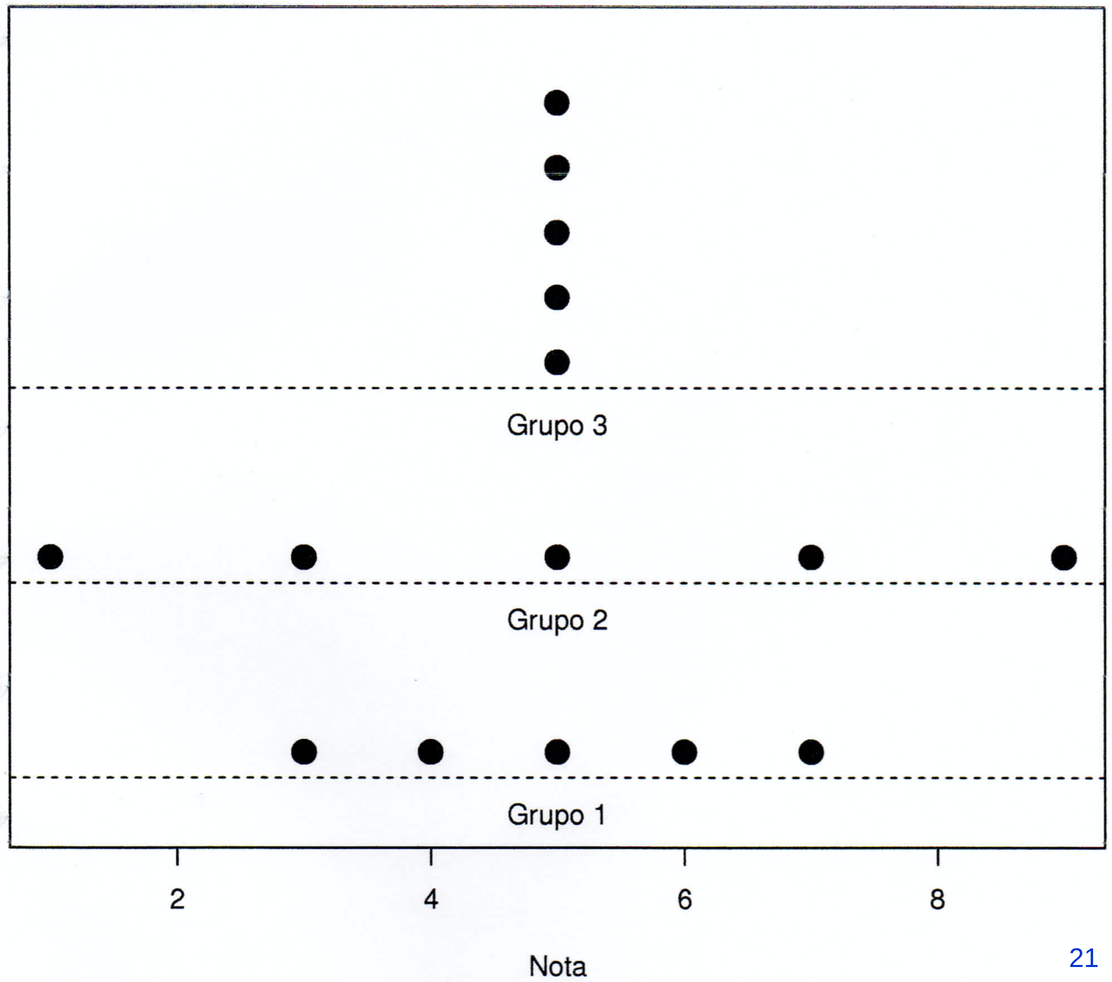
Amplitude

$$A = max - min$$

Grupo 3, $A = 0$

Grupo 2, $A = 8$

Grupo 1, $A = 4$



Medidas de dispersão

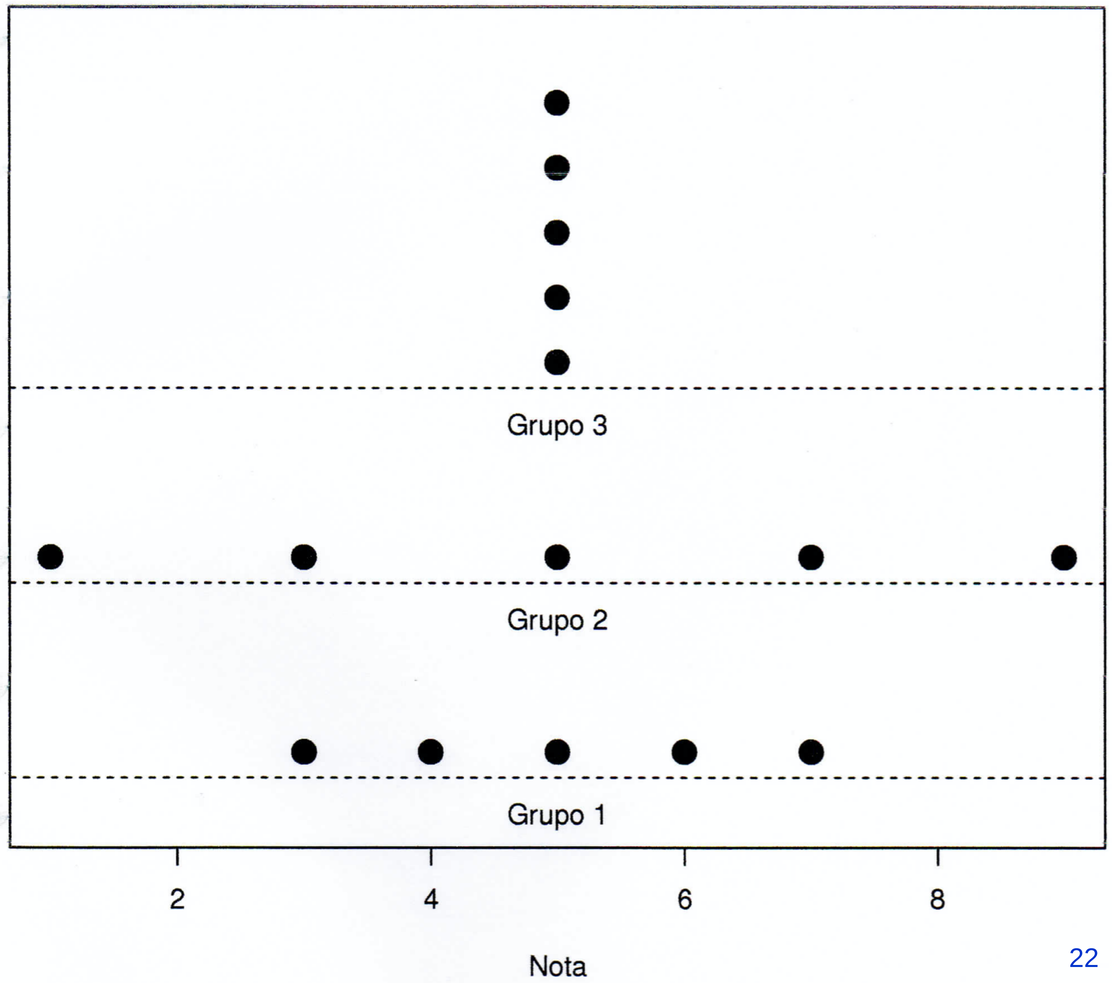
Diferença interquartil

$$DI = Q_3 - Q_1$$

Grupo 3, $DI = 0$

Grupo 2, $DI = 6$

Grupo 1, $DI = 3$



Medidas de dispersão

Variância e desvio padrão

- **Variância**:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- **Desvio padrão**:

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

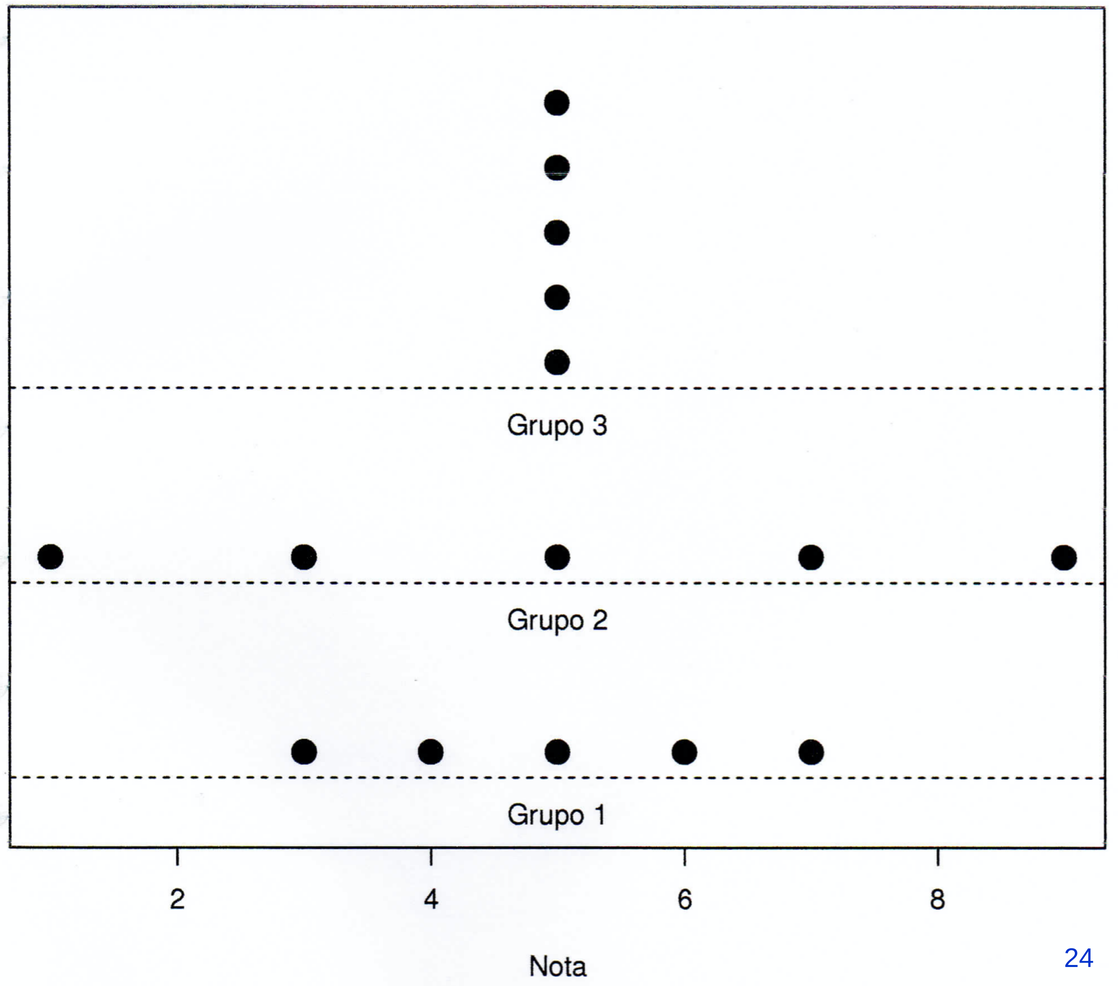
Medidas de dispersão

Variância e desvio padrão

Grupo 3: $s^2 = 0 \Rightarrow s = 0$

Grupo 2: $s^2 = 10 \Rightarrow s = 3{,}16$

Grupo 1: $s^2 = 2{,}5 \Rightarrow s = 1{,}58$



Fórmula alternativa para a variância

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

ou

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

Medidas de dispersão

Coeficiente de Variação

- É uma medida de dispersão relativa à média dos dados
- Elimina o efeito da magnitude dos dados
- É utilizada na comparação de grupos

$$cv = \frac{s}{\bar{x}} \times 100\,\%$$

Exemplo

Altura (em cm) de uma amostra de recém-nascidos e de uma amostra de adolescentes

	Média	**Desvio padrão**	**Coeficiente de variação**
Recém nascidos	50	6	12%
Adolescentes	160	16	10%

Conclusão: Em relação às respectivas médias, as alturas dos adolescentes e dos recém-nascidos apresentam variabilidades semelhantes.

Exemplo

Em um grupo de pacientes foram tomadas as pulsações (batidas por minuto) e dosadas as taxas de ácido úrico (mg/100 ml).

	Média	**Desvio padrão**	**Coeficiente de variação**
Pulsação	68,7	8,7	12,7%
Ácido úrico	4,47	1,03	23%

Conclusão: a variabilidade da pulsação é menor do que a do ácido úrico, sugerindo que a primeira medida é mais estável do que a segunda.