Fundamentos de Análise Exploratória de Dados

Conceitos e Aplicações

Encontro 1

Introdução, dados, gráficos e tabelas para variáveis qualitativas e quantitativas.

Prof. Me. Lineu Alberto Cavazani de Freitas

Introdução

Estatística

Conjunto de métodos e técnicas usados para organizar, descrever, analisar e interpretar dados.

- Compreende:
  1. Planejamento (delineamento) de estudos e coleta de dados (amostragem).
  2. Descrição, análise e interpretação dos dados.

- Permite:
  1. Extrair informações importantes para tomada de decisões.
  2. Avaliar evidências empíricas sob hipóteses de interesse.

Origem da Estatística

- A palavra **Estatística** (**Statistics**) vem do latim **Status**, que significa **Estado**.
- A Estatística tem sua origem em levantamentos de **informações** de interesse para o **Estado**.
- As informações coletadas eram usadas para fins **demográficos**, **militares** e de **taxação de impostos**.
- Existem **registros** de coletas de dados e algumas análises que datam de **3000 anos A.C.** em civilizações como, China, Egito, etc.
- Apenas no século XVII a Estatística passou a ser considerada **disciplina autônoma** e não uma sub-área de outra disciplina.

Origem da Estatística

- A Estatística como área se desenvolveu muito no último século.
- A **teoria das probabilidades** (fundamento matemático da estatística) foi desenvolvida entre os séculos XVII e XIX com base no trabalho de autores como Thomas Bayes, Pierre-Simon Laplace e Carl Gauss.
- Ao contrário da natureza puramente teórica da probabilidade, a **Estatística é uma teoria aplicada relacionada à análise e modelagem de dados**.
- A Estatística moderna tem sua origem no final dos anos 1800 com nomes como Francis Galton e Karl Pearson.
- No começo do século XX, R. A. Fisher liderou o desenvolvimento da Estatística apresentando ideias como design experimental e estimação por máxima verossimilhança.

Símbolo da Estatística

- Representa a importância da matemática na Estatística por meio do **Somatório** e da **Integral**.
- A **engrenagem** representa a indústria, a principal área que fazia uso de métodos estatísticos quando o símbolo foi proposto (1963).

Figura 1. Símbolo da Estatística.

Conceitos fundamentais

Conceitos fundamentais

- **População**: conjunto de seres, itens ou eventos com uma característica comum.
  - TODOS aqueles que possuem a característica de interesse pertencem à população.
- **Amostra**: subconjunto da população.
- **Variáveis**: características observadas em cada elemento.

Em Estatística tentamos entender o que acontece na população com base no que observamos em uma amostra.

População x Amostra

- O objetivo de qualquer estudo é avaliar a **população**.
- Nem sempre é possível coletar dados de toda a população.
- A alternativa é trabalhar com uma **amostra**.
- Caso toda a população seja acessível no estudo, fazemos um estudo censitário (**censo**).

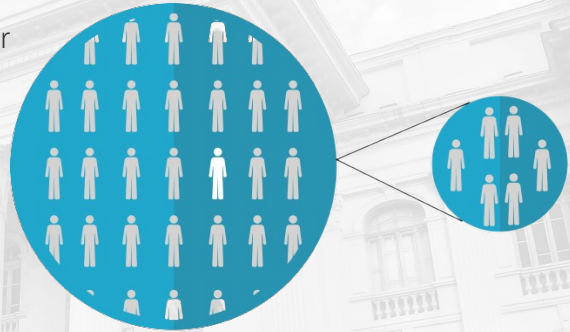


Figura 2. Representação população/amostra. Extraído de pixabay.com.

Exemplos

- Existe interesse em avaliar a opinião dos **alunos da UFPR** a respeito de determinada política.
  - **População**: todos os alunos da UFPR.
  - **Amostra**: parte dos alunos da UFPR.

- Um pesquisador propôs uma nova droga que tem como objetivo reduzir **cólicas menstruais**.
  - **População**: todos os indivíduos que apresentam cólicas menstruais.
  - **Amostra**: parte da população de indivíduos que apresentam cólicas.

Etapas da análise estatística

De forma geral, as etapas para análise de um conjunto de dados são:

1. Definição do problema.
   - Hipóteses, objetivos, população e variáveis de interesse.
2. Planejamento do estudo.
3. Coleta, limpeza e validação de dados.
4. Análise dos dados
   - Análise exploratória.
   - Aplicação de métodos mais sofisticados que permitam generalizar os resultados para a população.
5. Interpretação dos resultados.

Alguns exemplos de aplicações de Estatística

- **Medicina**: eficácia de tratamentos propostos.
- **Indústria**: avaliação de qualidade de itens produzidos.
- **Negócios**: análise do perfil dos indíviduos para concessão de crédito.

Figura 3. Extraído de pixabay.com.

Figura 4. Extraído de pixabay.com.

Figura 5. Extraído de pixabay.com.

Áreas da Estatística

Áreas da Estatística

Em livros de Estatística básica:

1. **Estatística descritiva ou exploratória.**
   - Coleta, organização, tratamento, análise e apresentação de dados.
2. **Probabilidade.**
   - Modelagem de fenômenos aleatórios para estudar a chance de ocorrência de desfechos.
3. **Inferência estatística.**
   - Estudo da população por meio de evidência fornecida pela amostra.

Áreas da Estatística

Fora da Estatística “básica”, existem diversos temas:

- Métodos de amostragem.
- Planejamento de experimentos.
- Controle estatístico de qualidade.
- Modelos de regressão.
- Análise de sobrevivência.
- Análise de dados correlacionados.
- Análise de séries temporais.
- Inferência Bayesiana.
- Aprendizado de máquina.
- Inferência causal.

Estatística e o desenvolvimento científico

Estatística e o desenvolvimento científico

- A Estatística está diretamente associada com o **método científico**.
  - Definimos uma **hipótese**.
  - Confrontamos esta hipótese com **evidências** (dados).
  - Com base nas evidências **rejeitamos** ou **não rejeitamos** as hipóteses iniciais.
  - Os resultados conduzem a **novas hipóteses** e o ciclo se reinicia.

- Praticamente todas as áreas do conhecimento humano requerem instrumentos para **análise de dados**.

A importância de resultados não significativos

- Muitos pesquisadores deixam de tornar públicos resultados não significativos.
- Contudo resultados não significativos são tão importantes quanto os significativos.
- A hipótese de interesse, rejeitada ou não rejeitada, fornece conhecimento a respeito do problema sob análise.

Figura 6. Extraído de pixabay.com.

Estatística e ética

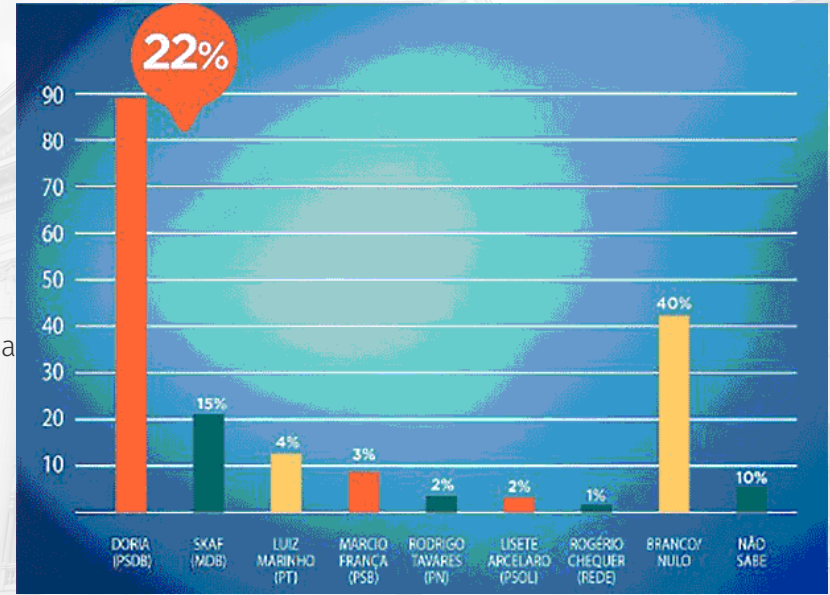
Estatística e ética

- Cuidados devem ser tomados na escolha do tipo análise a ser realizada.
- O uso e divulgação **ética** e **criteriosa** de dados e resultados de análises devem ser pré-requisitos **indispensáveis** e **inegociáveis** à qualquer analista.

Figura 7. Extraído de pixabay.com.

Estatística e ética

- Por exemplo, no contexto de gráficos, devemos evitar que o gráfico fique desproporcional ou privilegiando determinados valores a fim de induzir conclusões àqueles que utilizam o gráfico como forma de visualização.



Estatística e o desenvolvimento computacional

Estatística e o desenvolvimento computacional

- A popularização da Estatística se deu graças ao desenvolvimento computacional.
- Os computadores pessoais tornaram os métodos estatísticos mais acessíveis ao público geral por meio de softwares que implementam as metodologias.

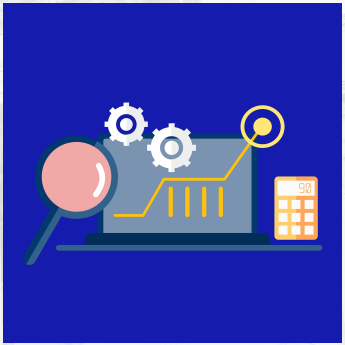


Figura 9. Extraído de pixabay.com.

Estatística e o desenvolvimento computacional

- Devido ao avanço computacional, houve um aumento considerável na capacidade de produzir e armazenar dados provenientes das mais diversas fontes.
- Graças ao avanço computacional podemos lidar com a manipulação de grandes conjuntos de dados.

Figura 10. Extraído de pixabay.com.

Estatística e o desenvolvimento computacional

- Este grande volume de dados também força o desenvolvimento dos métodos estatísticos e softwares para análise de dados.
- A capacidade computacional atual também desperta o interesse por métodos estatísticos computacionalmente intensivos.

Figura 11. Extraído de pixabay.com.

Considerações

- Onde há dados e incerteza, a Estatística pode ser usada.
- A Estatística vai muito além do senso comum: tabelas e gráficos em revistas esportivas e jornais ou pesquisas de intenção de voto em épocas de eleição.
- Existem diversas técnicas e possíveis áreas de aplicação.
- A Estatística está por trás de boa parte do desenvolvimento científico moderno.

Algumas leituras recomendadas

Livros técnicos

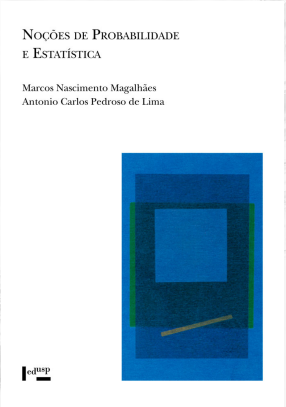


Figura 12. Noções de Probabilidade e Estatística.

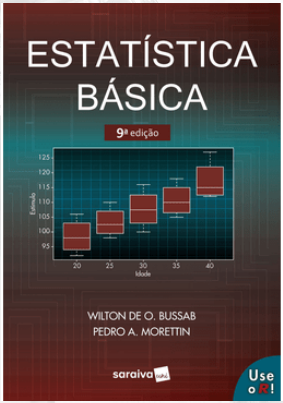


Figura 13. Estatística Básica.

Livros não técnicos

Figura 14. Uma senhora toma chá.

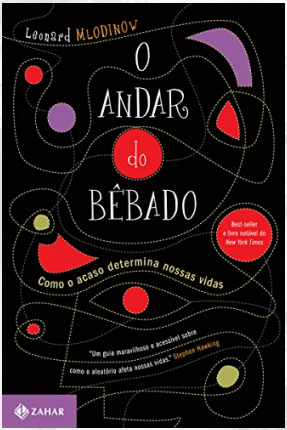


Figura 15. O andar do bêbado.

Livros não técnicos

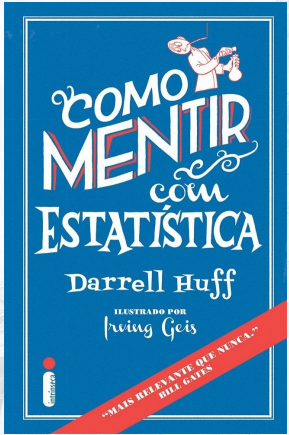


Figura 16. Como mentir com Estatística.

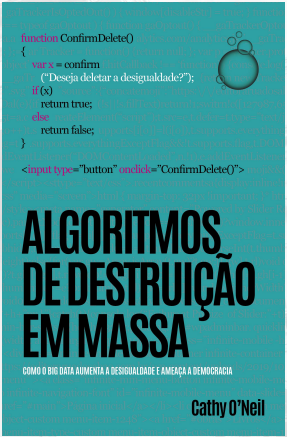
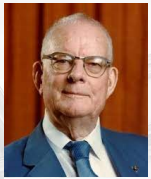


Figura 17. Algoritmos de destruição em massa.

Algumas frases para refletir

"Em Deus nós confiamos; todos os outros devem trazer dados." William Edwards Deming



"O trabalho do estatístico é o de catalisar o processo de construção do conhecimento científico." George E. P. Box

"A tentação de formular teorias prematuras sobre dados insuficientes é a ruína da nossa profissão." Sherlock Holmes, de Sir Arthur Conan Doyle

Dados

O que são dados?

- Dados são **conjuntos de valores**.
- Podem ser de diferentes fontes, tais como **estudos** e **experimentos**.
- Podem conter **variáveis** de diferentes tipos.
- Podem surgir em formatos **estruturados** e **não estruturados**.

Figura 18. Extraído de pixabay.com.

Conjunto de dados

- Em Estatística, em geral, lidamos com **dados estruturados em um formato tabular**.
- Os dados nem sempre começam nessa forma. Muitas vezes a informação deve ser processada e tratada de modo a chegar nesta estrutura.
- O conjunto de dados completo e sem tratamentos é denominado conjunto de **dados brutos**.
- Um conjunto de dados considerado **arrumado** é aquele em que:
  - Cada **coluna** representa uma **variável**.
  - Cada **linha** representa uma **observação**.
  - Cada **célula** representa o **valor** observado.

Conjunto de dados

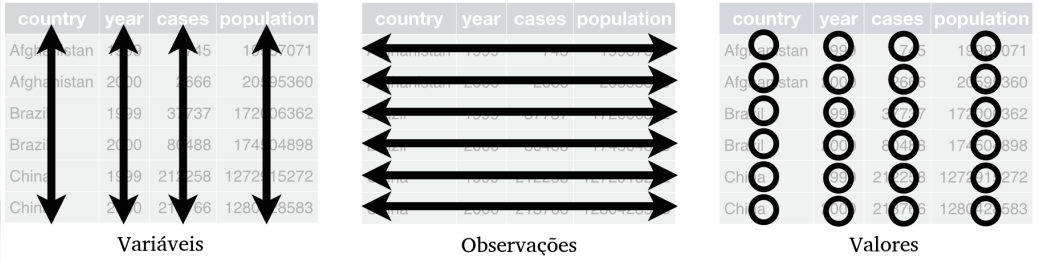


Figura 19. Adaptado de https://r4ds.had.co.nz.

Conjunto de dados

Tabela 1. Exemplo de conjunto de dados

ID	Sexo	Escolaridade	Altura	Peso	Irmãos
1	Masculino	Ensino superior	182	80	0
2	Feminino	Ensino médio	160	46	1
3	Feminino	Ensino superior	160	55	4
4	Feminino	Mestrado	165	58	3
5	Masculino	Ensino médio	183	55	1

Fontes de dados

De onde vêm os dados?

Alguns exemplos:

- Estudos de caso.
- Experimentos.
- Pesquisas.
- Registros administrativos.
- Dados em repositórios online.
- Bancos de dados corporativos.
- Sensores.
- Textos, imagens e vídeos.

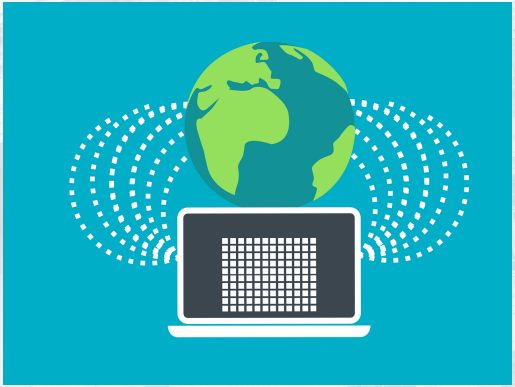


Figura 20. Extraído de pixabay.com.

Dados observacionais x dados experimentais

Dados observacionais

- **Observação passiva** da realidade.
- Sem modificação das condições.

Dados experimentais

- **Intervenção** na realidade.
- Condições controladas.
- Observação dos efeitos das intervenções.

Figura 21. Extraído de pixabay.com.

Dados observacionais x dados experimentais

- Cada tipo de estudo induz **relações** diferentes entre as observações e **modelos estatísticos** diferentes para modelar a incerteza destas relações.
- Um **conjunto de dados** é um dos subprodutos de um estudo. Ele contém as características principais (variáveis) que se tem interesse em estudar em uma população ou amostra.
- Estas características podem ser **qualitativas** ou **quantitativas** e a partir do conjunto de dados as análises inferenciais são feitas.
- As variáveis são assim chamadas porque seus valores não são constantes e variam segundo regras ou leis naturais que podem ser conhecidas ou desconhecidas.

Tipos de variáveis

Tipos de variáveis

- Na prática, podemos coletar variáveis de diferentes tipos e naturezas.
- Antes de de qualquer análise precisamos ser capazes de compreender os tipos de variáveis pois estes tipos conduzirão às análises e métodos estatísticos que poderão ser aplicados.

- Existem dois tipos (básicos) de variáveis:
  - Numéricas (**quantitativas**).
  - Não numéricas (**qualitativas**).

Figura 22. Tipos básicos de variáveis.

Variáveis quantitativas

- **Variáveis Quantitativas**: assumem valores numéricos.
  - **Discretas**: características mensuráveis que podem assumir apenas um número finito ou infinito **contável** de valores.
  - **Contínuas**: características mensuráveis que assumem valores em uma **escala contínua**, isto é, na reta real.

Exemplos

- Altura.
- Peso.
- Idade.
- Percentual de gordura corporal.
- Número de filhos.
- Número de fraturas.
- Número de faltas.
- Número de peças defeituosas em um lote.

Variáveis qualitativas

- **Variáveis Qualitativas**: são as características definidas por categorias, ou seja, representam uma classificação dos indivíduos e não uma característica numérica.
  - **Nominais**: não existe ordenação nem peso entre as categorias.
  - **Ordinais**: existe uma ordenação entre as categorias.

Exemplos

- Estado civil.
- Orientação sexual.
- Turma.
- Posição em que joga em um time.
- Severidade de uma lesão.
- Escolaridade.
- Grau de proficiência em língua inglesa.
- Risco de infarto.

Cuidados com variáveis

- Existem particularidades na classificação de variáveis devido a situações como:
  - Discretização de variáveis contínuas.
  - Limitações em instrumentos de mensuração.
  - Utilização de quantidades numéricas para representação de variáveis categóricas.
  - Dentre outras.
- Deve-se sempre estar atento a este tipo de situação pois podem levar a implicações nas análises e consequentemente nos resultados.
- Existem outros tipos de variáveis que ocorrem em situações particulares que requerem técnicas específicas de análise.

Análise de dados

No que devemos pensar antes de analisar nossos dados?

- O que estamos interessados em avaliar?
- Quais são as variáveis de interesse?
- Quais são as variáveis que queremos avaliar se influenciam a variável de interesse?
- Quais são os métodos disponíveis para análise de variáveis deste tipo?
- Quais os métodos disponíveis que permitem responder nossa pergunta de pesquisa?
- Como coletar os dados?
- Os dados são válidos?

Métodos de amostragem

Amostras

- Uma amostra é um **subconjunto da população**.
- Na prática costuma ser inviável trabalhar com a população toda.
- A alternativa então é trabalhar com uma **amostra** e **inferir** os resultados para a população.
- A seleção da amostra pode ser feita de diversas maneiras.

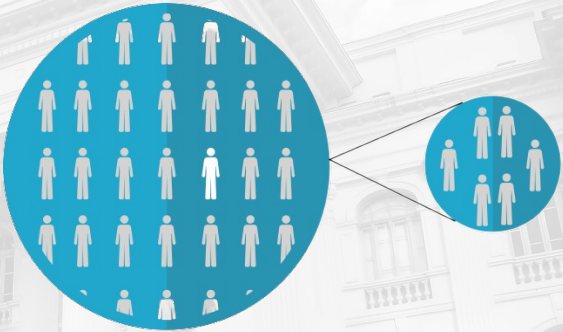


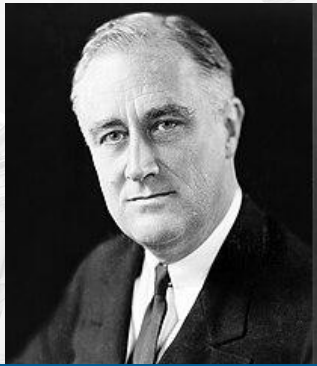
Figura 23. Extraído de pixabay.com.

Amostras

- Os métodos de amostragem servem para selecionar subconjuntos da população de forma mais **representativa** possível.
- A forma adequada de amostragem conduz a um **menor tamanho amostral** para obtenção de uma **precisão satisfatória**.
- São características desejáveis de uma amostra:
  - Capacidade de generalização.
  - Imparcialidade e representatividade.
  - Capacidade de medir a precisão das estimativas.
- Podemos dividir os métodos em:
  - Amostragem probabilística.
  - Amostragem não probabilística.

Um caso clássico: a história do Literary Digest

- O **Literary Digest** era uma revista americana de publicação semanal fundada em 1890.
- Em 1936 ocorreu a 38ª **eleição presidencial** dos Estados Unidos.
- Como candidatos haviam nomes como: Franklin Roosevelt, Alf Landon, William Lemke, Norman Thomas, dentre outros.
- **Roosevelt** e **Landon** eram vistos como os favoritos.




Um caso clássico: a história do Literary Digest

- No ano da eleição, o Literary Digest conduziu uma pesquisa de intenção de votos com **mais de 10 milhões de respondentes** com base em sua base de assinantes e outras listas de indivíduos.
- Enquanto isso, George Gallup, fundador da Gallup Poll, conduziu pesquisas quinzenais com apenas **2 mil indivíduos**.
- O Literary Digest previu a vitória de Landon, Gallup previu a vitória de Roosevelt. Qual dos dois acertou?

Um caso clássico: a história do Literary Digest

O resultado da eleição foi:

1. **Franklin D. Roosevelt, 27.752.648 de votos.**
2. Alf Landon, 16.681.862 de votos.
3. William Lemke, 892.378 votos.
4. Norman Thomas, 187.910 votos.
5. Outros, 132.901 votos.

- Gallup acertou, Literary Digest errou.
- O que deu errado na pesquisa do Literary Digest?
  - A resposta é: a **composição da amostra**.

Um caso clássico: a história do Literary Digest

- O Literary Digest optou por **quantidade**, prestando pouca atenção ao método de seleção.
- A amostra foi de **conveniência** e representava apenas o grupo da população com nível socioeconomico relativamente alto: seus próprios assinantes e pessoas que possuiam luxos da época como telefones.
- Isso gerou um **viés de amostragem**, ou seja, a amostra era diferente, de modos importantes e não aleatórios, da população que deveria representar. Ou simplesmente: **a amostra não era representativa**.
- Por outro lado, a amostra de Gallup era bem mais modesta, contudo o método de seleção gerou uma **amostra representativa da população** em que todas as camadas de votantes estavam presentes.

Amostragem probabilística

- Amostragem probabilística deve ser usada **sempre que possível**.
- O objetivo é dimensionar amostras que sejam capazes de **estimar** as quantidades de interesse com uma certa **precisão** desejada.
- Existem diversos métodos disponíveis.

Alguns métodos são:

- Amostragem aleatória simples (com ou sem reposição).
- Amostragem sistemática.
- Amostragem estratificada.
- Amostragem por conglomerados.

Amostragem não probabilística

- Em muitos casos não é possível fazer uso de métodos de amostragem probabilística.
- Surgem então os métodos de amostragem não probabilística, como amostragem por conveniência, intencional/julgamento, bola de neve.
- Uma avaliação da "representatividade" dos métodos de amostragem não probabilística não pode ser feita.
- Devemos tomar muito cuidado ao interpretar resultados baseados em métodos de amostragem não probabilísticos.
- Em geral, estas amostras carregam um alto risco de não serem representativas.
- Não há métodos para análise probabilística ou inferencial dos resultados.

Análise exploratória

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**.

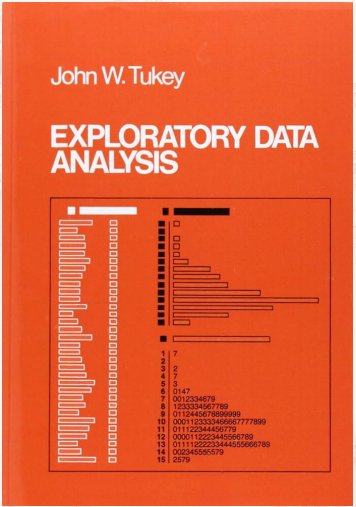


Figura 25. Capa do livro Exploratory Data Analysis de John Tukey.

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 26. 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 27. 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 2. 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 28. 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 3. 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 4. 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 5. 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 6. 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 7. 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

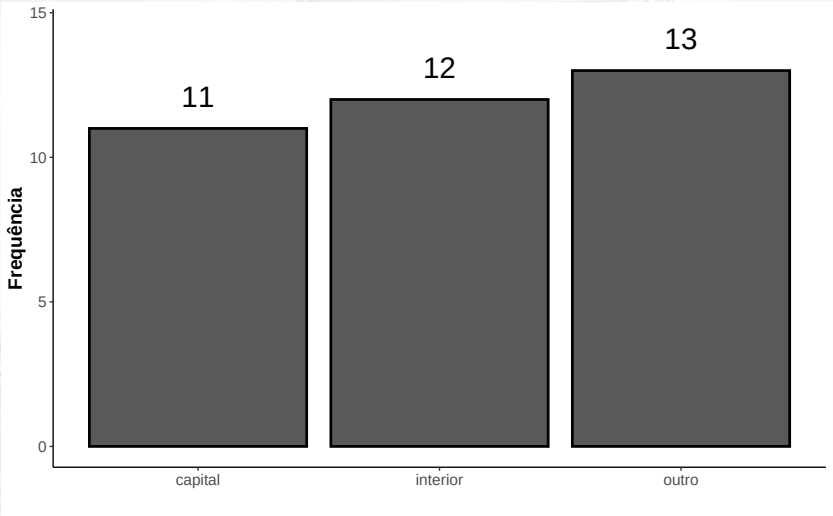


Figura 29. Gráfico de barras verticais para a região.

Gráfico de barras verticais

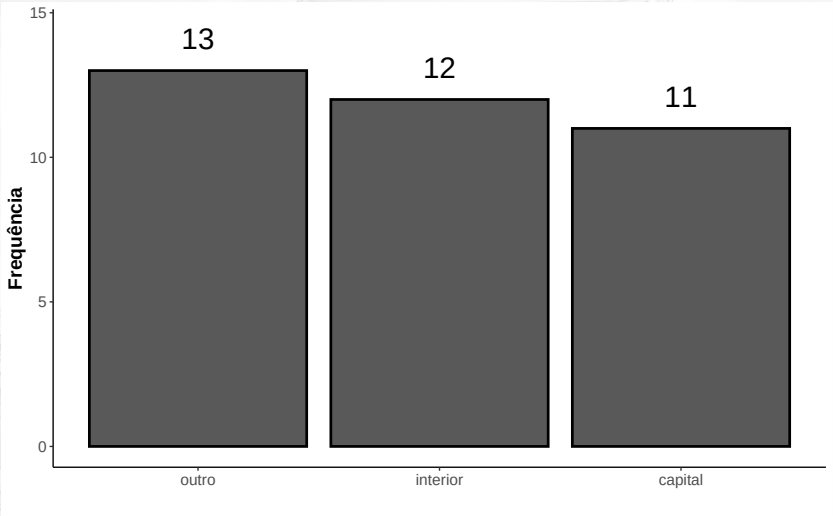


Figura 30. Gráfico de barras verticais para a região.

Gráfico de barras horizontais

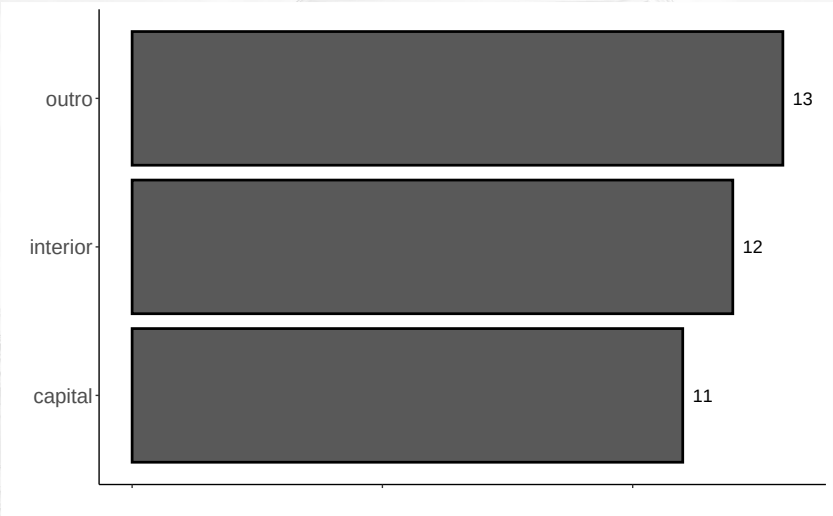


Figura 31. Gráfico de barras horizontais para a região.

Gráfico de barras empilhadas

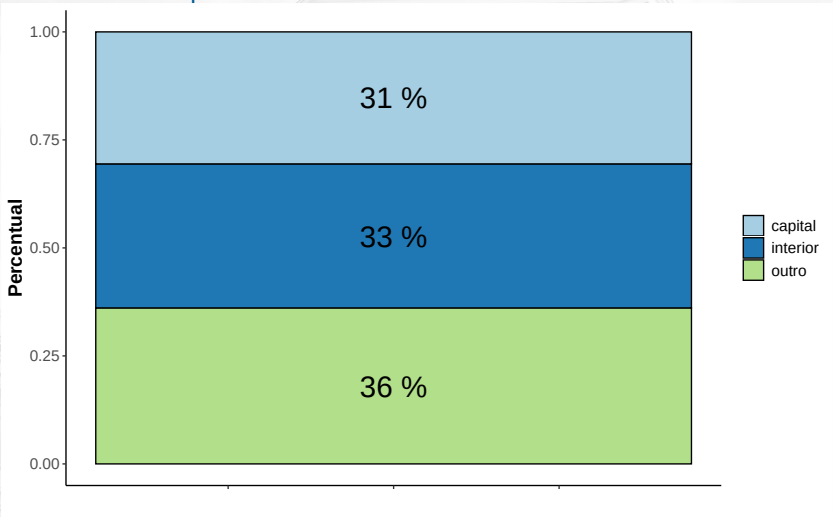


Figura 32. Gráfico de barras empilhadas para a região.

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

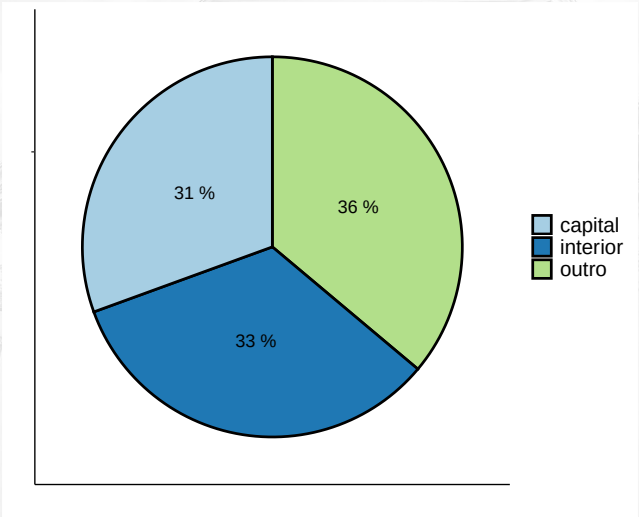


Figura 33. Gráfico de setores para a região

Análise descritiva univariada para variáveis quantitativas

Análise descritiva univariada para variáveis quantitativas

- Uma variável quantitativa é uma **característica** que pode ser **mensurada** e representada **numericamente**.
- Podem ser classificadas em **discretas** (finitos valores em um dado intervalo) ou **contínuas** (infinitos valores em um dado intervalo).
- Quando estamos lidando com **variáveis quantitativas discretas com poucos possíveis valores**, as técnicas apresentadas para variáveis qualitativas se aplicam.

Tabelas de frequência

Tabela 8. Tabela de frequências para o número de filhos (desconsiderando dados ausentes).

Filhos	Frequência	Percentual	Freq. Acumulada	Percentual Acumulado
0	4	20 %	4	20 %
1	5	25 %	9	45 %
2	7	35 %	16	80 %
3	3	15 %	19	95 %
4	0	0 %	19	95 %
5	1	5 %	20	100 %
Total	20	100 %	20	100 %

Gráfico de barras verticais

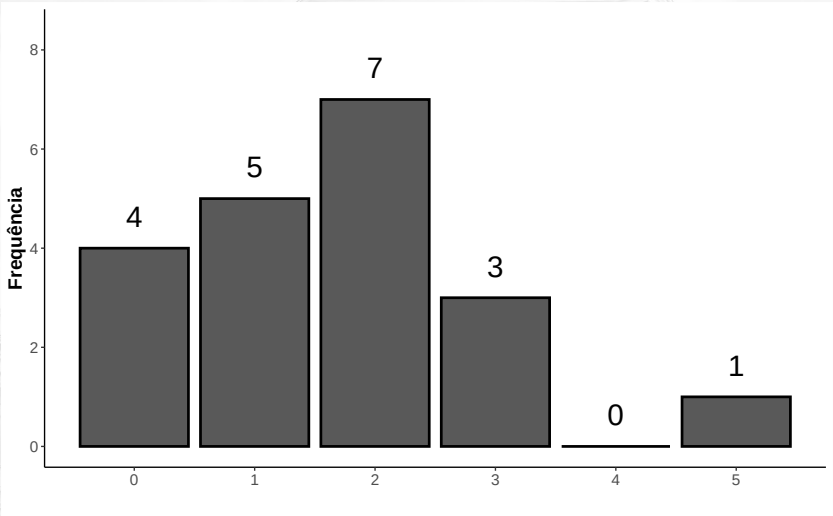


Figura 34. Gráfico de barras verticais para o número de filhos.

Análise descritiva univariada para variáveis quantitativas

- Para variáveis quantitativas **contínuas** ou **discretas com muitos possíveis valores**, precisamos de técnicas específicas.
- Uma estratégia comum é o **agrupamento em faixas de valores**, e avaliação das frequências nestas faixas.
- Podem ser usadas **tabelas de frequências** absolutas, relativas e acumuladas para as faixas de valores.
- Utilizando a **razão entre frequência relativa e a amplitude das faixas** de valores, geramos a **densidade**.

Análise descritiva univariada para variáveis quantitativas

Faixas de valores

- Cuidados devem ser tomados quanto às notações e tipos de faixas (aberto e fechado à esquerda ou direita).
- Diferentes pessoas e softwares podem usar intervalos distintos.
- Em geral usaremos intervalos **fechados à esquerda** e **abertos à direita**.
- Considerando dois valores a e b, em que $a < b$, os intervalos consideram que a **não** está incluído na faixa, b está.
- Notações usuais:
  - $a \leq y < b$.
  - $a \vdash b$.
  - $[a,b)$.
  - $[a,b[$.
- Exemplo:
  - $5 \leq y < 10$.
  - $5 \vdash 10$.
  - $[5,10)$.
  - $[5,10[$.
  - Valores maiores ou iguais a 5 até valores menores que 10 (10 não está no intervalo).

Análise descritiva univariada para variáveis quantitativas

Perguntas que surgem são:

- Como agrupar em classes?
- Qual o tamanho ideal das faixas de valores?

- Classes definidas com a **mesma amplitude** é o procedimento mais usual, apesar de ser possível definir classes com tamanhos diferentes.
- Existem procedimentos que podem ser usados para obter a amplitude, como **Sturges**.
- Em geral, **5** a **15** faixas são suficientes.

Tabelas de frequência para uma variável quantitativa

Tabela 9. Tabela de frequências usando faixas de salários.

Faixas	Frequência	Freq. Relativa	Freq. Acumulada	Freq. Rel. Acumulada
[4,6)	4	0.11	4	0.11
[6,8)	6	0.17	10	0.28
[8,10)	8	0.22	18	0.5
[10,12)	4	0.11	22	0.61
[12,14)	5	0.14	27	0.75
[14,16)	3	0.08	30	0.83
[16,18)	3	0.08	33	0.91
[18,20)	2	0.06	35	0.97
[20,22)	0	0.00	35	0.97
[22,24]	1	0.03	36	1
Total	36	1.00		

Tabelas de frequência para uma variável quantitativa

Tabela 10. Tabela de frequências usando faixas de salários.

Faixas	Frequência	Percentual	Freq. Acumulada	Percentual Acumulado
[4,6)	4	11 %	4	11 %
[6,8)	6	17 %	10	28 %
[8,10)	8	22 %	18	50 %
[10,12)	4	11 %	22	61 %
[12,14)	5	14 %	27	75 %
[14,16)	3	8 %	30	83 %
[16,18)	3	8 %	33	91 %
[18,20)	2	6 %	35	97 %
[20,22)	0	0 %	35	97 %
[22,24]	1	3 %	36	100 %
Total	36	100%		

Tabelas de frequência para uma variável quantitativa

Tabela 11. Tabela de frequências usando faixas de salários.

Faixas	Frequência	Percentual	Freq. Acum.	Perc. Acum.	Amplitude	Densidade
[4,6)	4	11 %	4	11 %	2	0.055
[6,8)	6	17 %	10	28 %	2	0.085
[8,10)	8	22 %	18	50 %	2	0.11
[10,12)	4	11 %	22	61 %	2	0.055
[12,14)	5	14 %	27	75 %	2	0.07
[14,16)	3	8 %	30	83 %	2	0.04
[16,18)	3	8 %	33	91 %	2	0.04
[18,20)	2	6 %	35	97 %	2	0.03
[20,22)	0	0 %	35	97 %	2	0
[22,24]	1	3 %	36	100 %	2	0.015
Total	36	100%				

Gráficos para representação de frequências de uma variável quantitativa

- Assim como no caso de variáveis qualitativas ou quantitativas discretas com poucos possíveis valores, a representação por meio de gráficos pode ser bastante benéfica para análise de variáveis quantitativas.

Algumas possibilidades são

- Histograma.
- Gráfico de densidade empírica.
- Box-plot

Histograma

- Consiste em **retângulos contíguos** de base dada pelas faixas de valores definindas para uma variável.
- Algumas possibilidades são:
  - A **área** representar a **frequência** da respectiva faixa.
  - A **altura** representar a **frequência** absoluta na faixa.
  - A **altura** representar o quociente da área pela amplitude da faixa: a **densidade**.

Histograma

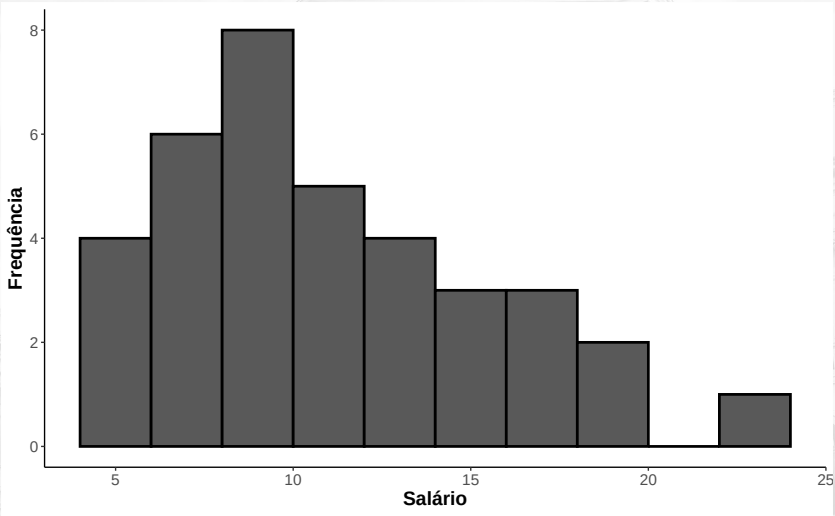


Figura 35. Histograma dos salários.

Efeito do número de classes

- O número de classes pode afetar diretamente as tabelas e gráficos.
- Com poucas classes, os dados ficam excessivamente resumidos e as classes ficam muito heterogêneas.
- Com muitas classes, os dados ficam segmentados em excesso e as representações são comprometidas.

Efeito do número de classes

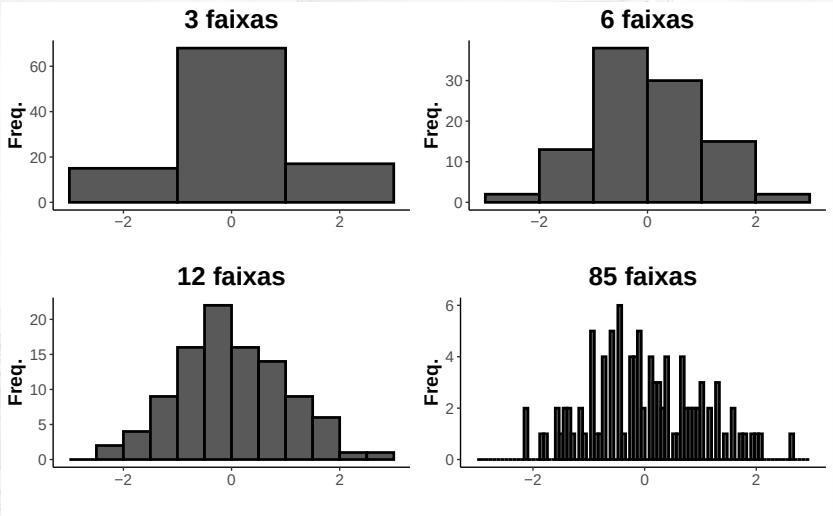


Figura 36. Efeito do número de classes em histogramas.

Gráfico de densidade empírica

Intuição

- Imagine uma sequência de histogramas de densidade em que o número de observações aumenta, juntamente com o número de faixas.
- No limite, teremos uma **curva**.
- Esta curva é chamada de gráfico de **densidade empírica**.
- É um gráfico "computacionalmente intensivo", depende da definição de uma função kernel e do tamanho da banda.
- A área sob a curva é igual a 1.
- Outra forma de ver o gráfico de densidade empírica é como um **histograma suavizado**.

Gráfico de densidade empírica

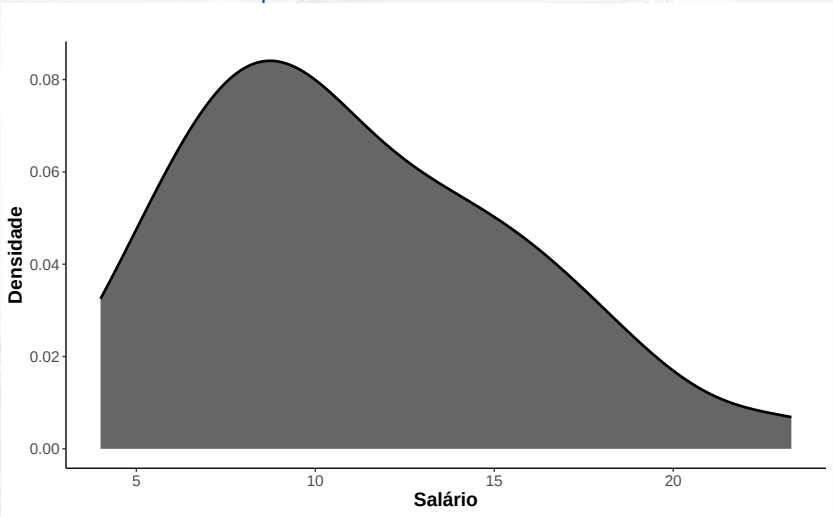


Figura 37. Gráfico de densidade dos salários.

Box-plot

- Outra importante visualização é o **box-plot**.
- É possível analisar a **distribuição** dos dados, aspectos quanto a **posição**, **variabilidade**, **assimetria** e também a presença de **valores atípicos**.
- Retomaremos o box-plot após estudar quartis, em medidas descritivas.

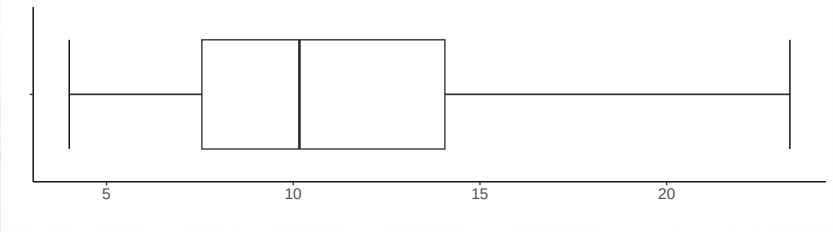


Figura 38. Box-plot dos salários.

Histograma, densidade e box-plot

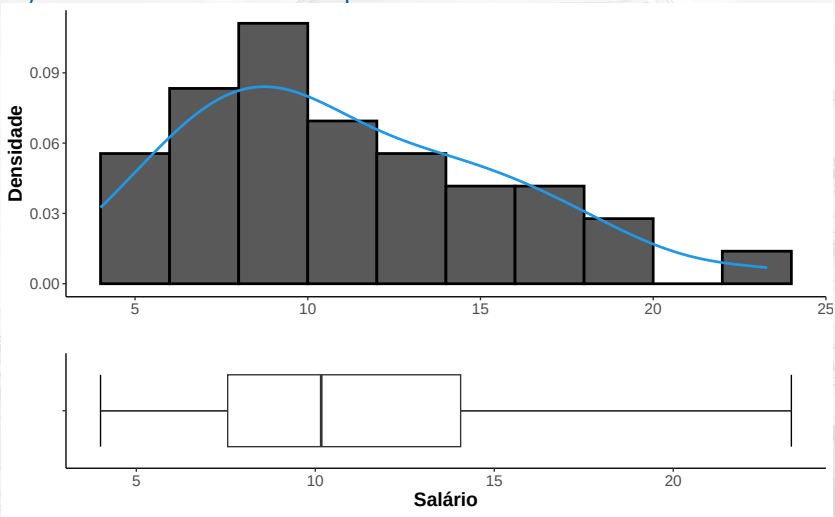


Figura 39. Combinação de representações.

Assimetria

- Um conjunto de valores pode ser aproximadamente **simétrico**, **assimétrico** à esquerda ou à direita.
- Tais características são facilmente diagnosticadas por meio de **análise gráfica** usando um histograma, gráfico de densidade ou box-plot.
- Futuramente veremos como diagnosticar assimetria por meio de **medidas descritivas**.

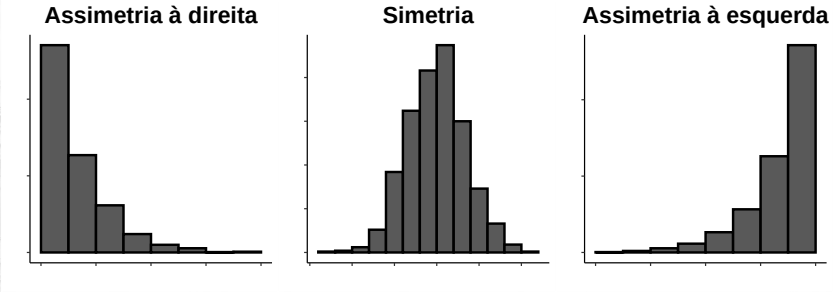


Figura 40. Ilustração assimetria.

O que foi visto:

- Introdução e conceitos fundamentais.
- Áreas da Estatística.
- Estatística e desenvolvimento científico, ética e desenvolvimento computacional.
- Leituras recomendadas.
- Dados e fontes de dados.
- Tipos de variáveis.
- Análise de dados.
- Considerações sobre amostragem.
- Introdução à análise exploratória.
- Análise exploratória univariada para variáveis qualitativas e quantitativas.

Próximos assuntos:

- Resumos numéricos.
- Medidas de posição central.
- Medidas de posição relativa.
- Medidas de dispersão.
- Análises bivariadas.