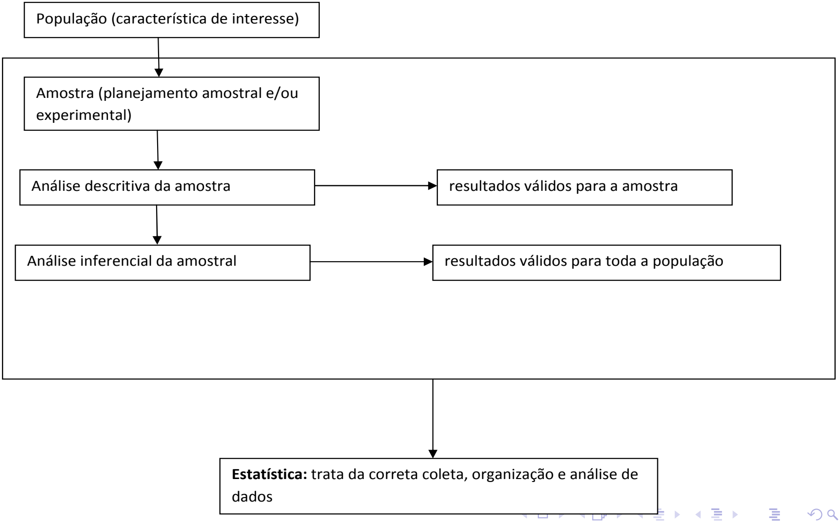
Estatística (Análise de dados)



Tipos de Variáveis

Problema Motivador:

Um pesquisador está interessado em fazer um levantamento sobre aspectos sócio-econômicos dos empregados da seção de orçamentos de uma companhia (vide tabela).

Algumas variáveis como sexo, escolaridade e estado civil, têm como possíveis respostas uma descrição ou qualidade do indivíduo, e portanto são chamadas de **variáveis qualitativas**. Já variáveis como número de filhos e salário têm como possíveis respostas um número, um valor, uma quantidade, e portanto são chamadas de **variáveis quantitativas**.

Variáveis

Qualitativa

- Nominal

  Não existe ordenação nas possíveis categorias de resposta (ex: sexo, estado civil)

- Ordinal

  Existe uma certa ordem nas possíveis categorias de respostas (ex: escolaridade)

Tipos de Variáveis

Quantitativa

- Discreta

  Os possíveis valores formam um conjunto finito ou enumerável de números, em geral são variáveis associadas à contagens (ex: número de filhos)

- Contínua

  Os possíveis valores pertencem à um intervalo, aberto ou fechado, dos números reais (ex: peso de um indivíduo)

Distribuição de Frequências (DF)

- Objeto de estudo: variável (ex: peso)
- Elementos para construir uma DF: realizações (valores observados) da variável
- Objetivo conhecer a distribuição dessa variável (aleatória)

Distribuição de Frequências

- Exemplo: Grau de escolaridade (variável qualitativa ordinal)

  $\#$ total de empregados $= 36$

  $\#$ empregados com Ensino Fundamental $= 12$

  $\#$ empregados com Ensino Médio $= 18$

  $\#$ empregados com Ensino Superior $= 6$

Distribuição de Frequências

Grau de Instrução	Frequência (n_i)	Proporção (f_i)	% ($100 \times f_i$)
Ensino Fundamental	12	0.3333	33.33
Ensino Médio	18	0.5000	50.00
Ensino Superior	6	0.1667	16.67
Total	36	1.0000	100.00

$f_i = \frac{n_i}{36}$

Distribuição de Frequências

- Exemplo: Salário (variável quantitativa contínua)

  Agrupar os dados por faixas de valores

  # total de empregados = 36

  # empregados com salário na faixa 4.00-8.00 = 10

  # empregados com salário na faixa 8.00-12.00 = 12

  # empregados com salário na faixa 12.00-16.00 = 8

  # empregados com salário na faixa 16.00-20.00 = 5

  # empregados com salário na faixa 20.00-24.00 = 1

Distribuição de Frequências

Faixa salarial	Frequência (n_i)	Proporção (f_i)	% ($100 \times f_i$)
4.00-8.00	10	0.2778	27.78
8.00-12.00	12	0.3333	33.33
12.00-16.00	8	0.2222	22.22
16.00-20.00	5	0.1389	13.89
20.00-24.00	1	0.0278	2.78
Total	36	1.0000	100.00

$$f_i = \frac{n_i}{36}$$

Distribuição de Frequências

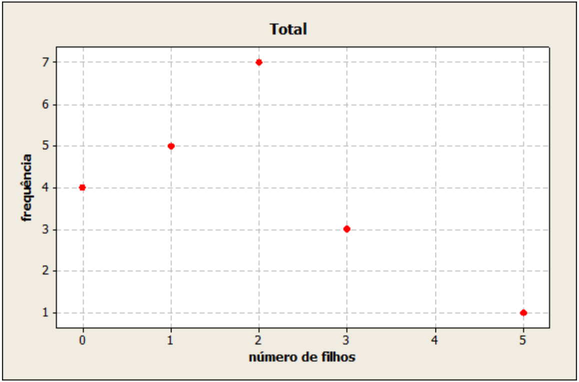
- Escolha dos intervalos: arbitrária ou seguindo algum indicador
  - um número pequeno de classes $\rightarrow$ perda de informação
  - um número grande de classes $\rightarrow$ perda da visão geral dos dados como um conjunto
  - sugestão: 5 a 15 classes com a mesma amplitude
  - $\sqrt(n)$
  - Regra de Sturges: $\ln(n)$

Representação Gráfica das Variáveis Quantitativas

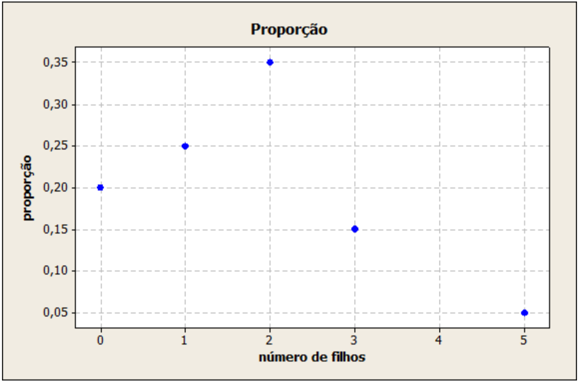
- Objetivo: estudar a distribuição de frequências de uma variável
- Exemplo: número de filhos dos empregados casados

Número de filhos (x_i)	Frequência (n_i)	Proporção (f_i)	% ($100 \times f_i$)
0	4	0.20	20
1	5	0.25	25
2	7	0.35	35
3	3	0.15	15
5	1	0.05	5
Total	20	1.0000	100.00

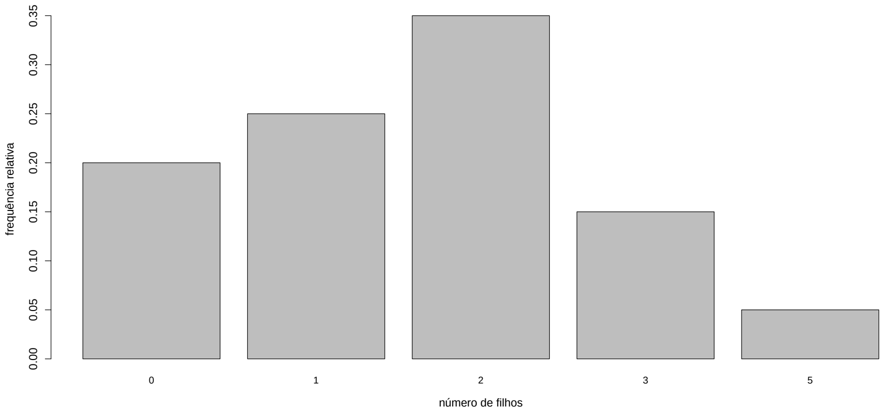
Representação Gráfica de Variáveis Quantitativas



Representação Gráfica de Variáveis Quantitativas



Representação Gráfica de Variáveis Quantitativas

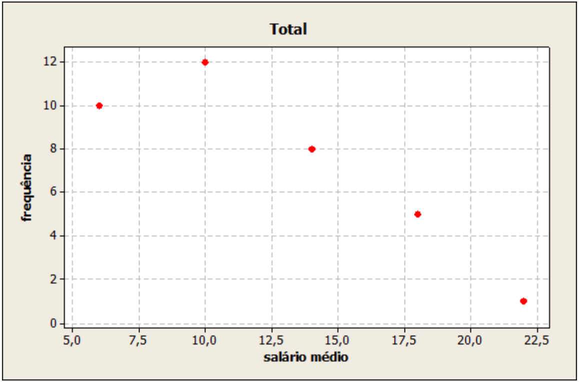


Representação Gráfica de Variáveis Contínuas

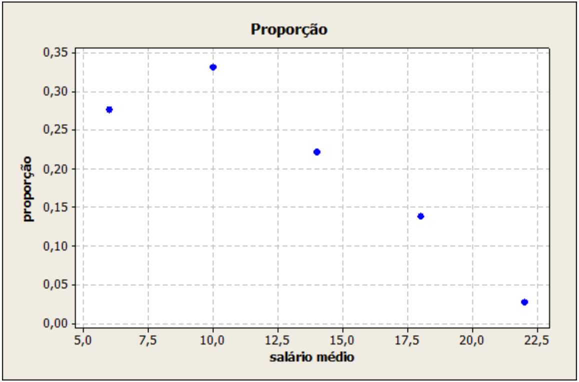
- Dados de salário: são utilizados os pontos médios das faixas salariais

Salário médio	Frequência (n_i)	Proporção (f_i)	% ($100 \times f_i$)
6.00	10	0.2778	27.78
10.00	12	0.3333	33.33
14.00	8	0.2222	22.22
18.00	5	0.1389	13.89
22.00	1	0.0278	2.78
Total	36	1.0000	100.00

Representação Gráfica de Variáveis Contínuas

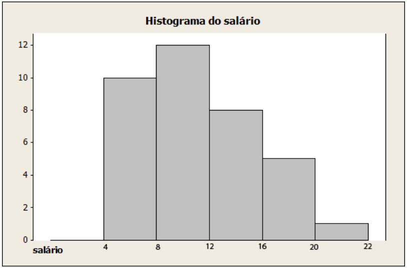


Representação Gráfica de Variáveis Contínuas



Representação Gráfica de Variáveis Contínuas

- Melhor representação dos dados: Histograma
- Associa a frequência aos intervalos de valores, e não mais ao ponto médio



Representação Gráfica de Variáveis Contínuas

- Ramo e Folhas
- Objetivo: obter informação da distribuição dos dados
- Característica: Não perde informação sobre os dados
- Cada informação é dividida em duas partes: a primeira (ramo) é colocada à esquerda da linha vertical, e a segunda (folhas) à direita

Representação Gráfica de Variáveis Contínuas

4	00	56		
5	25	73		
6	26	66	86	
7	39	44	59	
8	12	46	74	95
9	13	35	77	80
10	53	76		
11	06	59		
12	00	79		
13	23	60	85	
14	69	71		
15	99			
16	22	61		
17	26			
18	75			
19	40			
20				
21				
22				
23	30			

Medidas de Posição

- Propósito: resumir os dados, através de valores que representem o conjunto de dados em relação à alguma característica (posição, dispersão)
- Medidas de posição central
    - Média aritmética (Me)
    - Mediana (Md)
    - Moda (Mo)

Medidas de Posição

Moda

- Resultado mais frequente, obtido em um conjunto de dados observados
- No exemplo do número de filhos, $Mo = 2$
- É interessante notar que um conjunto de dados pode apresentar mais de uma moda, sendo então bimodal, trimodal, etc.

Medidas de Posição

Mediana

- Resultado que ocupa a posição central em um conjunto de dados ordenados de forma crescente
- Número ímpar de observações: utiliza-se a observação central
  - ex: 3, 4, 7, 8, 8
  - $Md = 7$
- Número par de observações: utiliza-se a média aritmética das duas observações centrais
  - ex: 3, 4, 7, 8, 8, 9
  - $Md = \frac{7+8}{2} = 7.5$

Medidas de Posição

Média

- Soma dos valores observados dividida pelo número total de observações
- ex: 3, 4, 7, 8, 8 $\rightarrow Me = \frac{3+4+7+8+8}{5} = \frac{30}{5} = 6$
- No exemplo do número de filhos $Me = 1.65$
- Expressão geral

$$Me\,(X) = \frac{x_1 + ... + x_k}{k} = \frac{1}{k}\sum_{i=1}^{k} x_i$$

$x_1, ..., x_k$ são os valores observados para uma variável de estudo X

Medidas de Posição

- Caso particular:

  n_1 observações são iguais a x_1

  n_2 observações são iguais a x_2

  $\vdots$

  n_k observações são iguais a x_k

  tal que: $n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = n$

$$Me(X) = \frac{n_1x_1 + n_2x_2 + \ldots + n_kx_k}{n} = \frac{1}{n}\sum_{i=1}^{k} n_ix_i = \sum_{i=1}^{k} \frac{n_i}{n}x_i = \sum_{i=1}^{k} f_ix_i$$

Medidas de Posição

- No exemplo do número de filhos

$$\left.\begin{array}{ll} n_1 = 4, & x_1 = 0 \\ n_2 = 5, & x_2 = 1 \\ n_3 = 7, & x_3 = 2 \\ n_4 = 3, & x_4 = 3 \\ n_5 = 1, & x_5 = 5 \end{array}\right\} n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = n$$

então,

$$Me(X) = \frac{4 \times 0 + 5 \times 1 + 7 \times 2 + 3 \times 3 + 1 \times 5}{20} = 1.65$$

Medidas de Posição

Análise 1

- Conjunto de dados $D_1 = \{2, 2.5, 3, 4.3, 2.9\}$
- Ordenando de forma crescente $D_1^{'} = \{2, 2.5, 2.9, 3, 4.3\}$
- $Md = 2.9$
- $Me = \frac{2+2.5+2.9+3+4.3}{5} = 2.94$

Medidas de Posição

Análise 2

- Conjunto de dados $D_2 = \{2, 7, 3, 4.3, 2.9\}$
- Ordenando de forma crescente $D_2^{'} = \{2, 2.8, 3, 4.3, 7\}$
- $Md = 3$
- $Me = \frac{2+2.8+3+4.3+7}{5} = 3.84$

Medidas de Posição

Observação

Na primeira e segunda análise a mediana tem valores próximos (2.9 e 3), no entanto, a média tem uma diferença de quase 1 unidade (2.94 e 3,84). Com isso em vista, podemos definir a propriedade de robustez da mediana.

A mediana é uma medida mais robusta que a média, quando submetida a mudanças nos valores observados, ou a incorporação de mais observações no conjunto de dados original.

Medidas de Posição

Análise 3

- Conjunto de dados $D_3 = \{2, 2.5, 3, 4.3, 2.9, 7\}$
- Ordenando de forma crescente $D_3^{'} = \{2, 2.5, 2.9, 3, 4.3, 7\}$
- $Md = \frac{2.9+3}{2}$
- $Me = \frac{2+2.5+2.9+3+4.3+7}{6} = 3.62$

Medidas de Posição

Comparação entre as análises dos conjuntos de dados

Dados	Md	Me
D_1	2.90	2.94
D_2	3.00	3.84
D_3	2.95	3.62

Medidas de Dispersão

- Propósito: obter uma medida que represente a variabilidade, uma vez que conjuntos de dados diferentes podem apresentar uma mesma medida de posição.
- Por exemplo, $A = \{3, 4, 5, 6, 7\}$ e $B = \{5, 5, 5, 5, 5\}$ têm a mesma média: $Me = 5$

Medidas de Dispersão

- Desvio: afastamento de uma observação de uma determinada medida de posição
  - ex: $A = \{3, 4, 5, 6, 7\}$

    $Me = \bar{x} = 5$

    $Desvios = \{3-5, 4-5, 5-5, 6-5, 7-5\} = \{-2, -1, 0, 1, 2\}$
  - ex: $B = \{5, 5, 5, 5, 5\}$

    $Me = \bar{x} = 5$

    $Desvios = \{5-5, 5-5, 5-5, 5-5, 5-5\} = \{0, 0, 0, 0, 0\}$

Medidas de Dispersão

- Medidas "globais" de desvio na amostra de dados:
  - $\sum_{i=1}^{5} |x_i - \bar{x}|$
  - $\sum_{i=1}^{5} (x_i - \bar{x})^2$
- Ambas as medidas evitam que desvios iguais em módulo, mas com sinais opostos, se anulem
- Desvio Médio

  $DM(X) = \sum_{i=1}^{n} \frac{|x_i - \bar{x}|}{n}$
- Variância

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

Medidas de Dispersão

- ex: $A = \{3, 4, 5, 6, 7\}$

  $DM(A) = \frac{|-2|+|-1|+|0|+|1|+|2|}{5} = \frac{6}{5} = 1.2$

  $Var(A) = \frac{(-2)^2+(-1)^2+0^2+1^2+2^2}{5} = \frac{10}{5} = 2$

- ex: $B = \{5, 5, 5, 5, 5\}$

  $DM(A) = \frac{|0|+|0|+|0|+|0|+|0|}{5} = \frac{0}{5} = 0$

  $Var(A) = \frac{0^2+0^2+0^2+0^2+0^2}{5} = \frac{0}{5} = 0$

Medidas de Dispersão

- Desvio Padrão

  $DP(X) = \sqrt{Var(X)}$

  - ex: $DP(A) = \sqrt{2} = 1.41$
  - ex: $DP(B) = \sqrt{0} = 0$

Medidas Complementares para Análise de Dados

- Extremos

  O menor e o maior valor do conjunto de dados

- Quartis (Q) ou Juntas (J)
  - 1^o Quartil: deixa um quarto dos valores abaixo, e três quartos acima dele
  - 2^o Quartil $=$ Mediana: deixa metade dos valores abaixo, e metade acima dele
  - 3^o Quartil: deixa três quartos dos valores abaixo, e um quarto acima dele

Medidas Complementares para Análise de Dados

- Exemplo: Variável Salário
  - $Md = \frac{9.8+10.53}{2} = 10.17$
  - $Q_1 = J_1 = \frac{7.44+7.59}{2} = 7.52$
  - $Q_3 = J_3 = \frac{13.85+14.69}{2} = 14.27$
  - $E_i = 4.00$ (menor valor)
  - $E_s = 23.30$ (maior valor)

Medidas Complementares para Análise de Dados

Esquema dos Cinco Números

	36	
Md	10.17	
J	7.52	14.27
E	4.00	23.30

Cada uma das componentes do esquema dos cinco números é uma medida robusta de dados, e é também uma estatística de ordem.

Medidas Complementares para Análise de Dados

- Intervalo Interquartílico: A medida de dispersão “intervalo interquartílico” pode ser considerada uma medida robusta de dispersão.

  $d_J = J_3 - J_1 = Q_3 - Q_1$

  - No exemplo do salário: $d_J = 14.27 - 7.52 = 6.75$
- Dispersão Inferior: $J_2 - E_i$
- Dispersão Superior: $E_s - J_2$

Análise de dados

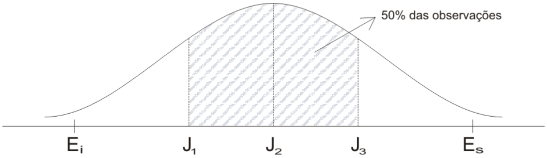
Se a distribuição dos dados que estudamos é simétrica, esperamos que:

- a distribuição inferior seja aproximadamente igual à superior

  $J_2 - E_i \approx E_s - J_2$

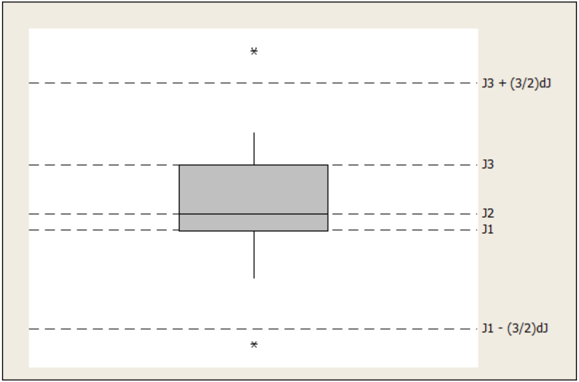
- $J_2 - J_1 \approx J_3 - J_2$

- $J_1 - E_i \approx E_s - J_3$



Análise de dados

Box Plot



Análise de dados

- Os valores que estão muito distantes de J_1 e J_3 são chamados *outliers* (observações discrepantes)
  - observações menores que $J_1 - \frac{3}{2}d_J$
  - observações maiores que $J_3 + \frac{3}{2}d_J$
- A partir do retângulo, para cima e para baixo, seguem linhas até o ponto de observação mais remoto, que não seja outlier

Análise de dados

- O desenho dá uma idéia de:
  - posição: J_1, J_2, J_3
  - dispersão: d_J
  - assimetria: $J_3 - J_2$; $J_2 - J_1$
  - caudas: comprimento das linhas que seguem desde o retângulo
  - dados discrepantes

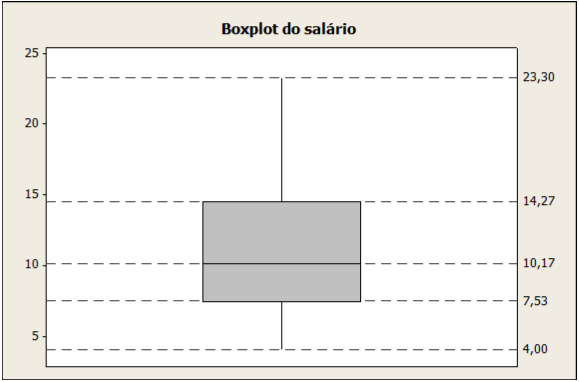
Análise de dados

- Exemplo

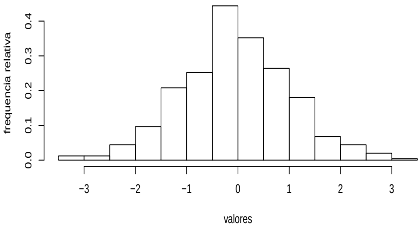
$$J_1 = 7.52 \qquad E_i = 4.00$$
$$J_2 = 10.17 \qquad E_s = 23.30$$
$$J_3 = 14.27 \qquad d_J = 6.75$$

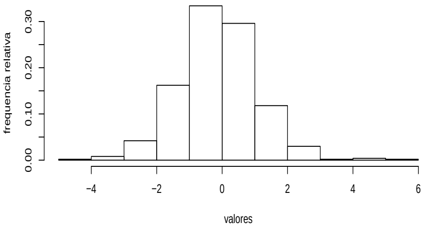
$$J_2 - J_1 = 2.65$$
$$J_3 - J_2 = 4.1$$
$$J_1 - \frac{3}{2}d_J = -2.605$$
$$J_3 + \frac{3}{2}d_J = 24.395$$

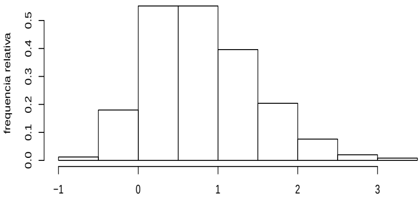
Análide de dados

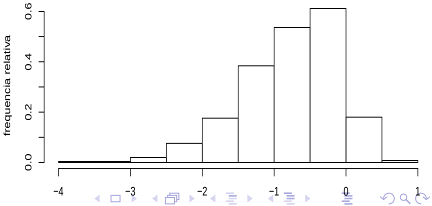


Análise de dados









Análise de dados