SENAI

Ciência de Dados (CIDA)

Aula 1 – Introdução à Estatística

Prof.: Hugo S. Idagawa

Introdução

- **O que é estatística?**

“A estatística é uma coleção de métodos para planejar experimentos, obter dados, organizá-los, resumi-los, interpretá-los e, a partir deles, extrair conclusões.”

Descritiva: Tem por objetivo descrever o conjunto de dados observado. Envolve a organização, o resumo e a representação desses dados.

Inferencial: Tem por objetivo estabelecar comparações entre os dados. Uma das aplicações é chegar a conclusões sobre um conjunto de dados maior ao qual não temos acesso.

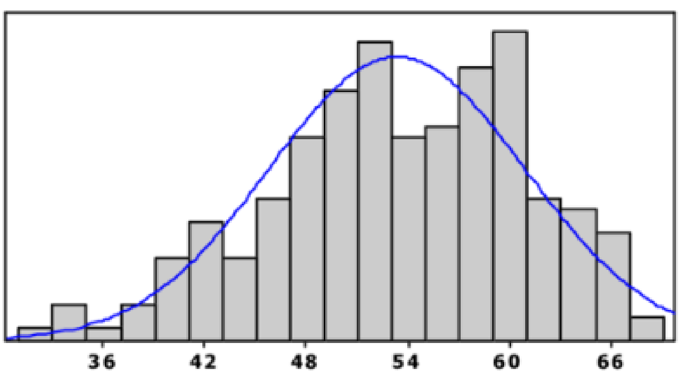
Introdução

Estatística

Descritiva | **Inferencial**

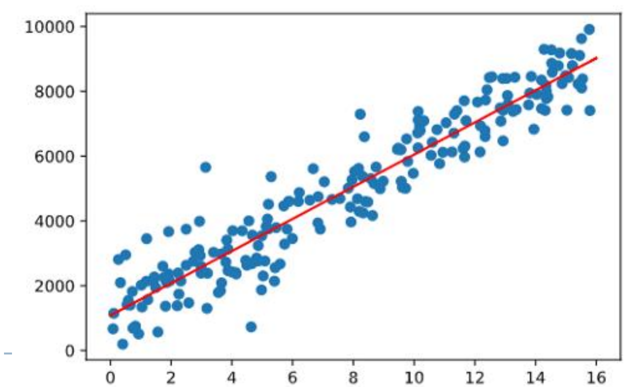
Principais conhecimentos:

- Medidas resumo
- Tendências
- Dispersão



Principais conhecimentos:

- Probabilidades
- Intervalos de confiança
- Testes de hipóteses
- Regressões



Estatística Descritiva

- A estatística descritiva está preocupada em representar os dados usando gráficos e diagramas, cujo objetivo é resumir todos os valores em um (ou alguns) números.

- **Conceitos básicos:**

1) **População:** todos os indivíduos (ou elementos) alvos do estudo.
2) **Amostra:** parte de uma população
3) **Parâmetro:** característica da população. Em geral não é possível (ou é muito caro) encontrar esse valor.
4) **Estimativa:** característica da amostra. Geralmente usamos uma estimativa para aproximar um parâmetro.
5) **Variável:** característica de um elemento/indivíduo da população. Exemplo: “idade dos alunos”: x= 19anos.

Estatística Descritiva

➢ **<u>Medidas Resumo:</u>**

As medidas resumo permitem resumir as informações de uma variável da amostra. Essas medidas são úteis pois dificilmente conseguimos extrair alguma informação quando olhamos para um banco de dados com muitas informações.

Medidas Resumo

Posição

- Média
- Moda
- Mediana

Dispersão

- Variância
- Desvio-padrão
- Desvio médio

Quantis

- Quantis
- Diagrama de caixa (boxplot)

Estatística Descritiva – Medidas Resumo

Para explicar os conceitos das medidas resumo a seguir, vamos utilizar o exemplo a seguir: deseja-se estudar a renda média mensal de uma cidade do Brasil. Como não é possível perguntar para todos os habitantes a renda de cada um, decidiu-se entrevistar uma amostra de 9 indivíduos e organizar o resultado dessas entrevistas na tabela abaixo:

Pessoa	Renda (R$)
1	1500,00
2	3000,00
3	3000,00
4	2200,00
5	5000,00
6	1750,00
7	2800,00
8	4500,00
9	15000,00

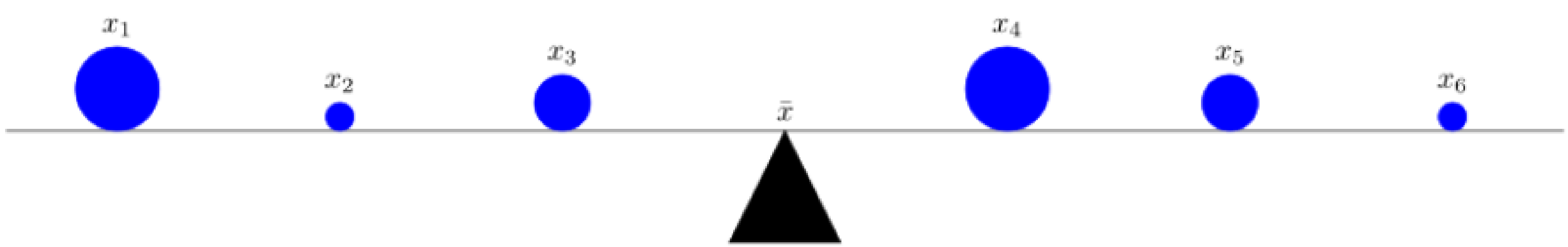
Estatística Descritiva – Medidas Resumo

- **Moda:** a moda de uma variável X é o valor que aparece mais vezes na amostra. Matematicamente, a moda de X é representada por:

mo(X) = valor.

- **Média:** a média pode ser interpretada como sendo o centro de massa de uma barra em que diferentes pesos foram colocados espaçados igualmente sobre ela. Matematicamente, a média é calculada como:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$



Estatística Descritiva – Medidas Resumo

- **Mediana:** a mediana é o valor que divide a sequência ordenada dos valores amostrado em duas partes iguais. A mediana é representada por: md(X) = valor.

OBS: se o número de medições for ímpar, o valor da mediana é o próprio valor central. Se o número de medições for par, o valor da mediana corresponde à média dos dois valores mais próximos do centro da amostragem.

$$md(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & se\ n\ é\ ímpar, \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & se\ n\ é\ par. \end{cases}$$

Estatística Descritiva – Medidas Resumo

- Apesar das medidas de posição serem úteis, elas podem ser insuficientes para representar de forma fiel todos os valores de uma variável. **Exemplo:** resultados das notas de 4 testes diferentes:

Aluno	Teste 1	Teste 2	Teste 3	Teste 4
1	3	1	5	4
2	4	3	5	5
3	5	5	5	5
4	6	7	5	6
5	7	9	5	5
Média				
Moda				
Mediana				

Estatística Descritiva – Medidas Resumo

- No exemplo anterior, observe que todas as medidas de posição são iguais, porém a distribuição dos resultados é bastante diferente entre os testes.

- Assim, para dar uma idéia melhor da distribuição dos valores, utilizamos as medidas de dispersão.

- **Desvio médio:** o desvio médio é a média dos desvios absolutos das amostras. Ele é representado por dm(X) e calculado conforme a equação abaixo:

$$dm(x) = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n}$$

Estatística Descritiva – Medidas Resumo

- **Variância:** a variância é uma medida de dispersão que é útil para determinar o afastamento da média que os dados de uma amostra apresentam. Ela é definida como var(x) e calculada conforme a fórmula abaixo:

$$var(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

- **Desvio padrão:** o desvio padrão é a raiz quadrada da variância. Ela é definida como dp(X) e calculada conforme a fórmula abaixo:

$$dp(x) = \sqrt{var(x)}$$

OBS: a interpretação para todas as medidas de dispersão é a seguinte: quanto menor o desvio padrão (ou a variância, ou o desvio médio), mais homogênea é a variável em estudo.

Estatística Descritiva – Medidas Resumo

- Até o momento, consideramos apenas medidas de resumo para representar informações em relação ao centro dos dados. Assim, para os casos de posições não centrais, é comum utilizar os quantis.

- Um quantil-p ou quantil de ordem p é um valor que deixa 100*p% (0 < p < 1) das observações à sua esquerda.

- Um quantil-p também é representado por Q(p). Por exemplo: um quantil-0,6 pode ser representado por Q(0,6).

- Matematicamente, o Q(p) é definido conforme abaixo:

$$Q(p) = \begin{cases} x_{(i)}, & \text{se } \; p = p_i = (i - 0{,}5)/n,\ i = 1, \ldots, n \\ (1 - f_i)Q(p_i) + f_i Q(p_{i+1}), & \text{se } \; p_i < p < p_{i+1} \\ x_{(1)}, & \text{se } \; 0 < p < p_1 \\ x_{(n)}, & \text{se } \; p_n < p < 1, \end{cases}$$

Estatística Descritiva – Medidas Resumo

- **Exemplo:** determine o quantil-0,90 para os dados abaixo que representam o peso de 16 alunos:

52	56	62	54	52	51	60	61
56	55	56	54	57	67	61	49

- **Passo 1:** ordenação em ordem crescente dos dados

- **Passo 2:** determinação do valor do quantil, sabendo-se que n=16 (número de observações) e p=0,90.

Estatística Descritiva – Medidas Resumo

- **Exemplo:** determine o quantil-0,90 para os dados abaixo que representam o peso de 16 alunos:

52	56	62	54	52	51	60	61
56	55	56	54	57	67	61	49

- **Passo 1:** ordenação em ordem crescente dos dados

49	51	52	52	54	54	55	56
56	56	57	60	61	61	62	67

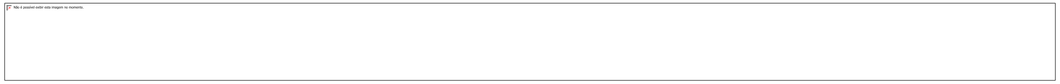
- **Passo 2:** determinação do valor do quantil, sabendo-se que n=16 e p=0,90.

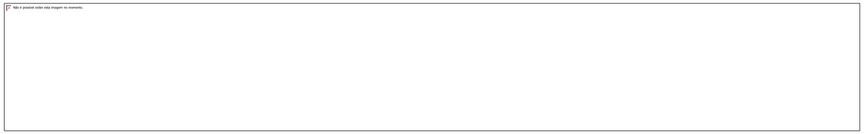
Estatística Descritiva – Medidas Resumo

- **Interpretação do resultado:** Q(0,90)=62 significa que 90% dos dados de peso estão abaixo de 62kg.

49	51	52	52	54	54	55	56
56	56	57	60	61	61	62	67

- **Exemplo 2:** determine o quantil-0,75 para os dados do exemplo anterior:





Interpretação: 75% dos dados estão abaixo de 60,5.

Estatística Descritiva – Medidas Resumo

- **Regra geral para a determinação de quantis:**

  a) Se a posição (i) do quantil for um número inteiro, o valor do quantil será a média dos dados que ocupam as posições *i* e *i+1.*

  b) Se a posição (i) do quantil não for um número inteiro, devemos arredondar para cima e usar o valor do dado da posição *i*.

- Entre os vários quantis existentes, aqueles que **dividem os dados em 4 partes iguais são chamados de quartis** e muitas vezes são utilizados para a geração de gráficos de distribuição.

  - ✓ Q(0,25): primeiro quartil ou Q_1
  - ✓ Q(0,50): segundo quartil ou Q_2 ou mediana
  - ✓ Q(0,75): terceiro quartil ou Q_3
  - ✓ Q(1,0): quarto quartil ou Q_4

Estatística Descritiva – Medidas Resumo

- **<u>Regra geral para a determinação de quantis:</u>**

  a) Se a posição (i) do quantil for um número inteiro, o valor do quantil será a média dos dados que ocupam as posições *i* e *i+1.*

  b) Se a posição (i) do quantil não for um número inteiro, devemos arredondar para cima e usar o valor do dado da posição *i*.

- Entre os vários quantis existentes, aqueles que **dividem os dados em 4 partes iguais são chamados de quartis** e muitas vezes são utilizados para a geração de gráficos de distribuição.

  - ✓ Q(0,25): primeiro quartil ou Q_1
  - ✓ Q(0,50): segundo quartil ou Q_2 ou mediana
  - ✓ Q(0,75): terceiro quartil ou Q_3
  - ✓ Q(1,0): quarto quartil ou Q_4

Estatística Descritiva – Medidas Resumo

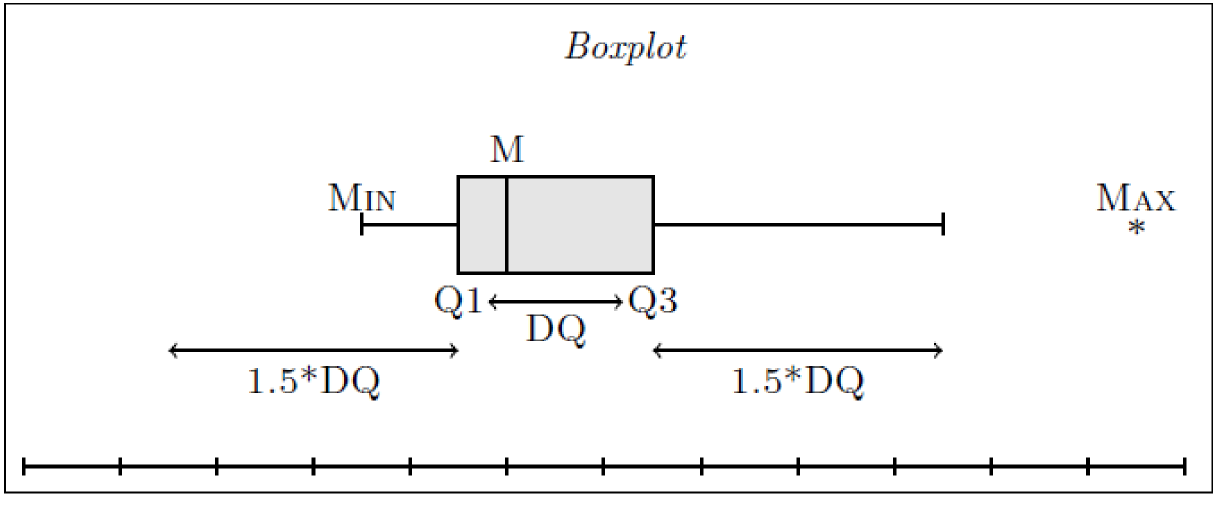
- **Vantagem do uso de quantis:**
  - ✓ Permite definir intervalos de mapas de modo equilibrado (cada classe tem aproximadamente a mesma quantidade de dados)
- **Desvantagem do uso de quantis:**
  - ✓ Pode resultar em separação de dados semelhantes em classes diferentes, ou seja, os quantis podem acabar colocando na mesma classe dados diferentes e separando dados semelhantes (ex: distribuições bi-modais por exemplo).

Estatística Descritiva – Medidas Resumo

➢ **Diagramas de caixa (Boxplots):**

O boxplot é um gráfico baseado nos quantis que serve como uma alternativa para resumir a distribuição dos dados. Ele é um retângulo com bases determinadas pelos quartis Q1 e Q3. Além dos quartis, o boxplot possui a marcação da mediana (M), e de dois “bigodes” (ou “whiskers”) que auxiliam na identificação de pontos atípicos (“outliers”). Esses bigodes são limitados pela seguinte fórmula:

- ✓ **Limite inferior:** max(x_1; Q1 -1,5*dq)
- ✓ **Limite superior:** min(x_n; Q3 + 1,5*dq)



Estatística Descritiva – Medidas Resumo

<u>Exemplo:</u> A tabela abaixo apresenta a participação de mercado das 11 principais modalidades de seguros em % do valor total dos prêmios emitidos (outras modalidades correspondem à 6,9%). Apresente o boxplot desses dados:

RAMO	%
Automóvel	33,6
Saúde	14,0
Incêndio	12,9
Vida	12,2
Riscos Diversos	5,5
Habitação	5,3
Transporte	3,1
Acidentes Pessoais	2,9
Obrigatório Veículos	1,7
Riscos de Engenharia	1,0
Responsabilidade Civil *	0,9

Fonte (Fenaseg, in Exame, Fev / 93)

Estatística Descritiva – Medidas Resumo

<u>Exemplo:</u> Os dados da tabela abaixo correspondem a população (em 10000 habitantes) de 30 municípios brasileiros (IBGE, 1996). Apresente o boxplot desses dados:

Município	População	Município	População
São Paulo (SP)	988,8	Nova Iguaçu (RJ)	83,9
Rio de Janeiro (RJ)	556,9	São Luís (MA)	80,2
Salvador (BA)	224,6	Maceió (AL)	74,7
Belo Horizonte (MG)	210,9	Duque de Caxias (RJ)	72,7
Fortaleza (CE)	201,5	S, Bernardo do Campo (SP)	68,4
Brasília (DF)	187,7	Natal (RN)	66,8
Curitiba (PR)	151,6	Teresina (PI)	66,8
Recife (PE)	135,8	Osasco (SP)	63,7
Porto Alegre (RS)	129,8	Santo André (SP)	62,8
Manaus (AM)	119,4	Campo Grande (MS)	61,9
Belém (PA)	116,0	João Pessoa (PB)	56,2
Goiânia (GO)	102,3	Jaboatão (PE)	54,1
Guarulhos (SP)	101,8	Contagem (MG)	50,3
Campinas (SP)	92,4	S, José dos Campos (SP)	49,7
São Gonçalo (RJ)	84,7	Ribeirão Preto (SP)	46,3