Especialização em *Data Science* e Estatística Aplicada

Módulo I - Estatística descritiva para *Data Science*

Profa. Dra. Amanda Buosi Gazon Milani

Goiânia, 2024

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
FEN FACULDADE DE ENFERMAGEM
UFG UNIVERSIDADE FEDERAL DE GOIÁS

Conteúdo Programático

- Introdução à estatística: população, amostra, natureza dos dados, tipos de variáveis;
- Descrição de dados em tabelas: frequências absoluta, relativa e acumulada;
- Visualização de dados: gráfico em barras, colunas, setores, dispersão, histograma e polígonos de frequências;
- Medidas de tendência central: média, mediana e moda;
- Medidas de dispersão: amplitude, variância, desvio padrão e coeficiente de variação;
- Medidas separatrizes: quartis e percentis;
- Outliers e box-plot;
- Análise bidimensional: tabelas e gráficos.

Conteúdo - Aula 2

1. Visualização de dados
   - Tipos de Gráficos
   - Construção de gráficos
   - Conjunto de dados de Exemplo
   - Gráfico em barras
   - Gráfico em colunas
   - Gráfico em setores
   - Gráfico de dispersão
   - Histograma
   - Polígonos de frequências

2. 0.6 - Aula Prática: Visualização de dados através de gráficos

Visualização de dados: Gráficos

As representações gráficas de tabelas de distribuições de frequências permitem que se tenha uma rápida e concisa visualização da distribuição da variável.

Gráficos para variáveis qualitativas

- Gráfico em barras;
- Gráfico em colunas;
- Gráfico em setores (pizza).

Gráficos para variáveis quantitativas

- Gráfico em colunas;
- Gráfico de dispersão;
- Histograma;
- Gráfico de ogiva (frequência acumulada);
- Polígono de frequências;
- Box-plot.

Construção de gráficos

Alguns pontos que devem ser respeitados na construção de gráficos:

1. Devem ser claros e simples, atraindo a atenção e inspirando confiança;
2. Devem ser de tamanho adequado à sua publicação em revistas, periódicos, cartazes, livros, etc;
3. Devem sempre ter um título completo, o qual deve ser colocado na parte superior do gráfico;
4. Devem ser construídos numa escala que não desfigure os fatos ou as relações que se deseja destacar;
5. Seus eixos devem sempre ser especificados (dar nome) e graduados (criar escala);
6. Quando os dados não são próprios, deve-se citar a fonte, a qual deve ser colocada na parte inferior do gráfico.

Tabela 1: Informações sobre estado civil, grau de instrução, número de filhos, salários (expresso como fração do salário mínimo), idade (medida em anos e meses) e procedência de 36 empregados da seção de orçamentos da Companhia MB.

Nº	Estado Civil	Grau de instrução	Nº de filhos	Nº salários	Idade	Região de procedência
1	solteiro	fundamental	-	4,00	26	interior
2	casado	fundamental	1	4,56	32	capital
3	casado	fundamental	2	5,25	36	capital
4	solteiro	médio	-	5,73	20	outra
5	solteiro	fundamental	-	6,26	40	outra
6	casado	fundamental	0	6,66	28	interior
7	solteiro	fundamental	-	6,86	41	interior
8	solteiro	fundamental	-	7,39	43	capital
9	casado	médio	1	7,59	34	capital
10	solteiro	médio	-	7,44	23	outra
11	casado	médio	2	8,12	33	interior
12	solteiro	fundamental	-	8,46	27	capital
13	solteiro	médio	-	8,74	37	outra
14	casado	fundamental	3	8,95	44	outra
15	casado	médio	0	9,13	30	interior
16	solteiro	médio	-	9,35	38	outra
17	casado	médio	1	9,77	31	capital

Fonte: Bussab, W.O., Morettin, P.A. Estatística Básica, 2010

Tabela 2: (Continuação da tabela anterior)

Nº	Estado Civil	Grau de instrução	Nº de filhos	Nº salários	Idade	Região de procedência
18	casado	fundamental	2	9,80	39	outra
19	solteiro	superior	-	10,53	25	interior
20	solteiro	médio	-	10,76	37	interior
21	casado	médio	1	11,06	30	outra
22	solteiro	médio	-	11,59	34	capital
23	solteiro	fundamental	-	12,00	41	outra
24	casado	superior	0	12,79	26	outra
25	casado	médio	2	13,23	32	interior
26	casado	médio	2	13,60	35	outra
27	solteiro	fundamental	-	13,85	46	outra
28	casado	médio	0	14,69	29	interior
29	casado	médio	5	14,71	40	interior
30	casado	médio	2	15,99	35	capital
31	solteiro	superior	-	16,22	31	outra
32	casado	médio	1	16,61	36	interior
33	casado	superior	3	17,26	43	capital
34	solteiro	superior	-	18,75	33	capital
35	casado	médio	2	19,40	48	capital
36	casado	superior	3	23,30	42	interior

Fonte: Bussab, W.O., Morettin, P.A. Estatística Básica, 2010

Gráficos em barras - Variável qualitativa

- É um gráfico formado por retângulos horizontais de larguras iguais, onde o comprimento de cada retângulo representa a intensidade de um atributo.
- É recomendável para variáveis cujas categorias tenham nomes extensos.

Exemplo: Se por exemplo considerarmos a variável grau de instrução, cuja tabela de frequências, conforme já vimos, é dada a seguir.

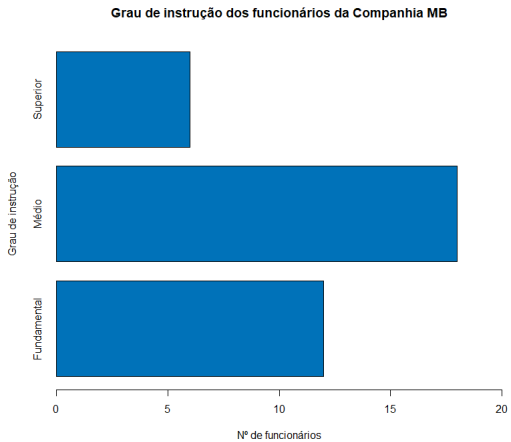
Tabela 3: Resumo das frequências da variável "Grau de Instrução"

Grau de Instrução	Frequência	Proporção	Porcentagem
Fundamental	12	0,333	33,33%
Médio	18	0,500	50,00%
Superior	6	0,167	16,67%
Total	36	1,000	100,00%

Fonte: Elaborado pela autora

Gráficos em barras - Variável qualitativa

Figura 1: Gráfico em barras para a variável “Grau de Instrução”



Fonte: Elaborado pela autora

Nota

O gráfico em barras ao lado foi construído utilizando a frequência absoluta para o eixo x, entretanto, uma das frequências relativas poderiam ter sido adotadas, proporção ou porcentagem.

Gráficos em colunas - Variáveis qualitativas e quantitativas

- É um gráfico formado por retângulos verticais de larguras iguais, onde as alturas de cada retângulo representam as intensidades de cada atributo.
- Se utilizado para variáveis qualitativas, é recomendável para variáveis cujas categorias tenham nomes breves.

Exemplo 1: Se por exemplo considerarmos a variável estado civil, cuja tabela de frequências, conforme já vimos, é dada a seguir.

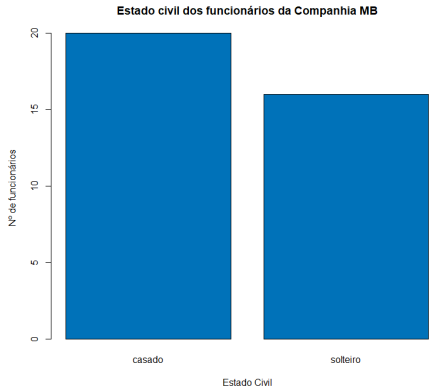
Tabela 4: Resumo das frequências da variável "Estado Civil"

Estado Civil	Frequência	Proporção	Porcentagem
Casado	20	0,556	55,56%
Solteiro	16	0,444	44,44%
Total	36	1,000	100,00%

Fonte: Elaborado pela autora

Gráficos em colunas - Variável qualitativa

Figura 2: Gráfico em colunas para a variável “Estado civil”



Fonte: Elaborado pela autora

Nota

O gráfico em colunas ao lado foi construído utilizando a frequência absoluta para o eixo y, entretanto, uma das frequências relativas poderiam ter sido adotadas, proporção ou porcentagem.

Gráficos em colunas - Variável quantitativa

Exemplo 2: Se por exemplo considerarmos a variável número de filhos, cuja tabela de frequências, conforme já vimos, é dada a seguir.

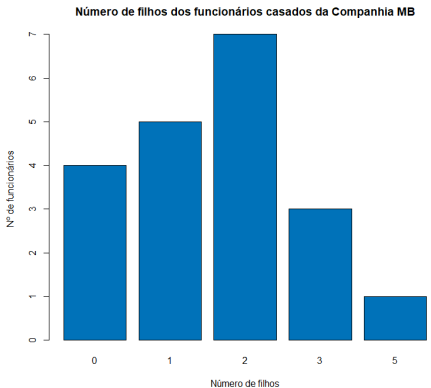
Tabela 5: Resumo das frequências da variável "Número de filhos dos empregados casados"

Nº de filhos	Frequência	Proporção	Porcentagem
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	1,00	100%

Fonte: Elaborado pela autora

Gráficos em colunas - Variável quantitativa

Figura 3: Gráfico em colunas para a variável "Número de filhos dos funcionários casados"



Fonte: Elaborado pela autora

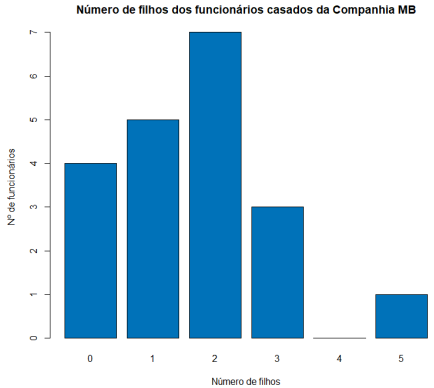
Atenção!

Note que o gráfico de barras ao lado não preserva a característica quantitativa (eixo x) da variável, uma vez que a ausência do valor 4 foi ignorada na execução do gráfico em colunas.

Veremos, a seguir uma versão correta do gráfico de colunas e, posteriormente, um gráfico de colunas mais adequado para variáveis quantitativas: o histograma.

Gráficos em colunas - Variável quantitativa

Figura 4: Gráfico em colunas para a variável “Nº de filhos dos funcionários casados”



Fonte: Elaborado pela autora

Gráficos em setores - Variável qualitativa

- A variável é projetada em um círculo dividido em setores com áreas proporcionais às frequências.
- O total é representado pelo círculo, que fica dividido em tantos setores quantas são as partes.
- Os setores são tais que suas áreas são respectivamente proporcionais aos dados da série.
- Obtemos cada setor por meio de uma regra de três simples, lembrando que o total da série corresponde a 360^o.

Exemplo: Se por exemplo considerarmos a variável região de procedência, cuja tabela de frequências é dada a seguir.

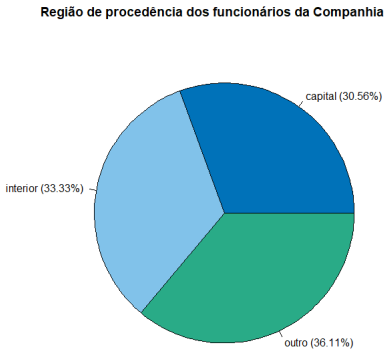
Tabela 6: Resumo das frequências da variável "Região de procedência"

Região de procedência	Frequência	Proporção	Porcentagem
Capital	11	0,3056	30,56%
Interior	12	0,3333	33,33%
Outra	13	0,3611	36,11%
Total	36	1,0000	100,00%

Fonte: Bussab e Morettin

Gráficos em setores - Variável qualitativa

Figura 5: Gráfico de setores para a variável “Região de procedência”



Fonte: Elaborado pela autora

Gráficos de dispersão unidimensional - Variável quantitativa

- Cada valor é representado por pontos ou segmentos verticais, localizados acima do eixo x.

Exemplo: Retomando o exemplo da variável número de filhos, cuja tabela de frequências, conforme já vimos, é dada a seguir.

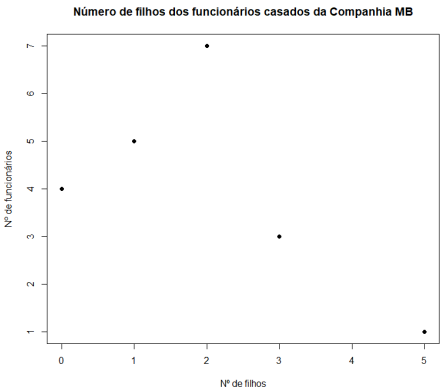
Tabela 7: Resumo das frequências da variável "Número de filhos dos empregados casados"

Nº de filhos	Frequência	Proporção	Porcentagem
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	1,00	100%

Fonte: Elaborado pela autora

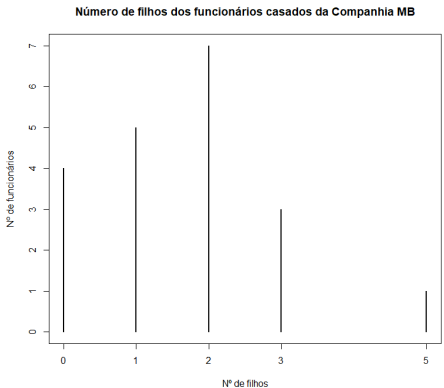
Gráficos de dispersão unidimensional - Variável quantitativa

Figura 6: Gráfico de dispersão unidimensional para a variável “Número de filhos” - Opção 1



Fonte: Elaborado pela autora

Figura 7: Gráfico de dispersão unidimensional para a variável “Número de filhos” - Opção 2



Fonte: Elaborado pela autora

Histogramas - Variável quantitativa com intervalo de classe

- É um gráfico de barras contíguas, com as bases proporcionais aos intervalos de classe e a área de cada retângulo proporcional à respectiva frequência.
- A área total do histograma é igual a 1.
- Considere:
  - fr_i : frequência relativa ou proporção do i-ésimo intervalo de classe;
  - Δ_i : amplitude da i-ésimo intervalo de classe.
- Para que a área do retângulo respectivo seja proporcional a fr_i, a sua altura deve ser proporcional a

$$\frac{fr_i}{\Delta_i}$$

que é chamada de **densidade de frequência** do i-ésimo intervalo de classe, para $i = 1, \ldots, k$, em que k é o número de classes da variável.

Histogramas - Variável quantitativa com intervalo de classe

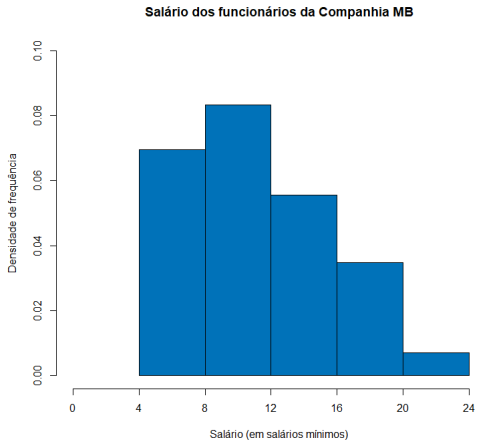
Tabela 8: Resumo das frequências da variável “Salário”

Salário	Frequência	Proporção	Amplitude da classe	Densidade de frequência
$4 \vdash 8$	10	0,2778	4	0,070
$8 \vdash 12$	12	0,3333	4	0,083
$12 \vdash 16$	8	0,2222	4	0,055
$16 \vdash 20$	5	0,1389	4	0,005
$20 \vdash 24$	1	0,0278	4	0,008
Total	36	1,0000		

Fonte: Elaborado pela autora

Histogramas - Variável quantitativa com intervalo de classe

Figura 8: Histograma para a variável “Salário”



Fonte: Elaborado pela autora

Nota

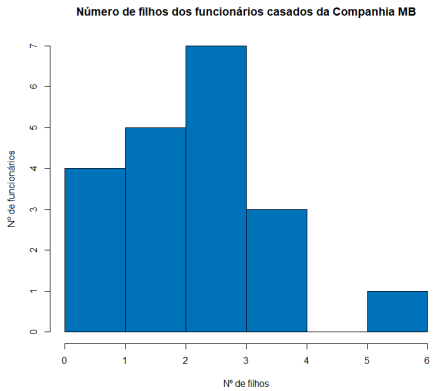
Embora o histograma seja definido como um gráfico com área 1 e, portanto, deva apresentar a densidade de frequências no eixo y, é comum vermos a utilização do histograma com as frequências (absoluta ou relativa) em vez das densidades de frequências no eixo y.

Histogramas - Variável quantitativa discreta

Podemos fazer histograma para variáveis discretas também!

No caso das variáveis discretas, sem intervalos de classes, como é o exemplo da variável número de filhos, o histograma apresenta no eixo x do gráfico os seguintes intervalos: $[0, 1)$, $[1, 2)$, $[2, 3)$, $[3, 4)$, $[4, 5)$ e $[5, 6)$, correspondendo aos valores $0, 1, 2, 3, 4, 5$ discretos da variável.

Figura 9: Histograma para a variável “Número de filhos”



Fonte: Elaborado pela autora

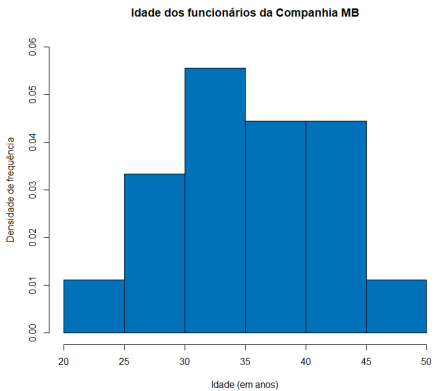
Histogramas - Variável quantitativa discreta

Quando as variáveis discretas apresentam muitos valores distintos, também podemos fazer intervalos de classes e construir histogramas para representá-las graficamente.

Tabela 9: Resumo de frequências da variável "Idade"

Idade	Frequência	Proporção
20 ⊢ 25	2	0,06
25 ⊢ 30	6	0,16
30 ⊢ 35	10	0,28
35 ⊢ 40	8	0,22
40 ⊢ 45	8	0,22
45 ⊢ 50	2	0,06
Total	36	1,00

Figura 10: Histograma para a variável "Número de filhos"



Fonte: Elaborado pela autora

Gráfico da frequência acumulada ou Ogiva

(ou Polígono de frequências acumuladas) - Variável quantitativa

- É um gráfico que une os pontos cujas abscissas (eixo x) são os limites superiores das classes e as ordenadas (eixo y) suas respectivas frequências acumuladas.
- O ponto inicial do gráfico é o limite inferior da primeira classe, com frequência acumulada zero, pois não existe valor inferior a ele.

Exemplo:

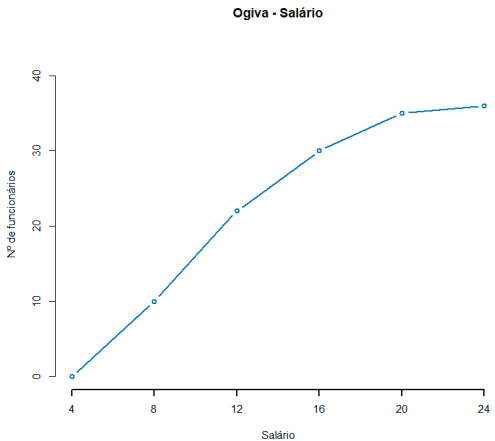
Tabela 10: Distribuição de frequências para variável "Salário"

Salário	Frequência	Frequência acumulada
4 ⊢ 8	10	10
8 ⊢ 12	12	22
12 ⊢ 16	8	30
16 ⊢ 20	5	35
20 ⊢ 24	1	36
Total	36	

Fonte: Elaborado pela autora

Gráfico da frequência acumulada ou Ogiva

Figura 11: Gráfico da frequência acumulada (ogiva) para a variável “Salário”



Fonte: Elaborado pela autora

Gráfico Polígono de frequências - Variável quantitativa

- É um gráfico que une por linhas retas os pontos médios das bases superiores dos retângulos do histograma.

Exemplo: Considerando novamente a variável salário, cuja tabela de frequências, conforme já vimos, é dada a seguir.

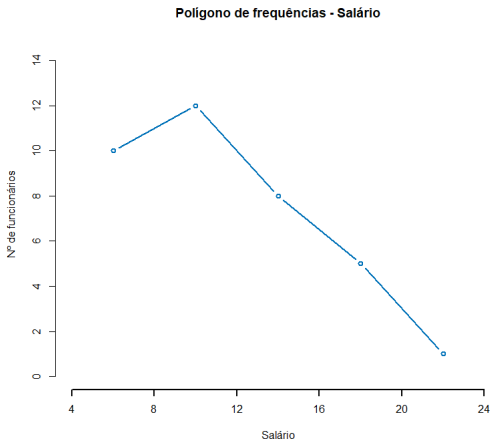
Tabela 11: Distribuição de frequências para variável "Salário"

Salário	Frequência	Freq. acumulada	Ponto médio da classe
4 ⊢ 8	10	10	6
8 ⊢ 12	12	22	10
12 ⊢ 16	8	30	14
16 ⊢ 20	5	35	18
20 ⊢ 24	1	36	22
Total	36		

Fonte: Elaborado pela autora

Gráfico Polígono de frequências - Variável quantitativa

Figura 12: Gráfico polígono de frequências para a variável “Salário”



Fonte: Elaborado pela autora

0.6 - Aula Prática: Visualização de dados através de gráficos

Agora, iremos aplicar o conteúdo teórico adquirido na aula prática, utilizando a linguagem de programação R, aplicando as técnica à base de nascidos vivos 2024. Vamos nos direcionar para o arquivo da `0.6_AulaPrática`.

Estatística descritiva para *Data Science*

0.6 - Aula Prática: Visualização de dados através de gráficos

Profa. Dra. Amanda Buosi Gazon Milani

2024-07-20

Especialização em *Data Science* e Estatística Aplicada

Módulo I - Estatística descritiva para *Data Science*

Profa. Dra. Amanda Buosi Gazon Milani

amandamilani@ufg.br

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

FEN FACULDADE DE ENFERMAGEM

UFG UNIVERSIDADE FEDERAL DE GOIÁS