Análise exploratória

Gráficos e tabelas para variáveis qualitativas

Prof. Me. Lineu Alberto Cavazani de Freitas

Departamento de Estatística
Laboratório de Estatística e Geoinformação

Análise exploratória

- Parte primordial de qualquer análise estatística é chamada **análise descritiva** ou **exploratória**.
- Consiste basicamente de **tabelas**, **resumos numéricos** e **análises gráficas** das variáveis disponíveis em um conjunto de dados.
- Trata-se de uma etapa de extrema importância e deve preceder qualquer análise mais sofisticada.
- As técnicas de análise exploratória visam **resumir** e **apresentar** as informações de um conjunto de dados brutos.

Análise exploratória

- A análise exploratória de dados é uma área relativamente nova.
- Nasceu do clássico livro **Exploratory Data Analysis** de **John Tukey** em 1977.
- Algo curioso é que Tukey tinha uma relação próxima com a Ciência da Computação e definiu os termos **bit** e **software**.

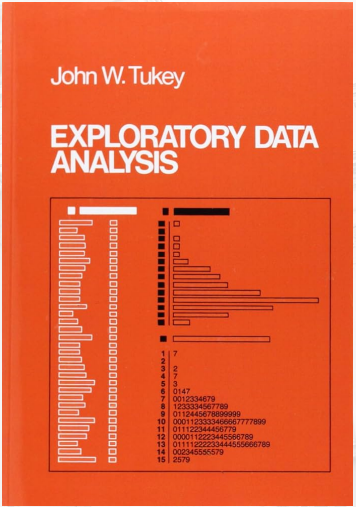

Figura 1. Capa do livro Exploratory Data Analysis de John Tukey.

Análise exploratória

- Como quase tudo em análise de dados, o **avanço computacional** permitiu com que a análise exploratória evoluísse substancialmente.
- Por exemplo: historicamente o processo de criação de um gráfico era reservado a pessoas qualificadas pois a produção de uma visualização era difícil.
- Hoje qualquer pessoa pode inserir dados em um aplicativo e gerar um gráfico.
- Este tipo de facilidade é importante para disseminação e democratização dos métodos, porém abre margem para certas práticas inadequadas.

Análise exploratória

- Tentar compreender um conjunto de dados sem algum método que permita resumir as informações é inviável.
- A análise exploratória é a primeira forma de tentarmos entender o que acontece nos nossos dados.
- Uma das tarefas é a etapa de consistência dos dados, isto é, verificar se os dados coletados são condizentes com a realidade.

Figura 2. Extraído de pixabay.com.

Análise exploratória

- O conjunto de técnicas aplicáveis está diretamente associado ao **tipo das variáveis de interesse** (quantitativas x qualitativas) e suas ramificações.
- Podemos conduzir análises focadas nas variáveis uma a uma (**análises univariadas**).
- Também podemos conduzir análises focadas em avaliar a relação entre as variáveis (**análises multivariadas**).

Figura 3. Extraído de pixabay.com.

Análise exploratória

Podemos fazer uso diversas técnicas, tais como

- Tabelas de frequência absolutas.
- Tabelas de frequência relativas.
- Tabelas de frequência acumuladas.
- Tabelas para múltiplas variáveis.
- Gráficos.
- Medidas de posição central.
- Medidas de posição relativa.
- Medidas de forma.
- Medidas de dispersão.
- Medidas de associação.

Análise exploratória

- Para ilustrar as técnicas de análise exploratória de dados, usaremos o conjunto de dados “milsa”.
- Este conjunto de dados aparece no livro “Estatística Básica” de W. O. Bussab e P. A. Morettin.
- Conjunto de dados hipotético de atributos de 36 funcionários da companhia “Milsa”.

O conjunto possui as seguintes variáveis:

- **Funcionário**: identificadora de funcionário.
- **Estado civil**: casado ou solteiro.
- **Instrução**: 1º grau, 2º grau, superior.
- **Filhos**: número de filhos.
- **Salário**: salário do funcionário.
- **Anos**: idade em anos completos.
- **Meses**: meses além dos anos completos.
- **Região**: capital, interior, outro.

Análise exploratória

Tabela 1. Primeiras linhas do conjunto de dados Milsa.

Funcionário	Estado civil	Instrução	Filhos	Salário	Anos	Meses	Região
1	solteiro	1o Grau	NA	4.00	26	3	interior
2	casado	1o Grau	1	4.56	32	10	capital
3	casado	1o Grau	2	5.25	36	5	capital
4	solteiro	2o Grau	NA	5.73	20	10	outro
5	solteiro	1o Grau	NA	6.26	40	7	outro
6	casado	1o Grau	0	6.66	28	0	interior
7	solteiro	1o Grau	NA	6.86	41	0	interior
8	solteiro	1o Grau	NA	7.39	43	4	capital
9	casado	2o Grau	1	7.59	34	10	capital
10	solteiro	2o Grau	NA	7.44	23	6	outro

Análise descritiva univariada para variáveis qualitativas

Análise descritiva univariada para variáveis qualitativas

- Uma variável qualitativa representa um atributo que pode ser expresso por meio de **rótulos** ou **categorias**.
- Podem ser classificadas em **nominais** (sem ordenação natural entre as categorias) ou **ordinais** (com ordenação natural entre as categorias).
- As categorias também são chamadas de **classes** ou **níveis**.
- Na análise descritiva de uma variável qualitativa estamos interessados em avaliar as **frequências** das classes.

Tipos de frequência

- **Frequência absoluta** (f_a): número de observações no conjunto de dados que pertence a uma determinada classe.
- **Frequência relativa** (f_r): frequência de classe dividida pelo número total de observações no conjunto de dados.
  - Pode ser apresentada em forma de percentual, quando multiplicada por 100.
- **Frequência acumulada** (F_a ou F_r): frequência absoluta ou relativa acumulada conforme disposição das classes.
  - Não faz muito sentido para variáveis qualitativas nominais.

Tabelas de frequência para uma variável qualitativa

- Utlizando apenas os dados brutos é difícil responder questões de interesse.
- Para reduzir os dados originais de forma que fique mais claro o entendimento dos mesmos são utilizadas as **tabelas de frequência**.
- No caso de variáveis qualitativas consiste em listar os possíveis níveis da variável e fazer a contagem de quantas vezes cada nível aparece nos dados brutos.

Figura 4. Extraído de pixabay.com.

Tabelas de frequência para uma variável qualitativa

- Cada **linha** da tabela diz respeito a um **nível** da variável.
- As **colunas** podem apresentar diferentes tipos de **frequência** (absoluta, relativa).
- Alguns cuidados para a apresentação dos resultados dizem respeito ao tipo de variável em questão: nominal ou ordinal.
- Os níveis de variáveis **nominais não apresentam uma ordenação natural**, portanto, na apresentação dos resultados pode ser interessante **ordenar** os níveis **por frequência** ou **por ordem alfabética**.
- Esta estratégia não é recomendada para variáveis **ordinais**, pois estas **apresentam uma ordenação natural** e esta ordenação deve ser preferencialmente mantida na exposição dos resultados.

Tabelas de frequência para uma variável qualitativa nominal

Tabela 2. Tabela de frequências para a região.

Região	Frequência	Freq. Relativa
capital	11	0.31
interior	12	0.33
outro	13	0.36
Total	36	1.00

Tabelas de frequência para uma variável qualitativa nominal

Tabela 3. Tabela de frequências para a região.

Região	Frequência	Freq. Relativa
outro	13	0.36
interior	12	0.33
capital	11	0.31
Total	36	1.00

Tabelas de frequência para uma variável qualitativa nominal

Tabela 4. Tabela de frequências para a região.

Região	Frequência	Percentual
outro	13	36 %
interior	12	33 %
capital	11	31 %
Total	36	100 %

Tabelas de frequência para uma variável qualitativa ordinal

Tabela 5. Tabela de frequências para o grau de instrução.

Instrução	Frequência	Freq. Relativa	Freq. Acumulada	Freq. Rel. Acumulada
10 Grau	12	0.33	12	0.33
20 Grau	18	0.50	30	0.83
Superior	6	0.17	36	1.00
Total	36	1.00	36	1.00

Tabelas de frequência para uma variável qualitativa ordinal

Tabela 6. Tabela de frequências para o grau de instrução.

Instrução	Frequência	Percentual	Freq. Acumulada	Percentual Acumulado
10 Grau	12	33 %	12	33 %
20 Grau	18	50 %	30	83 %
Superior	6	17 %	36	100 %
Total	36	100 %	36	100 %

Gráficos para representação de frequências de uma variável qualitativa

- A representação por meio de tabelas é útil mas nem sempre eficiente.
- Em diversos casos pode ser mais conveniente utilizar um **gráfico**.
- “Uma imagem vale mais que mil palavras”.
- Os cuidados com a ordenação dos níveis de acordo com o tipo da variável se mantém.

Algumas possibilidades são:

- Gráfico de barras verticais.
- Gráfico de barras horizontais.
- Gráfico de barras empilhadas.
- Gráfico de setores.

Gráfico de barras

Gráfico de barras verticais ou horizontais.

- Utiliza os possíveis **níveis** das variáveis **em um eixo**.
- As **frequências ou porcentagens** ficam **no outro eixo**.
- O tamanho da barra correspondente à frequência ou percentual.

Gráfico de barras empilhadas.

- Usa-se **uma única barra**.
- A barra é dividida de acordo com a **contribuição relativa** de cada nível da variável.
- Representa-se a frequência relativa ou o percentual.

Gráfico de barras verticais

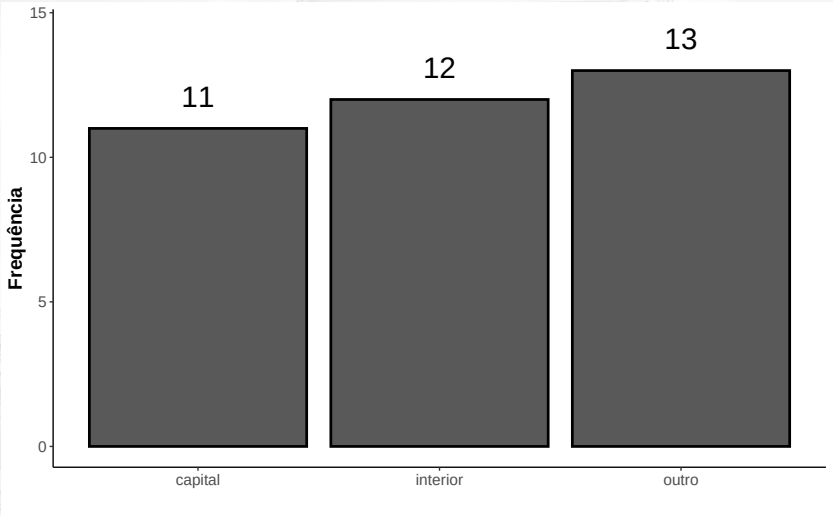

Figura 5. Gráfico de barras verticais para a região.

Gráfico de barras verticais

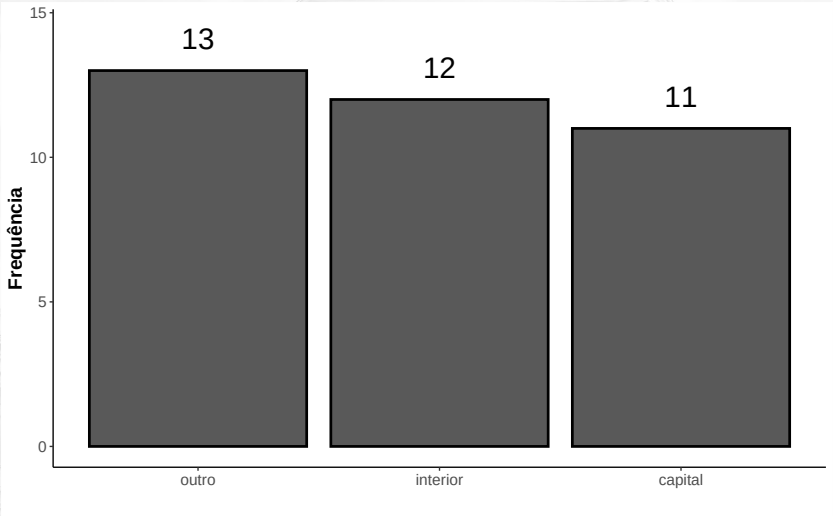

Figura 6. Gráfico de barras verticais para a região.

Gráfico de barras horizontais

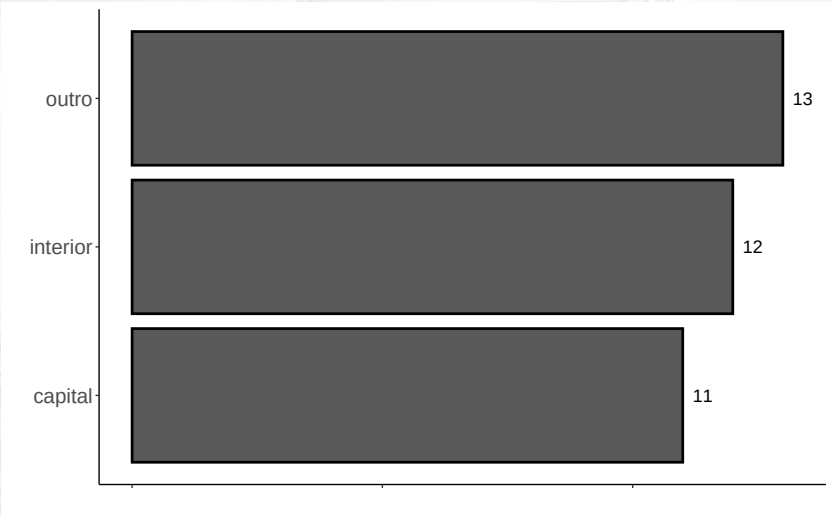

Figura 7. Gráfico de barras horizontais para a região.

Gráfico de barras empilhadas

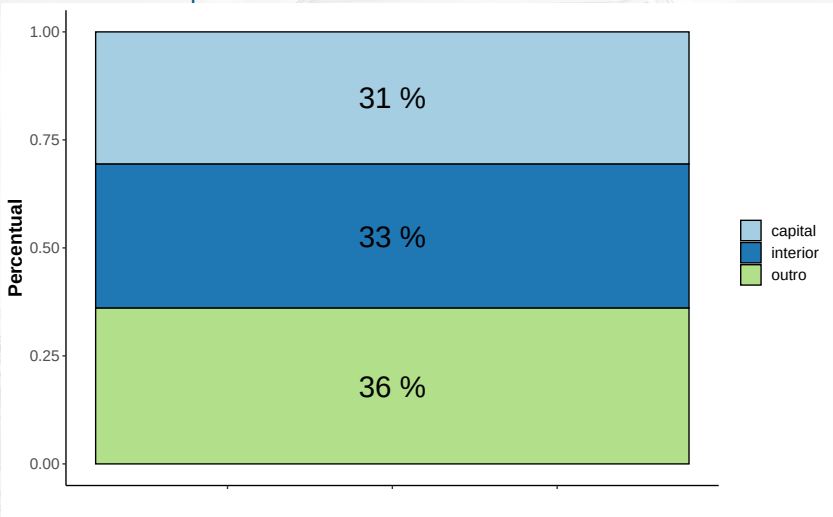

Figura 8. Gráfico de barras empilhadas para a região.

Gráfico de setores

- Consiste em **repartir um círculo** em setores de tamanhos proporcionais às **frequências relativas** ou às **porcentagens** de cada valor.
- Pode ser usados para representar variáveis com **poucos níveis**.
- Apesar de muito usado e preferido em diversas áreas, **deve ser evitado**.
- O cérebro humano tem dificuldade em relacionar **frequências** com **áreas relativas**.
- Para variáveis com muitos níveis, o gráfico tende a ficar **visualmente poluído** e **pouco informativo**.
- Outro problema é que níveis com **frequências iguais a 0 deixam de aparecer no gráfico**, diferente de um gráfico de barras.

Gráfico de setores

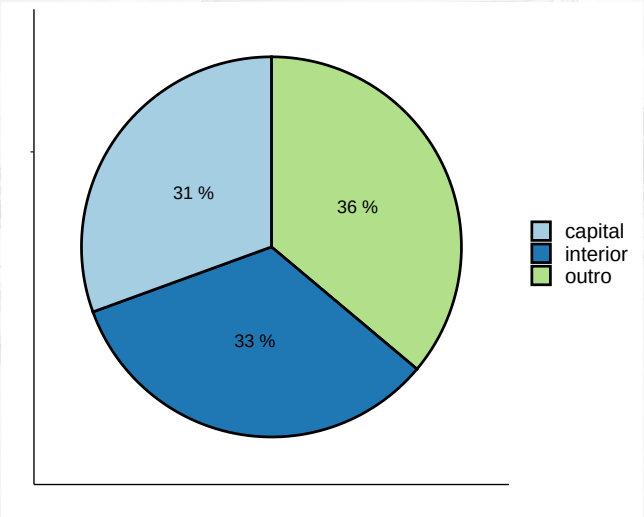

Figura 9. Gráfico de setores para a região

O que foi visto:

- Introdução à análise exploratória.
- Análise exploratória univariada para variáveis qualitativas.

Próximos assuntos:

- Análise exploratória univariada para variáveis quantitativas.