WBA0752_v1.1

TÉCNICAS ESTATÍSTICAS: TEORIA E PRÁTICA (*R PROGRAMMING*)

Marcelo Tavares de Lima

A Técnicas estatísticas: teoria e prática (*R Programing*)

1ª edição

Londrina
Editora e Distribuidora Educacional S.A.
2019

Dados Internacionais de Catalogação na Publicação (CIP)

Lima, Marcelo Tavares de
L732t Técnicas estatísticas: teoria e prática (R Programing)/
Marcelo Tavares de Lima, – Londrina: Editora e Distribuidora Educacional S.A. 2019.
155 p.

ISBN 978-85-522-1650-6

1. DW (Data Warehouse). 2. Linguagem R. I. Lima, Marcelo Tavares de. Título.

CDD 004

Thamiris Mantovani CRB: 8/9491

2019
Editora e Distribuidora Educacional S.A.
Avenida Paris, 675 – Parque Residencial João Piza
CEP: 86041-100 — Londrina — PR
e-mail: editora.educacional@kroton.com.br
Homepage: http://www.kroton.com.br/

SUMÁRIO

Apresentação da disciplina

Caro aluno,

Em tempos de grandes volumes de informações, o correto tratamento e a correta manipulação destas torna-se imprescindível para quem as utiliza. Para tanto, faz-se necessária a busca por técnicas e ferramentas apropriadas para isso, assim como a formação de profissionais preparados e habilitados.

As técnicas estatísticas existem há bastante tempo e, desde a sua criação, são utilizadas de maneira exaustiva para o tratamento e manipulação de dados. Seu uso tem aumentando quase que de maneira exponencial com o passar dos anos, o que é uma consequência do aumento da busca por este conhecimento.

O profissional que precisa lidar com bases de dados, sejam elas pequenas, médias ou grandes, precisa saber utilizar corretamente as técnicas de extração de informações e de manipulações desses dados, pois o uso incorreto da técnica pode ter como resultado algo inesperado, incorreto e até mesmo completamente diferente do que se esperava no início do trabalho.

O conjunto de técnicas estatísticas existentes para tratamento e manipulação de dados é bastante amplo, que vai desde técnicas descritivas e exploratórias até técnicas mais sofisticadas, como a inferência e modelagem.

O produto do uso das técnicas estatísticas pode ser apresentado em tabelas e gráficos apropriados para o tipo de informação tratada e manipulada. A elaboração de tabelas e gráficos pode ser realizada com o suporte de um programa computacional, que pode ser uma planilha eletrônica, como o MS Excel ou um programa mais apropriado ao trabalho, como o programa R.

O R é um programa de código aberto e gratuito utilizado para o tratamento de bases de dados, com o uso de técnicas estatísticas, mineração de dados, elaboração de gráficos e de tabelas, assim como relatórios gerenciais diversos.

A proposta desta disciplina é apresentar ao aluno as técnicas estatísticas mais utilizadas para o tratamento de dados, com a apresentação de um suporte teórico, mas com principal enfoque na aplicação em programação R para obtenção de resultados analíticos, assim como produtos para subsidiar relatórios como tabelas e gráficos.

Estatística e linguagem R: apresentação, histórico e principais conceitos

Autor: Marcelo Tavares de Lima

Objetivos

- Descrever um breve histórico da estatística.
- Apresentar os principais conceitos de estatística.
- Apresentar um breve histórico sobre o R e a estatística.

1. Introdução

Neste texto serão apresentados os principais conceitos de estatística, tais como população, amostra, variável, dentre outros, para que você se torne familiarizado com os mesmos e compreenda com facilidade quando forem citados.

Também será apresentado um breve histórico da estatística, assim como um breve histórico sobre o programa computacional R, o que é um programa de código aberto (*open source*) e gratuito, bastante utilizado para o tratamento e análise estatística de dados.

Ao final deste texto, pretendemos deixá-lo familiarizado com os principais conceitos de estatística, assim como com os principais conceitos de R, para que ao longo do curso você possa utilizá-los de maneira automática e otimizada.

PARA SABER MAIS

O aprendizado do R, segundo Aquino (2014, p. 1), "é difícil no início devido à necessidade de se adaptar à sua lógica de funcionamento, se acostumar com a estrutura dos seus documentos de ajuda e memorizar alguns comandos básicos". No entanto, após superado esse contato inicial, é possível perceber as vantagens do seu uso. Uma das principais se refere ao fato de ser um programa gratuito e de código aberto (*open source*).

ASSIMILE

Os métodos estatísticos compõem uma metodologia que possui processos apropriados para coletar, organizar e interpretar dados, estimar quantidades relevantes, investigar hipóteses, realizar inferências sobre uma população com base em amostras e, também, pode auxiliar na identificação de fatores importantes em um estudo, possibilitando a realização de previsões, etc.

1.1 Principais conceitos em estatística e um pouco de história

A palavra estatística surgiu pela primeira vez no século XVIII, sugerida pelo alemão Gottfried Achemmel (1719-1772) (MEDEIROS, 2009). É uma palavra derivada do termo *status*, que significa estado em latim, o que a relaciona, basicamente, com as atividades do Estado, pois na época do surgimento do termo era uma atividade exclusivamente realizada pelo Estado.

No entanto, antes de Achemmel sugerir o termo estatística, em 1662, John Graunt publicou informes sobre nascimentos e mortes, ou seja, relatórios estatísticos. Seu trabalho foi complementado por estudos de mortalidade e taxas de morbidade, tamanhos populacionais, informações de renda e taxas de desemprego. Tudo isso, sem ter feito citação sobre o termo estatística (TRIOLA, 2008).

A história da estatística tem sua origem nos negócios do Estado (MEDEIROS, 2009). No entanto, podia ser encontrada nas mais diversas atividades, como na agricultura, na saúde, educação, ciências sociais, etc.

Ao longo de sua utilização e desenvolvimento, a estatística passou a ser aplicada em ramos diversos, como citado anteriormente, e se tornou, juntamente com as ciências econômicas, uma ciência social por

excelência (MEDEIROS, 2009). Tal mudança aconteceu porque passou a lidar com grandes quantidades numéricas e grandes volumes de dados, tornando-a, praticamente, obrigatória para tratamento adequado dessas grandes massas de dados.

Medeiros (2009) apresenta um relato histórico sobre os primeiros levantamentos estatísticos que se tem registrado na história.

> O primeiro levantamento estatístico de que se tem conhecimento se deve a Heródoto e se refere a um estudo da riqueza da população do Egito, cuja finalidade era averiguar quais eram os recursos humanos e econômicos disponíveis para a construção das pirâmides, isso no ano de 3.050 a.C. No ano de 2.238 a.C., o imperador Chinês Yao ordenou a realização de uma Estatística com fins industriais e comerciais. No ano de 1.400 a.C., o famoso faraó egípcio Ramsés II ordenou um levantamento das terras do Egito. (MEDEIROS, 2009, p. 17)

Ao longo dos tempos, é possível apresentar a importância da estatística, assim como seus principais interesses, de forma sucinta, em quatros fases, descritas por Medeiros (2009) conforme o Quadro 1.

Quadro 1 – As fases de desenvolvimento da estatística

Primeira fase	Pepino, no ano de 758, e Carlos Magno, em 762, realizaram estatísticas sobre as terras que eram propriedade da Igreja. Essas foram as únicas estatísticas importantes desde a queda do Império Romano, que se deu a partir do século III d.C.
Segunda fase	Na Inglaterra, no século XVII, já se analisavam grupos de observações numéricas referentes à saúde pública, nascimentos, mortes e comércio. Destacam-se, nesse período, John Graunt (1620-1674) e William Petty (1623-1687), que procuraram leis quantitativas para traduzir fenômenos sociais e políticos.
Terceira fase	Também no século XVII, inicia-se o desenvolvimento do cálculo das probabilidades que, juntamente com os conhecimentos estatísticos, redimensionou a estatística. Nessa fase, destacaram-se: Fermat (1601-1664), Pascal (1623-1662) e Huygens (1629-1695).

Quarta fase	No século XIX, inicia-se a última fase do desenvolvimento da estatística, alargando e interligando os conhecimentos adquiridos nas três fases anteriores. Nesta fase, a estatística não se limita apenas ao estudo da demografia e da economia, como antes. Agora, o seu campo de aplicação se estende à análise de dados em biologia, medicina, física, psicologia, indústria, comércio, meteorologia, educação, etc., e, ainda, a domínios aparentemente desligados, como estrutura de linguagem e estudo de formas literárias. Destacam-se, no período, Ronald Fisher (1890-1962) e Karl Pearson (1857-1936).

Fonte: adaptado de MEDEIROS (2009).

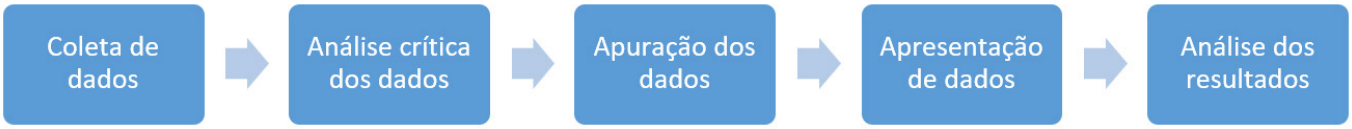
Ao longo do desenvolvimento histórico da estatística (Quadro 1), onde ocorreu a sua consolidação, algumas ferramentas passaram a ser utilizadas com maior frequência, como as tabelas mais complexas, os gráficos mais elaborados, fazendo assim, com que a estatística deixasse de ser uma simples catalogação de dados numéricos coletivos para se tornar um método de análise inferencial ou análise do todo a partir de dados.

Com o seu estabelecimento, a estatística passou a ter um conceito, de certa forma, padrão, o qual pode ser descrito como, segundo Medeiros (2009, p. 18 *apud* Crespo, 1995), "uma parte da Matemática Aplicada que fornece métodos para a coleta, organização, descrição, análise e interpretação de dados".

A partir de então, a estatística deixou de ser uma ferramenta limitada à organização e descrição de dados para proporcionar métodos inferenciais que pudessem permitir a extrapolação de conclusões a partir dos dados amostrais, podendo superar "achismos" e "casuísmos" corriqueiros.

Ainda, considerando as etapas descritas por Medeiros (2009), no conceito de estatística é possível descrevê-las com maiores detalhes, conforme apresentado a seguir e resumido na Figura 1.

Figura 1 – Etapas do trabalho estatístico



Fonte: elaborada pelo autor.

1.1.1 Coleta de dados

A coleta dos dados pode ser descrita como a busca ou a compilação das informações ou variáveis que representam o(s) fenômeno(s) a ser(em) estudado(s).

Uma coleta de dados pode ser direta ou indireta. Entende-se por coleta direta a coleta realizada sobre registros de informação obrigatória, por exemplo, como nascimentos, casamentos, óbitos, divórcios, etc., ou quando os dados são coletados pelo próprio pesquisador, por intermédio de inquéritos ou questionários, como, por exemplo, as pesquisas realizadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE) (CRESPO, 2009).

Ainda é possível classificar uma coleta direta de dados como contínua, periódica e ocasional. Entende-se como coleta direta contínua os registros realizados de forma contínua, como os registros de nascimentos. Já a coleta periódica de dados ocorre quando é realizada em intervalos constantes de tempo, como as pesquisas do IBGE. A coleta ocasional acontece quando se faz coleta extemporaneamente, com o intuito de atender uma demanda específica.

Diz-se que a coleta de dados é indireta quando é realizada sobre dados conhecidos existentes, coletados de forma direta por outras fontes e/ou quando se coleta de um fenômeno que não seja o de interesse, mas que se sabe que está relacionado com o objeto de interesse. Um exemplo seria o uso dos dados das pesquisas do IBGE para a realização de outras pesquisas.

1.1.2 Análise crítica dos dados

É um momento de verificação de erros e inconsistências. É a busca de falhas e imperfeições. Nessa etapa, os dados precisam ser analisados criticamente, de maneira cuidadosa, para que não ocorram erros grosseiros em análises posteriores. Falhas como dados ausentes ou faltantes, conhecidos como *missing data*, também devem ser verificados nesta etapa.

Segundo Crespo (2009, p. 5), "a crítica é externa quando visa às causas dos erros por parte do informante, por distração ou má interpretação das perguntas que lhe foram feitas". Ainda, segundo o mesmo autor, uma crítica é considerada "interna quando visa observar os elementos originais dos dados da coleta" (CRESPO, 2009, p. 5).

1.1.3 Apuração dos dados

Nesta etapa, os dados são processados segundo algum critério de classificação por operações matemáticas e estatísticas realizadas em programas computacionais, corriqueiramente, para obtenção de resultados. É nesta etapa que, de fato, se realiza a aplicação dos métodos de análise de dados (BUSSAB; MORETTIN, 2017).

1.1.4 Apresentação dos resultados

Nesta etapa os dados devem ser apresentados, após o devido tratamento, em tabelas e gráficos apropriados, com o propósito de tornar fácil a sua apresentação e a análise dos resultados encontrados (BUSSAB; MORETTIN, 2017).

1.1.5 Análise dos resultados

Análise de parte da população (amostra) para tirar conclusões para o todo (população) realizada a partir dos resultados encontrados nas etapas 3 e 4. A partir disto, busca-se alcançar conclusões sobre o todo (BUSSAB; MORETTIN, 2017). Para exemplificar, considere as pesquisas de intenção de voto realizadas pelos institutos de pesquisa. Os dados

coletados são obtidos por amostragem e seus resultados são aplicados à população geral.

Diante dos passos apresentados, pode-se concluir que, basicamente, a estatística tem interesse em métodos científicos associados com a coleta, a organização, o resumo, a apresentação e a análise dos dados manipulados e relacionados com o objetivo de pesquisa, assim como a obtenção de conclusões para a tomada de decisões.

Quando se consegue examinar todos os elementos ou indivíduos de um grupo de interesse, está sendo realizada uma análise populacional, ou seja, estão sendo manipulados dados de toda a população-alvo ou de interesse, ou ainda, o universo. No entanto, sabe-se que isso é quase sempre impossível de se realizar.

Na impossibilidade de manipulação de dados de uma população, faz-se um exame em um subconjunto da população alvo, ou seja, em uma amostra, a qual deve ser obtida segundo regras rígidas metodológicas, com o intuito de garantir a representatividade populacional, ou seja, garantir que os resultados encontrados a partir da amostra tratada possam ser estendidos para a população de onde ela foi extraída.

Toda pesquisa desenvolvida, em alguma etapa de sua execução, depara-se com a necessidade de análise e de entendimento dos dados resultantes da coleta das informações sobre o objeto de estudo. Diante desta situação, o pesquisador necessitará trabalhar os dados com a intenção de transformá-los em informações, com o intuito de realizar comparações com outros resultados, ou, para avaliar a sua adequação com alguma teoria ou hipótese (BUSSAB; MORETTIN, 2017).

De forma geral, pode-se dizer que a essência da ciência é a observação e que o seu objetivo geral é a inferência de suas análises e manipulações. Entende-se por inferência como a inferência estatística, a qual faz parte da estatística como um todo. A estatística, por sua vez, faz parte da

metodologia científica, a qual tem como propósito, a coleta, redução, análise e modelagem de dados.

Para dar continuidade ao estudo introdutório da estatística, faz-se necessária a apresentação de alguns conceitos básicos e fundamentais associados a ela, os quais são imprescindíveis para a continuidade de seu estudo (SIQUEIRA; TIBÚRCIO, 2011).

1. População ou universo: conjunto de elementos (pessoas, objeto ou até mesmo, conjunto de valores) que tenham alguma(s) característica(s) em comum ao objeto do estudo.

2. População-alvo: população de interesse, ou que se pretende atingir, para se chegar a alguma conclusão documentada nos objetivos da pesquisa.

3. Amostra: subconjunto ou subcoleção da população, ou seja, é uma parte dos elementos pertencentes à população, o qual é extraído para estudo.

4. Variável: característica de interesse a ser analisada. Existe uma classificação para os tipos de variáveis, a qual deve ser aplicada sempre no planejamento da pesquisa, pois a técnica estatística a ser utilizada dependerá diretamente desta informação.

5. Variável qualitativa ou categórica: informação que pode ser classificada como um atributo ou característica. É uma informação não numérica que pode ser separada em diferentes categorias. Podem ser classificadas em nominal e ordinal.

6. Variável qualitativa nominal: caracterizada por dados que consistem em rótulos, não podendo ser dispostos segundo algum tipo de ordenação, como, por exemplo, o sexo (F = feminino e M = masculino).

7. Variável qualitativa ordinal: informação não numérica que pode apresentar algum tipo de ordenação, como, por exemplo, a escolaridade (Fundamental, Médio e Superior).

8. Variável quantitativa: informação numérica que representa contagens ou medidas. Pode ser dividida em duas categorias: discreta e contínua.

9. Variável quantitativa discreta: informação numérica resultante de um conjunto finito de valores possíveis, ou de um conjunto que pode ter seus valores enumerados, como, por exemplo, o número de filhos de uma família (0, 1, 2, ...).

10. Variável quantitativa contínua: informação numérica resultante de um conjunto infinito de valores possíveis, em geral, oriunda de mensuração dentro de um intervalo de valores. Um exemplo deste tipo de variável é o índice de massa corporal (IMC), o peso corporal, a altura, a distância entre duas cidades, etc.

11. Variável quantitativa contínua com nível de mensuração intervalar: Variável originária de uma mensuração onde as operações matemáticas de soma (+) e de subtração (–) podem ser realizadas, mas as operações de multiplicação (×) e divisão (÷) não são apropriadas. Um exemplo deste tipo de variável é a temperatura ambiente, a qual quando registra 0 ºC não significa ausência de temperatura, ou quando se compara uma temperatura de 10ºC com 30ºC se pode afirmar que existe diferença de 20ºC, no entanto, não se pode afirmar que 30ºC é três vezes mais quente que 10ºC. Não existe um zero absoluto em temperatura medida em graus Celsius (SIQUEIRA; TIBÚRCIO, 2011).

12. Variável quantitativa contínua com nível de mensuração razão: variável originária de uma mensuração onde é possível realizar as quatro operações matemáticas básicas: soma, subtração, multiplicação e divisão. Para este tipo de variável, é possível

identificar o zero absoluto. Por exemplo, a idade fornecida em anos, pode-se dizer que uma pessoa que tem 30 anos já viveu 20 anos a mais do que uma pessoa que tem 10 anos. Ainda é possível afirmar que a pessoa de 30 anos tem o triplo de idade da pessoa que tem 10 anos (SIQUEIRA; TIBÚRCIO, 2011).

13. Parâmetro: característica que descreve uma população, como, por exemplo, a média populacional ou a proporção populacional.

14. Estimativa ou estatística: característica que descreve uma amostra, como, por exemplo, a média amostral ou a proporção amostral.

15. Amostragem: método científico elaborado para a obtenção de amostras. Existem vários métodos de amostragem e os principais são: amostragem aleatória simples, amostragem estratificada, amostragem sistemática, amostragem por conglomerados e a combinação destes tipos.

O planejamento de um estudo está associado com a maneira de sua organização. Independente da área de pesquisa, um planejamento de estudos pode ser descritivo ou comparativo.

Os estudos “descritivos descrevem uma determinada situação, sem preocupação ou possibilidade de comparação e, como o nome indica, as conclusões são meramente descritivas” (SIQUEIRA e TIBÚRCIO, 2011, p. 5).

Para a realização de um estudo comparativo, há a necessidade da existência de um grupo de comparação e, ainda dentro dos estudos comparativos, é possível fazer a distinção entre estudos observacionais, os quais são feitos com dados produzidos pelo histórico dos sujeitos de análise e, os estudos experimentais, onde há uma intervenção feita pelo pesquisador.

Em se tratando de análise de dados, é possível classificar a análise estatística em clássica, também conhecida como frequentista, e

estatística bayesiana. O termo bayesiano é originário do teorema de Bayes, criado por Thomas Bayes (1702-1761), o qual foi um reverendo presbiteriano inglês e matemático (SIQUEIRA; TIBÚRCIO, 2011). Para a realização de uma boa estratégia de análise de dados, sugere-se que ela seja separada em duas etapas: a análise preliminar e a análise definitiva.

Segundo Siqueira e Tibúrcio (2011), para a realização de uma análise preliminar, faz-se necessária a realização das seguintes etapas: a) processamento dos dados de uma forma conveniente para análises posteriores; b) verificação da qualidade dos dados através de uma crítica detalhada, como a verificação de existência de erros, observações atípicas, ausência de dados, entre outros. Ainda na etapa de crítica dos dados, também é importante verificar se os dados necessitam de algum tipo de modificação, como uma transformação, por exemplo; c) fazer uma análise descritiva dos dados.

Ainda segundo Siqueira e Tibúrcio (2011, p. 6), para a realização de uma análise definitiva, deve-se "ter em mente que a escolha do procedimento apropriado depende de vários fatores". Dentre estes fatores, as autoras descrevem a importância de se realizar os seguintes questionamentos:

> Existe alguma publicação sobre o assunto? Você, alguém que conhece ou algum centro de pesquisa já enfrentou problema parecido? Há alguma informação a priori? É possível reformular o problema de maneira que o torne mais simples para ser resolvido? É possível dividir os problemas em partes disjuntas e resolver cada uma delas por vez? Quais são os objetivos do estudo? Qual é a estrutura dos dados? Houve pareamento? (SIQUEIRA e TIBÚRCIO, 2011, p. 6-7)

Depois da coleta, tendo à disposição os dados brutos, estes devem ser organizados de maneira apropriada em um banco de dados para serem trabalhados posteriormente, com o intuito de consolidação dos mesmos.

De forma geral, as etapas de análise de dados podem ser divididas em duas: análise descritiva de dados, por onde sempre se deve começar, e a inferência estatística.

Uma análise descritiva de dados pode ser compreendida como a análise realizada após a produção de tabelas, gráficos e medidas descritivas como média, variância, mediana, coeficiente de variação, etc. Estas medidas, ainda, também são chamadas de estatísticas, as quais podem ser definidas como função (em termos matemáticos) da amostra.

De maneira geral, não é suficiente apenas a descrição de um conjunto de dados, o interessante é obter conclusões que possam ser extrapoladas para a população de onde os dados foram retirados, pois resultados que se limitam a valer apenas para a amostra são, de certa forma, bastante limitados. A inferência estatística é apropriada para retirar conclusões para a população da qual a amostra foi retirada (SIQUEIRA e TIBÚRCIO, 2011). Os procedimentos básicos da inferência estatística são: estimação pontual, estimação por intervalo (intervalo de confiança – IC) e teste de hipóteses.

Durante a exploração de uma amostra é importante fazer uso de diferentes análises e escolher aquela que apresenta a melhor descrição dos dados, além de atender aos pressupostos das técnicas estatísticas utilizadas.

Como etapa final, Siqueira e Tibúrcio (2011, p. 7) afirmam que, “após tirar conclusões e discutir os resultados, o próximo passo é preparar o relatório final, expondo toda a pesquisa de forma concisa e precisa”.

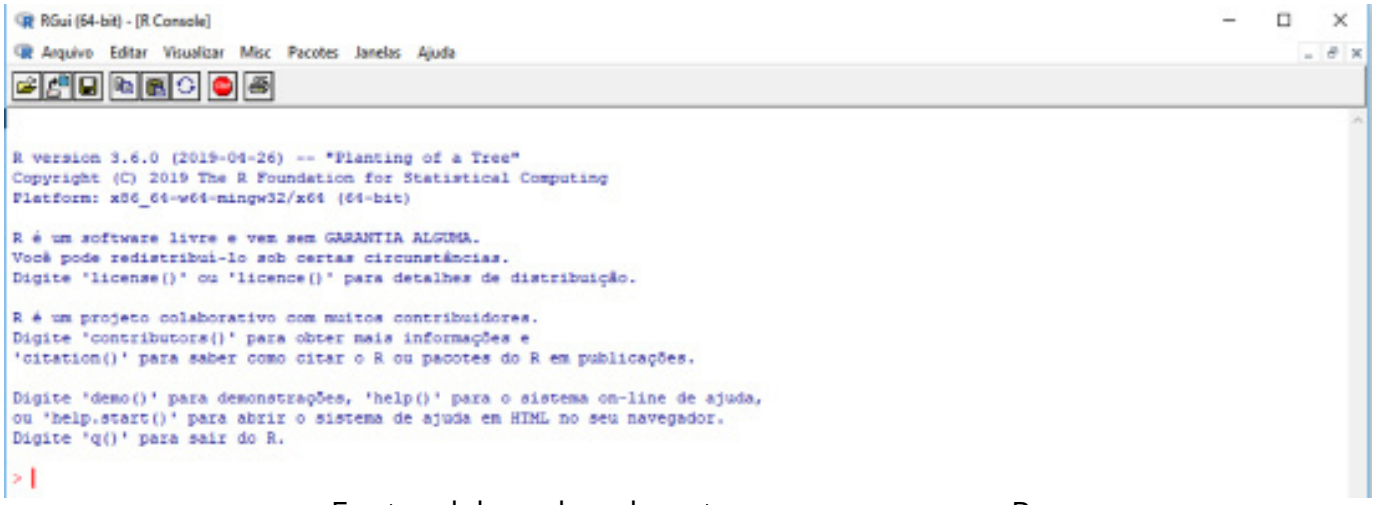
2. Linguagem R

O programa R é um *software* estatístico gratuito e de código aberto (*open source*). Ele é resultante de um projeto elaborado por diversos

colaboradores ao redor do mundo, composto por implementações para análises estatísticas simples até as mais complexas (SIQUEIRA. TIBÚRCIO, 2011).

A linguagem R é semelhante à linguagem S desenvolvida no Bell Laboratories (antiga AT&T e, agora, Lucent Technologies) por John Chambers e colaboradores. A linguagem R pode ser considerada com uma implementação diferenciada da linguagem S, pois existem diferenças significativas entre elas, no entanto, o código escrito para S não se alterou na mudança para a linguagem R. A Figura 2 mostra a tela de interface do programa R.

Figura 2 – Tela de interface do programa R



Fonte: elaborada pelo autor com o programa R.

Existe uma grande quantidade de procedimentos estatísticos de pacotes computacionais livremente disponíveis na internet para a linguagem R, os quais podem ser carregados opcionalmente. Toda função digitada no R deve aparecer entre ( ), onde são inseridos um ou mais argumentos.

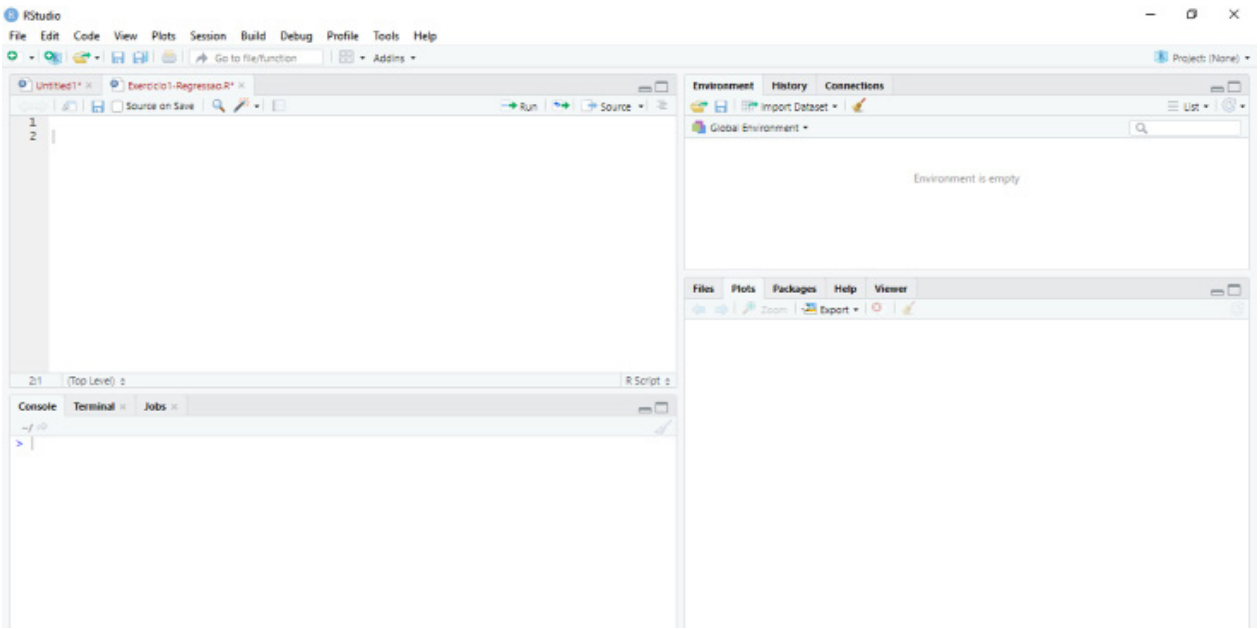
Para realizar o *download* do R é possível seguir as instruções apresentadas no material de Santana (2017), o qual apresenta um passo a passo de como efetuar a instalação do programa. No mesmo material, é possível seguir as instruções para realizar o *download* de uma interface para uso do R conhecida como RStudio, a qual é uma interface funcional e muito mais amigável para uso do R, que, também, é conhecida como ambiente

de desenvolvimento integrado (IDE). Assim como para o R, o material também apresenta um passo a passo para a instalação e para o uso do RStudio. Existe uma versão gratuita, mas, também, existem versões pagas, as quais possuem recursos adicionais que facilitam a usabilidade do usuário. O uso do RStudio não é obrigatório. Portanto, o usuário poderá utilizar apenas o programa R original para realizar suas análises estatísticas. Uma interface padrão do RStudio é apresentada na Figura 2. Atente-se para escolher a versão apropriada para a instalação de seu uso.

O RStudio não é a única interface existente para o R. No entanto, é uma das mais conhecidas e utilizadas. Algumas dessas interfaces funcionam apenas em sistema operacional Linux, outras apenas em Windows ou Mac OS, mas existem aquelas que funcionam em qualquer sistema operacional.

O R será utilizado, neste curso, para demonstrar a utilização de funções estatísticas já implementadas no *software* e, também, para desenvolver a habilidade do aluno com exemplos aplicados com a metodologia estatística apresentada em paralelo.

Figura 3 – Interface padrão do RStudio



Fonte: elaborada pelo autor com a IDE RStudio.

Este texto apresentou um contexto histórico da estatística e apresentou alguns de seus conceitos fundamentais. Também apresentou um pouco do histórico do programa R, um programa gratuito e de código aberto (*open source*), assim como apresentou uma das interfaces existentes e bastante utilizada, o RStudio.

TEORIA EM PRÁTICA

Você trabalha em uma empresa de pesquisa de mercado e faz uso de ferramentas de *business intelligence* para manipular dados. Sua responsabilidade é realizar o *analytics* e apresentar resultados para os clientes de sua empresa.

A demanda de trabalho na sua empresa tem crescido bastante, então você sugeriu ao seu superior aumentar a sua equipe de trabalho em mais dois colaboradores. Após analisarem seu pedido e terem aceitado sua proposta, você ficou responsável por treinar os dois novos colaboradores com conhecimentos básicos de estatística e de linguagem R, que é a ferramenta de manipulação de dados utilizada por você.

Então, você vai elaborar um material para treinamento de seus novos colaboradores. Esse material deve conter conceitos básicos de estatística, como definições gerais de estatística, tipos de variáveis, principais métodos de análise de dados, etc.

Nesse mesmo material você precisa apresentar a linguagem R, suas funcionalidades, sua IDE RStudio e alguns comandos básicos. Esse material precisa ser um material muito didático para facilitar a aprendizagem dos novos colegas de trabalho.

Com esse desafio em mãos, como você pretende fazer? Por onde pretende iniciar? Tem a intenção de elaborar um material curto, médio ou longo?

VERIFICAÇÃO DE LEITURA

1. Medeiros (2009) descreve a história da estatística em etapas que marcam caracteristicamente cada uma delas. Em quantas etapas ele apresenta esse período histórico? Assinale a alternativa correta.

   a. Duas etapas.

   b. Três etapas.

   c. Quatro etapas.

   d. Cinco etapas.

   e. Seis etapas.

2. Um dos conceitos fundamentais em estatística é aquele que se refere ao conceito de variável. Inicialmente, pode-se dizer que existem dois tipos de variáveis. Quais são esses tipos? Assinale a alternativa correta.

   a. Nominal e razão.

   b. Qualitativa e quantitativa.

   c. Qualitativa e nominal.

   d. Quantitativa e contínua.

   e. Intervalar e razão.

3. A verificação da qualidade dos dados de uma pesquisa faz parte de qual etapa de uma análise estatística? Assinale a alternativa correta.

   a. Análise definitiva.

   b. Amostragem.

   c. Coleta de dados.

   d. Análise preliminar.

   e. Análise de tabelas.

Referências bibliográficas

AQUINO, Jakson Alves de. **R para cientistas sociais** Ilhés: UESC. 2014. Disponível em: http://www.uesc.br/editora/livrosdigitais_20140513/r_cientistas.pdf. Acesso em: 9 jun. 2019.

BUSSAB, Wilton.; MORETTIN, Pedro A. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017. 554p.

CRESPO, Antonio Arnot. **Estatística fácil.** 19. ed. São Paulo: Saraiva, 2009.

MEDEIROS, Carlos Augusto de. **Estatística aplicada à educação.** Brasília: Universidade de Brasília, 2009. 136p. Disponível em: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=598-estatistica-aplicada-a-educacao&Itemid=30192. Acesso em: 1 jun. 2019.

SANTANA, Verônica. **Tutorial R/RStudio.** São Paulo. FEA-USP. 2017. Disponível em: https://edisciplinas.usp.br/pluginfile.php/2996937/mod_resource/content/1/Tutorial.pdf. Acesso em: 9 jun. 2019.

SIQUEIRA, Arminda. L., TIBÚRCIO, Jacqueline. D. **Estatística na área da saúde**: conceitos, metodologia, aplicações e prática computacional. Belo Horizonte: Coopmed, 2011. 520 p.

THE R FOUNDATION.**The R Project for Statistical Computing**. Disponível em: https://www.r-project.org/. Acesso em: 9 jun. 2019.

TRIOLA, Mario F. **Introdução à estatística.** 10. ed. Brasil: LTC (Grupo GEN), 2008. 696 p.

Gabarito

Questão 1 – Resposta: D

Resolução: Medeiros (2009) apresentou a história da estatística em quatro fases principais, descrevendo as características marcantes em cada uma delas.

Feedback de reforço: Lembre-se do que foi apresentado no Quadro 1 da Leitura Fundamental. Avalie os fatos históricos ali apresentados.

Questão 2 – Resposta: B

Resolução: O conceito de variável em estatística divide-se em dois tipos: qualitativa e quantitativa.

Feedback de reforço: Lembre-se das diferenças entre os tipos de dados. São quatro os tipos existentes.

Questão 3 – Resposta: D

Resolução: A verificação da qualidade dos dados faz parte da análise preliminar de dados.

Feedback de reforço: Lembre-se que existem duas etapas principais para uma análise de dados.

Estatística: média, mediana e moda. Desvio padrão e conceitos de amostra. Intervalos de confiança, hipóteses, correlação, causalidade e regressões lineares

Autor: Marcelo Tavares de Lima

Objetivos

- Apresentar conceitos de estatística descritiva.
- Apresentar conceitos de medidas de dispersão.
- Apresentar conceitos fundamentais de amostragem.
- Apresentar conceitos fundamentais de inferência estatística.
- Apresentar conceitos e medidas fundamentais de correlação, causalidade e regressão linear.

1. Introdução

Toda análise de dados deve ser iniciada com uma descrição detalhada dos dados juntamente com uma exploração inicial. O intuito é conhecer os padrões existentes. Um conjunto de métodos estatísticos são utilizados para a realização desta etapa, os quais são conhecidos como estatística descritiva.

Uma análise de dados mais profunda requer métodos indutivos para a validação de hipóteses de pesquisa, por exemplo. Os métodos estatísticos apropriados para esta fase são conhecidos como inferência estatística ou métodos inferenciais. Fazem parte deste conjunto de ferramentas as estimativas pontuais ou por intervalo, os testes de hipóteses e os modelos de regressão.

Este texto abordará ambos tipos de análise, com a apresentação de suas principais medidas estatísticas e, como calculá-las e interpretá-las.

2. Estatística descritiva e amostragem

Segundo Siqueira e Tibúrcio (2011, p. 52), “o uso de técnicas descritivas deve sempre preceder análises mais avançadas. A afirmação das autoras é impositiva porque é uma etapa da análise que deve ser cumprida exatamente no momento preliminar do trabalho de manipulação dos dados.

A análise descritiva proporciona a familiarização com os dados; permite detectar estruturas, como, por exemplo, a distribuição dos dados, a existência de valores atípicos ou mesmo incorretos.

Neste item serão apresentadas medidas descritivas de tendência central e medidas de dispersão, além de apresentar os principais métodos de obtenção de amostra, chamados de técnicas de amostragem.

2.1 Estatística descritiva

As medidas descritivas são importantes para produzir uma visão global dos dados (SIQUEIRA; TIBÚRCIO, 2011). Tais medidas recebem o nome genérico de estatísticas e se dividem em medidas de tendência central e medidas de dispersão. Um estudo inicial das medidas descritivas se dá com o uso de uma única variável por questões puramente didáticas. No entanto, é possível realizar análise descritiva com duas ou mais variáveis. Quando se trabalha com duas variáveis, a análise recebe o nome de bivariada. Quando se trabalha com três ou mais variáveis, a análise se chama multivariada.

Siqueira e Tibúrcio (2011, p. 58) definem dados brutos como "aqueles obtidos diretamente da pesquisa, isto é, que ainda não sofreram qualquer processo de síntese, consolidação ou análise". É com este tipo de dado que serão obtidas todas as medidas descritivas e inferenciais para a realização de uma análise de dados. Vale lembrar que o conjunto de dados é o que constitui uma amostra, o tamanho amostral tem como notação genérica a letra n.

As medidas descritivas que compõem as medidas de tendência central "resumem o conjunto de dados em um único número" (SIQUEIRA; TIBÚRCIO, 2011, p. 80). São medidas que buscam um valor que possa representar bem a distribuição de um conjunto de dados.

A medida de tendência central mais conhecida é a média aritmética, a qual é uma boa medida para representar conjuntos de dados que possuem distribuição simétrica.

Se a distribuição do conjunto de dados não for simétrica, a média aritmética deixa de ser uma boa representação, dando lugar a uma outra medida conhecida como mediana, ou até mesmo um percentil pode representar melhor os dados do que a média, nesse caso. No caso de distribuições simétricas, a média aritmética e a mediana devem ser aproximadamente iguais.

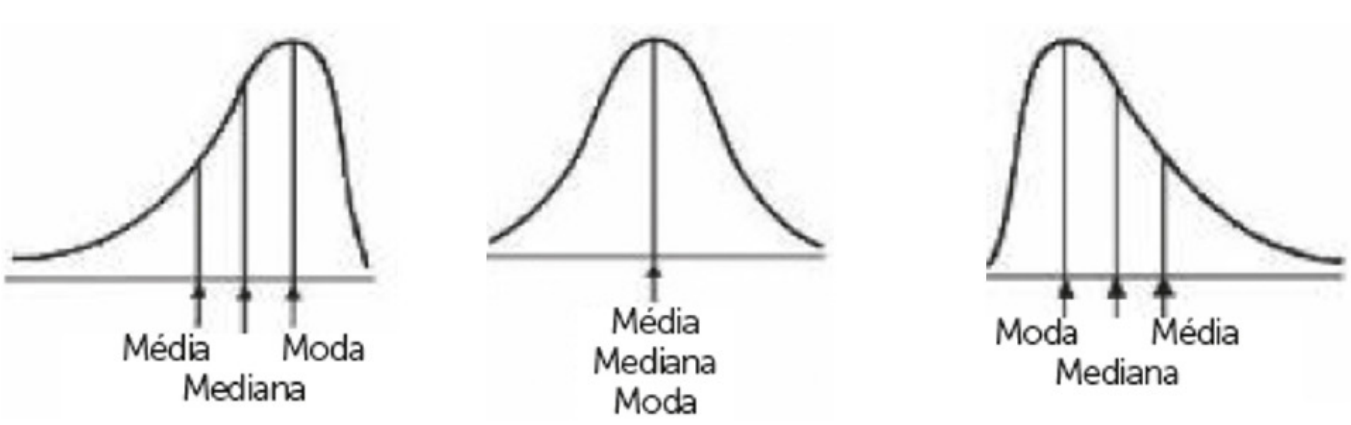
Consideremos que desejamos calcular a média aritmética de um conjunto de dados com n observações ($x_1, x_2,...,x_n$). Iremos representar a média de um conjunto de dados por $\bar{x}$ (lê-se x barra).

É possível que você saiba, no entanto, para formalizar o cálculo, somam-se todos os valores $\left(x_1, x_2, ..., x_n = \sum_{i=1}^{n} x_i\right)$ e, em seguida, divide-se pelo total de observações (n). Então, a fórmula será dada por:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

A Figura 1 apresenta a relação existente entre as principais medidas de tendência central, a média, mediana e a moda.

Figura 1 – Relação entre a média aritimética simples, mediana e moda



Fonte: RIBEIRO (2015).

De maneira geral, uma medida de tendência central não é o suficiente para descrever ou representar um conjunto de dados de forma satisfatória (SIQUEIRA; TIBÚRCIO, 2011).

Não é o suficiente saber o valor em torno do qual os dados estão concentrados. É necessário, também, conhecer o grau em que estão agregados. Em outras palavras, é preciso definir alguma medida da dispersão do conjunto de dados.

Existe um conjunto de medidas que mede a dispersão de um conjunto de dados. Esse conjunto é conhecido por medidas de dispersão e compõem, também, o conjunto de medidas descritivas, juntamente com as medidas de tendência central.

Assim como as medidas de tendência central, existem várias estatísticas que compõem o conjunto de medidas de dispersão. No entanto, este texto abordará apenas a variância e o desvio, que são as mais utilizadas para descrever conjuntos de dados.

"A variância é uma medida da variabilidade dos dados em torno da média" (SIQUEIRA; TIBÚRCIO, 2011, p. 84). Ela representa a média dos desvios ao quadrado das observações tendo como referência a média aritmética.

A variância de um conjunto de dados, denotada por s^2, pode ser representada matematicamente como:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} \qquad (2)$$

O desvio padrão, denotado por s_1 é a raiz quadrada positiva da variância. Essa operação é realizada para manter essa medida com a mesma unidade de medida que a média aritmética, pois, se não fizer isso, a medida ficaria alterada para a unidade original ao quadrado. Então, o desvio padrão pode ser calculado por:

$$s = \sqrt{s^2} \qquad (3)$$

Manualmente, para obter o desvio padrão, deve-se calcular primeiro a variância. No entanto, em um computador isso não é necessário. Pode-se calcular o desvio padrão diretamente.

PARA SABER MAIS

Outras medidas de tendência central e de variabilidade: além das médias e da mediana, também compõem o conjunto de medidas de tendência central, como a moda que representa o valor que mais se repete num conjunto de dados. Também existem outras medidas de variabilidade, como a amplitude e o coeficiente de variação.

ASSIMILE

Percentil: "É o valor que acumula determinado percentual" (SIQUEIRA; TIBÚRCIO, 2011, p. 91). De uma maneira geral, o percentil de uma determinada ordem θ, representado por P_{θ}, é o valor que é antecedido ($\geq$) por θ / 100 dos valores e sucedido ($\leq$) por [100 - θ] / 100. Por exemplo, o percentil de ordem θ / 25 é antecedido por 25 / 100 = 1 / 4 dos dados, depois de ordenados e, sucedido por [100 - 25] / 100 = 75 / 100 = 3/4 dos dados.

2.2 Amostragem

Amostragem é a coleta parcial da uma população-alvo sobre a qual se deseja estudar alguma(s) característica(s). Ela se justifica por vários fatores, porém os principais são a inacessibilidade da população estudada, o custo e o tempo disponível para a realização da pesquisa.

Os procedimentos de amostragem foram criados para garantir a representatividade de seus resultados em relação à população de onde foram extraídas amostras, ou seja, os resultados encontrados deverão ser um reflexo da população original (SIQUEIRA; TIBÚRCIO, 2011).

Antes de iniciar a coleta dos dados para compor a amostra é preciso deixar muito bem definida a unidade de análise ou unidade amostral. Pois, dependendo do tipo, alguns procedimentos de coleta serão melhores que outros.

Os procedimentos de amostragem se dividem em duas classes: probabilísticos e não probabilísticos. A amostragem probabilística atribui probabilidade de seleção aos elementos da população. Em geral, a probabilidade de seleção para compor a amostra é a mesma para todos os elementos. A amostragem não probabilística não atribui medida probabilística aos elementos da população. A seleção das unidades amostrais é feita por outros critérios, por exemplo, arbitrariedade, intenção, qualquer outra característica da pesquisa, etc. A escolha do tipo de amostragem a ser feita vai depender dos objetivos da pesquisa, dentre outras características da mesma.

Os tipos de amostragem probabilística e não probabilísticas ainda se subdividem da seguinte maneira: Amostragem probabilística pode ser aleatória simples; estratificada; conglomerados ou sistemática. Amostragem não probabilística pode ser acidental, intencional ou por cotas.

A principal vantagem de uma pesquisa feita por amostragem se dá por conta dos fatores aqui já mencionados: desnecessidade de acesso a toda a população, menor custo para a pesquisa e menor tempo de coleta. Mesmo assim, quando bem planejada, a amostra será uma cópia fidedigna da população de onde foi extraída.

A amostragem probabilística é um método eficiente para coletar amostras que reflitam, de forma correta, a variabilidade existente na(s) característica(s) de interesse, na população-alvo. O processo de definição de amostras probabilísticas está baseado em probabilidade de seleção dos seus componentes, onde cada um deles tem igual chance de ser selecionado (BUSSAB; MORETTIN, 2017).

As amostras construídas a partir de um procedimento probabilístico não necessitam ser representativas em todos os aspectos. Isto deve ocorrer apenas para as características que serão estudadas.

2.2.1 Amostragem aleatória simples

Técnica de amostragem probabilística mais básica e mais utilizada, em geral. A matemática por detrás deste método é particularmente complexa. Por isso, não vamos nos deter em muitos desses detalhes, pois não é o propósito deste texto.

O procedimento se baseia, de forma simplista, pela enumeração dos elementos da população, onde cada um recebe um número exclusivo e, em seguida, são sorteados aqueles que irão compor a amostra. No método de amostragem aleatória simples, todos os componentes da população têm igual chance de fazer parte da amostra. O mecanismo utilizado para sorteio pode ser uma tabela de números aleatórios ou um programa computacional apropriado.

Em situações onde se deseja determinar o tamanho da amostra, é necessário considerar uma possibilidade de errar com certa probabilidade. Tal erro, é chamado erro de estimação, o qual deve ser previamente determinado pelo responsável da pesquisa.

2.2.2 Amostragem estratificada

Técnica utilizada quando se deseja reduzir a variabilidade de alguma característica de interesse, dividindo a população em grupos (estratos) mais homogêneos com respeito a esta e, heterogêneos entre si. Por exemplo, uma pesquisa com estudantes universitários pode estratificá-los por turmas, para obter de cada uma delas as informações de interesse (SIQUEIRA; TIBÚRCIO, 2011).

Portanto, a técnica de amostragem estratificada só é cabível quando, no que se refere à variável de interesse, existirem diferenças significativas entre os elementos da população.

É possível fazer a estratificação por mais de uma característica, como, por exemplo, a estratificação feita com os estudantes universitários por turma, também pode ser feita por gênero, por notas em estatística, etc.

Para uma boa estratificação, é importante ter informações anteriores para a sua aplicação. Como dito anteriormente, é importante estar com informações relevantes com respeito à variabilidade populacional. Pois só assim estratificar pode ser uma saída para o problema, com a obtenção de uma amostra menor, comparativamente, se tivesse sido utilizada uma amostragem aleatória simples.

2.2.3 Amostragem sistemática

Uma lista ordenada de elementos pode ser mais útil do que uma lista não ordenada apresentada de forma aleatória. Isso reforça que nem sempre uma lista deve ser aleatorizada antes da realização de uma amostragem sistemática.

Este procedimento é útil quando se tem o intuito de selecionar, dentro de uma população de tamanho N, um número *n* de elementos a partir de um intervalo K, sendo este, um intervalo determinado pelo tamanho da população e pelo tamanho da amostra. Assim, temos que $K = N / n$. Portanto, a componente K é um intervalo de seleção sistemática. O tamanho amostral n é definido pela técnica de amostragem aleatória simples. Para exemplificar, suponha que se deseja selecionar itens de produção de uma linha de produção para avaliar se o processo está dentro do esperado, ou seja, se o produto está dentro dos padrões definidos. Decidiu-se selecionar os itens a serem inspecionados por amostragem sistemática, podendo ser selecionado um item a cada vinte fabricados.

2.2.4 Amostragem por conglomerados

O conjunto de elementos pertencentes a uma população é chamado de conglomerado. Se a intenção de compor uma amostra forem conglomerados, diz-se que a amostragem realizada é por conglomerados.

Uma justificativa para o seu uso está baseada na limitação por recursos financeiros, tempo, distâncias geográficas ou por combinações desses e outros motivos.

Neste método amostral, cada conglomerado é visto como uma espécie de miniatura da população, o que, intuitivamente, faz com que a ocorrência de heterogeneidade dentro de cada um seja desejável. Isto garante a representatividade da população. Seu uso é importante quando não é possível obter unidades amostrais mais simples, como uma pessoa (SIQUEIRA; TIBÚRCIO, 2011).

Um caso particular de amostragem por conglomerado é a amostragem por áreas, quando estas são divididas por área ocupada por uma população em estudo, sendo cada parte uma unidade amostral ou conglomerado.

Muitos institutos de pesquisa fazem uso do procedimento amostral por conglomerados, onde as unidades amostrais são quarteirões, bairros, distritos, setores censitários de uma cidade, etc.

3. Inferência estatística

“Inferência estatística é a metodologia estatística que possibilita, a partir de dados amostrais, fazer generalizações ou inferências sobre uma população, sempre com medida de precisão sobre sua veracidade (SIQUEIRA e TIBÚRCIO, 2011, p. 236).

A inferência estatística se divide, basicamente, em dois procedimentos: teste de hipóteses, também conhecido como teste de significância, e estimação, que se subdivide em pontual e por intervalo. Este último tipo é conhecido como intervalo de confiança (IC) para o parâmetro de interesse, como, por exemplo, média, mediana, desvio padrão, proporção, etc.

Um parâmetro, segundo Siqueira e Tibúrcio (2011, p. 236), "é um valor que descreve alguma característica da população". No entanto, em geral, é desconhecido e, por isso, torna-se necessário estimá-lo a partir de uma amostra.

Testar hipóteses trata-se de um procedimento de tomada de decisão. Como, por exemplo, é importante decidir se existe alguma diferença entre tratamentos que estejam sendo comparados.

No entanto, não basta apenas saber se os tratamentos diferem entre si, é preciso saber, também, o quanto eles diferem entre si, daí o procedimento de estimação. A estimação é um procedimento que determina uma previsão de um parâmetro populacional (desconhecido), baseado em informações contidas em amostras.

3.1 Teste de hipóteses

Agora será tratado um conceito importante da inferência estatística, conhecido como teste de hipóteses ou teste de significância. Basicamente, essa metodologia trata de procedimentos que podem ser utilizados para constatar ou refutar hipóteses de pesquisa através de amostras de dados. É um procedimento de tomada de decisão, como dito antes.

Segundo Pereira (2015, p. 154): "Em 1933, Jerzy Neyman (1894-1981) e Egon Pearson (1895-1980), ambos ex-alunos de Karl Pearson (1857-1936), pai do segundo, conceberam o que chamaram de teste de hipótese". Considere a tomada de decisão em concluir se há ou não diferença entre grupos de comparação.

Na prática, faz-se necessária a quantificação do efeito da intervenção e não apenas dizer se existe ou não diferença entre eles. É quando surge o procedimento de estimação. Este é um procedimento que permite fazer previsão de algum parâmetro populacional (desconhecido) baseado em informações contidas em amostras.

Alguns conceitos fundamentais necessários para o estudo de teste de hipóteses serão apresentados a seguir.

- **Hipótese nula (H_0):** é a hipótese que será testada pelo pesquisador. Por exemplo, na comparação entre dois tratamentos médicos, deseja-se provar que não há diferença entre eles.
- **Hipótese alternativa (H_1):** é a hipótese do problema a ser investigado, como, por exemplo, no caso dos dois parâmetros comparados, seria a hipótese de inexistência de igualdade entre eles.
- **Critério de decisão:** é baseado na estatística do teste após as hipóteses a serem testadas estarem definidas. "A estatística do teste mede a discrepância entre o que foi observado na amostra e o que seria esperado se a hipótese nula fosse verdadeira" (SIQUEIRA; TIBÚRCIO, 2011, p. 239).
- **Erro tipo I:** é a decisão em rejeitar a hipótese nula (H_0) quando de fato ela é verdadeira. Ou seja, pelo critério de decisão, os parâmetros testados são estatisticamente diferentes a partir da amostra utilizada, no entanto, sabe-se que isso não é para acontecer. Na literatura, a probabilidade de cometer o erro tipo I é conhecida como nível de significância, geralmente representado pela letra grega α (lê-se: alfa) (SIQUEIRA; TIBÚRCIO, 2011).
- **Erro tipo II**: é a decisão em não rejeitar a hipótese nula (H_0) quando de fato ela é falsa. Ou seja, é a aceitação da hipótese de que os parâmetros testados são idênticos, sabendo-se que essa hipótese é falsa. A probabilidade de cometer o erro do tipo II é representada pela letra β (lê-se beta).

- **Poder do teste:** é a capacidade de um teste identificar diferenças que realmente existem, em outras palavras, é a capacidade de rejeitar H_0 quando ela é realmente falsa. Ou seja, é a capacidade do resultado do teste representar a realidade, acertando a partir de uma amostra. Em termos de probabilidade, é definido como $(1 - \beta)$.

- **Probabilidade de significância**, nível descritivo, valor-p: é a probabilidade de ocorrência de valores iguais ou superiores ao assumido pela estatística do teste, sob a hipótese de que H_0 é verdadeira. Quanto mais baixo o valor-p, mais evidências para se rejeitar H_0. Ou seja, é a probabilidade de encontrar resultados semelhantes aos de H_0.

- **Hipóteses unilaterais e bilaterais**: baseiam como a hipótese alternativa H_1 é construída, que pode ser de três maneiras possíveis, segundo Pereira (2015), considerando-se a comparação entre dois grupos ou tratamentos:

  1. O parâmetro testado é diferente entre os grupos comparados (bilateral).
  2. O parâmetro no grupo de interesse é maior que no grupo de referência (unilateral).
  3. O parâmetro no grupo de interesse é menor que no grupo de referência (unilateral).

O Quadro 1 resume essas possibilidades, considerando hipóteses para a média populacional, representada pela letra grega (lê-se: mi) como parâmetro de teste.

Quadro 1 – Modalidades de teste de hipótese

Hipótese nula H_0	Hipótese alternativa H_1	Tipo de teste de hipótese
H_0: $\mu 1 = \mu 2$	H_1: $\mu 1 \neq \mu 2$	Bilateral
H_0: $\mu 1 \leq \mu 2$	$H_1 1$: $\mu 1 > \mu 2$	Unilateral
H_0: $\mu 1 \geq \mu 2$	H_1: $\mu 1 < \mu 2$	Unilateral

Fonte: adaptado de Pereira (2015).

3.2 Intervalos de confiança

Para a realização de uma tomada de decisão, o pesquisador deve se basear no valor do parâmetro de interesse, por exemplo, uma média. No entanto, na prática, o real valor do parâmetro nunca é conhecido. Para atender a essa demanda, desenvolveu-se métodos estatísticos chamados “teoria da estimação” (SIQUEIRA; TIBÚRCIO, 2011).

Para estudar esse campo dos métodos estatísticos, o conhecimento de alguns conceitos torna-se necessário. Segundo Siqueira e Tibúrcio (2011, p. 243), “o estimador é uma estatística (uma fórmula), enquanto a estimativa é um valor particular do estimador”.

Não existe uma notação padrão para um estimador, no entanto, existe uma notação usual para a representação de um estimador. A notação mais utilizada consiste em colocar o sinal de circunflexo “^” sobre o símbolo do estimador, chamado de chapéu. Como exemplo, o estimador de uma proporção populacional é representado por $\hat{P}$ (lê-se p-chapéu). Existe uma exceção, a média amostral é estimada pelo estimador x-barra ($\bar{X}$).

Como primeiro passo, faz-se necessária a identificação do parâmetro que se deseja estimar (média, mediana, desvio-padrão, proporção, risco relativo, *odds ratio*, etc.).

Em seguida, realizar a estimação, a qual pode ser desenvolvida de duas formas: (1) pontual, a qual fornece um único valor como estimativa; (2) intervalar, a qual fornece um intervalo de valores plausíveis para o parâmetro conhecido como intervalo de confiança (IC), obtidos usualmente com confiança de 95%.

A estimação por intervalo agrega ao estimador pontual a informação sobre sua variabilidade através da determinação de um limite inferior e um superior para a estimativa do parâmetro.

Um intervalo de confiança tem amplitude (A) definida como a diferença entre o limite superior e o inferior, onde é desejado que essa medida seja a menor possível. No entanto, no geral, ela depende do tamanho da amostra e da confiança pré-determinada. Quanto maior for a amostra, menor será a amplitude A. Esse resultado reforça a importância de uma boa execução do dimensionamento da amostra.

O coeficiente de confiança, representado pela letra γ (gama), é o complemento do conceito de nível de significância α, ou seja, $\gamma = 1 - \alpha$. Para exemplificar, considere que α = 0,05, então, $\gamma = 1 - 0{,}05 = 0{,}95$, ou seja, a confiança é de 95%.

3.3 Correlação

É uma medida estatística que avalia a existência de relação entre duas ou mais medidas numéricas e sua variação conjunta pode ser visualizada a partir de um diagrama de dispersão.

O coeficiente de correlação de Pearson (r) expressa a quantificação da relação linear entre duas variáveis numéricas, assim como o sentido dessa relação. Suas propriedades podem ser resumidas conforme abaixo (SIQUEIRA; TIBÚRCIO, 2011, p. 105).

- É uma quantidade adimensional, isto é, é um número puro.
- Varia entre -1 e 1.
- É invariante em relação à mudança de escala linear.

O coeficiente de correlação de Pearson é definido pela seguinte expressão:

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}} \quad (4)$$

Uma interpretação do resultado numérico do coeficiente de correlação de Pearson pode ser feita conforme apresentado no Quadro 2.

Quadro 2 – Intepretação do coeficiente de correlação de Pearson

Coeficiente de correlação (*r*)	Interpretação
r = +1	Perfeita positiva
0,8 ≤ r < 1	Forte positiva
0,5 ≤ r < 0,8	Moderada positiva
0,1 ≤ r < 0,5	Fraca positiva
0 < r < 0,1	Ínfima positiva
r = 0	Ausência de correlação
-0,1< r < 0	Ínfima negativa
-0,5 < r ≤ -0,1	Fraca negativa
-0,8 < r ≤ -0,5	Moderada negativa
-1 < r ≤ -0,8	Forte negativa
r = -1	Perfeita negativa

Fonte: adaptado SIQUEIRA e TIBÚRCIO (2011).

3.4 Causalidade e regressões lineares

A análise de regressão é um dos métodos mais importantes dentre os métodos estatísticos. Com sua utilização, é possível conhecer os efeitos que algumas variáveis exercem sobre outras. Mesmo que não haja relação significativa de causa e efeito entre as variáveis analisadas, com a análise de regressão é possível construir uma relação funcional expressa por equações matemáticas.

Como pressuposto, a análise de regressão considera que devem existir, no mínimo, duas variáveis para sua viabilidade de aplicação, em que uma delas é chamada dependente ou endógena (em geral denotada por Y) e a(s) outra(s), denominada(s) de independente(s) ou exógena(s) (em geral, denotada(s) por X).

De forma geral, a análise de regressão pode representar a relação entre as variáveis da seguinte maneira:

$$Y = f(X_1, X_2, \ldots X_n) \quad (5)$$

Onde Y representa a variável dependente ou endógena e os $X_k\ (k = 1, 2, \ldots, k)$ representam as variáveis explicativas ou exógenas. Considere como aplicação os seguintes exemplos: (1) O estudo do crescimento populacional (Y) em função dos anos analisados (X); (2) Estudo da variação da produção de um item (Y), segundo o preço de venda (X_1) e a renda dos potenciais consumidores (X_2).

Quando na análise de regressão tiver uma única variável independente, tem-se o caso particular chamado análise de regressão simples, e quando se tiver mais de uma variável independente, tem-se o caso de análise de regressão múltipla (BUSSAB; MORETTIN, 2017). Em toda análise de regressão, a relação funcional construída entre as variáveis dependentes e independentes considera um termo residual ou de erro, o qual significa um ajuste para equilibrar o modelo elaborado, ou seja, ele representa os fatores não considerados no processo de modelagem e que podem ser influentes na relação entre as variáveis analisadas, e por ter uma natureza aleatória, torna os modelos elaborados em probabilísticos, os quais sob esta condição recebem o nome de modelos estatísticos.

TEORIA EM PRÁTICA

Considere o exercício a seguir, disponível em Murolo e Bonetto (2013, p. 42), o qual descreve a situação de uma empresa de embalagens plásticas. A empresa está preocupada com a demanda (Y_i) do produto fabricado por ela. Então, resolveu fazer um estudo sobre as variações dos preços de venda (X_i), fez um levantamento de dados e obteve as informações apresentadas na Tabela 1.

Tabela 1 – Demanda de embalagens plásticas por preço

Preço de venda X_i	16	18	20	23	26	28	30	33	35
Demanda Y_i	1200	1150	950	830	800	760	700	690	670

Fonte: adaptada de MUROLO e BONETTI (2013, p. 42).

Reescrevendo os dados, serão calculadas algumas medidas que ajudarão a obter as estimativas dos parâmetros do modelo de regressão linear simples a ser ajustado pelo método de mínimos quadrados ordinários. Os resultados para essa etapa encontram-se na Tabela 2.

Tabela 2 – Dados auxiliares

Ordem	**Preço de venda** (X_i)	**Demanda** (Y_i)	X_i^2	Y_i^2	X_iY_i
1	16	1200	256	1440000	19200
2	18	1150	324	1322500	20700
3	20	950	400	902500	19000
4	23	830	529	688900	19090
5	26	800	676	640000	20800
6	28	760	784	577600	21280
7	30	700	900	490000	21000
8	33	690	1089	476100	22770
9	35	670	1225	448900	23450
Total	229	7750	6183	6986500	187290

Fonte: adaptada de MUROLO e BONETTI (2013, p. 42).

Com os cálculos construídos na tabela auxiliar (Tabela 2), pode-se calcular os valores das estimativas dos parâmetros com maior facilidade a partir da linha dos totais, como mostrado a seguir. Agora, você tem condições para realizar os cálculos das estimativas dos coeficientes do modelo. Vamos lá!

Suponha que a empresa deseja estimar a demanda para um determinado preço do produto plástico, por exemplo $x = \$31$. Então, utilizando a equação ajustada, será obtido o seguinte valor para a demanda (quantidade de produto).

VERIFICAÇÃO DE LEITURA

1. Para utilizar uma amostra em uma pesquisa é necessário que seja executado um dos procedimentos de técnica de amostragem para a obtenção de uma amostra confiável. Qual método de amostragem é apropriado para reduzir a variabilidade de uma característica de uma população em uma amostra?

   a. Aleatória simples.

   b. Estratificada.

   c. Conglomerados.

   d. Sistemática.

   e. Cotas.

2. O critério de decisão de um teste de hipóteses está baseado em qual medida estatística?

   a. Média populacional.

   b. Hipótese alternativa.

   c. Estatística do teste.

d. Nível de significância.

e. Poder do teste.

3. De forma padrão, para realizar um teste de hipóteses se utiliza quantas hipóteses de teste?

a. Uma.

b. Três.

c. Cinco.

d. Duas.

e. Quatro.

Referências bibliográficas

BUSSAB, W.; MORETTIN, P. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017. 554p.

MUROLO, A.F.; BONETTO, G. **Matemática aplicada** à **administração, economia e contabilidade.** São Paulo: Cengage Learning, 2013. 506p.

PEREIRA, J. C. R. **Bioestatística em outras palavras.** São Paulo: Universidade de São Paulo, FAPESP, 2015.

RIBEIRO, T. C. S. C. **Probabilidade e estatística.** Londrina: Editora e Distribuidora Educacional S.A., 2015.

SIQUEIRA, Arminda. L., TIBÚRCIO, Jacqueline. D. **Estatística na área da saúde**: conceitos, metodologia, aplicações e prática computacional. Belo Horizonte: Coopmed, 2011. 520 p.

Gabarito

Questão 1 – Resposta: B

Resolução: A técnica de amostragem estratificada é utilizada quando se deseja reduzir a variabilidade de alguma característica de interesse na população alvo.

Questão 2 – Resposta: C

Resolução: O critério de decisão de um teste de hipóteses é baseado na estatística do teste após as hipóteses a serem testadas estarem definidas.

Questão 3 – Resposta: D

Resolução: Todo procedimento de teste de hipóteses utiliza exatamente duas hipóteses para a sua realização.

R: Principais comandos e estrutura

Autor: Marcelo Tavares de Lima

Objetivos

- Apresentar os principais comandos do R.
- Apresentar como são estruturados os comandos no R.
- Apresentar aplicações diversas dos principais comandos do R.

1. Introdução

A linguagem de programação R, conforme definem Oliveira, Guerra e McDonnell (2018, p. 10), pode ser entendida como “um conjunto de pacotes e ferramentas estatísticas, munido de funções que facilitam sua utilização, desde a criação de simples rotinas até análises de dados complexas”.

A interface gráfica RStudio ajuda ao iniciante em linguagem de programação R a se familiarizar com a construção dos códigos e a inserção de informações necessárias para executar seus comandos. Apesar de ser um facilitador para o uso de linguagem R, o RStudio também tem uma série de funcionalidades que melhoram o uso da linguagem R. Portanto, também é utilizado por usuários avançados.

Este texto apresenta os principais comandos do R e, também, mostra como são estruturados. Para isso, será utilizada a interface RStudio versão para a plataforma Windows. No entanto, é possível replicar todos os comandos em outras plataformas com o uso da versão apropriada do programa.

2. A linguagem R

A linguagem R ou, como também é conhecido, o ambiente R, segundo Mello e Peternelli (2013, p. 15), “é uma linguagem de alto nível e um ambiente para análise de dados e geração de gráficos”. Uma de suas vantagens é a sua gratuidade e, por ter seu código fonte aberto, podendo ser manipulado por qualquer usuário para implementação de melhorias.

Neste texto são apresentados alguns comandos essenciais, definições importantes e os principais cuidados que se deve ter quando for utilizar a linguagem R, principalmente através da interface RStudio.

O R é inteiramente gratuito e o RStudio tem versão gratuita e versões pagas. Para baixar ambos, iniciando pelo R, basta acessar o endereço disponibilizado nas referências e escolher a versão adequada para o seu computador. Para baixar o RStudio, busque o link de *download* na internet e, também, escolha a versão adequada para o seu computador.

Para realizar a instalação, deve-se seguir as instruções mostradas em sua tela após clicar nos arquivos de instalação baixados em seu computador ou consultar manuais disponíveis na internet. Para utilizar este material com êxito é importante que seja concluída a etapa de instalação dos dois programas.

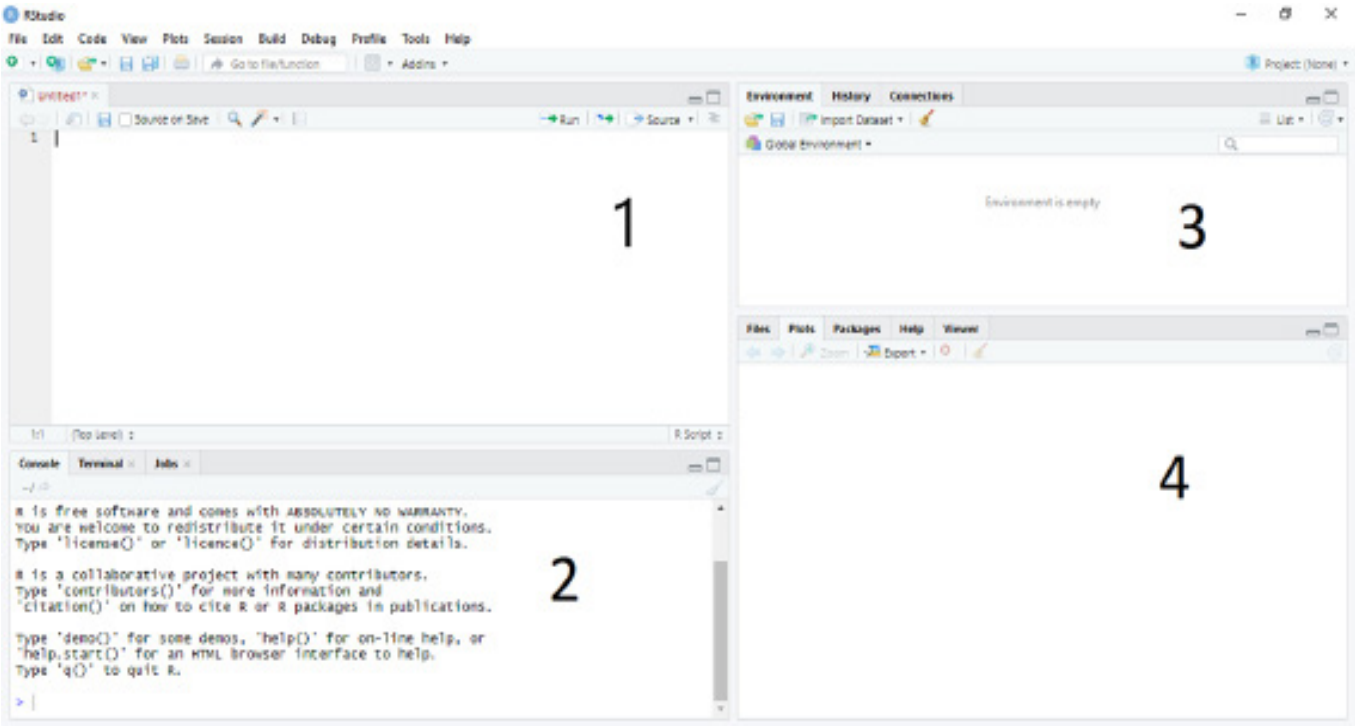
Tanto o R quanto o RStudio estão em língua inglesa. Portanto, faz-se necessário ter um mínimo de conhecimento de inglês técnico para sua utilização. A versão utilizada neste texto é a 1.2.1335 para Windows 64 bits.

2.1 O RStudio

O RStudio é uma das interfaces gráficas existentes para o R, no entanto é a mais utilizada atualmente e, também, a mais amigável para o uso da linguagem R. Além disso, ela tem um conjunto de funcionalidades que facilita o uso de comandos e obtenção de resultados, "é o que os especialistas em computação chamam de IDE (*Integrated Development Environment* ou Ambiente Integrado de Desenvolvimento)" (MELLO; PETERNELLI, 2013, p. 24).

A interface do RStudio é dividida em quatro partes principais, como mostrado na Figura 1 e detalhadas a seguir.

Figura 1 – Interface do RStudio



Fonte: elaborada pelo autor.

Parte 1. Editor de códigos: é onde se escreve e edita os códigos de linguagem R. Nesse mesmo espaço são criados os chamados *scripts*, ou seja, uma sequência de comandos que serão executados sequencialmente pelo R.

Parte 2. Console: é onde o R mostra a maioria dos seus resultados, ou seja, é onde são mostrados os resultados dos comandos executados. No console, também é possível escrever linhas de comando.

Parte 3. *Environment* e *History*: na aba *Environment* ficarão armazenados todos os objetos criados em uma sessão do R. Pode-se entender como sessão, valendo também para o RStudio, "o espaço de tempo entre o momento em que você inicia o R e o momento em que finaliza" (OLIVEIRA; GUERRA; McDONNELL, 2018, p. 11). Entende-se como objetos as variáveis declaradas nos comandos. Na aba History é criado um histórico dos comandos utilizados na sessão.

Parte 4. Abas Files, Plots, Packages, Help e Viewer: nesta parte, estão diversas funcionalidades do RStudio. Por exemplo, na aba Files você poderá fazer navegação de arquivos do seu computador, pois são mostradas algumas pastas do Explorer da máquina que está utilizando, o que permitirá definir o diretório de trabalho do R. A aba Plots mostra os gráficos gerados pelos comandos executados. A aba Packages mostra os pacotes instalados no R e a aba Help possui a documentação dos pacotes instalados e muitos exemplos que podem ajudar na construção de *scripts*.

2.2 Buscando ajuda (uso do *help*)

Todo usuário de programas computacionais, em algum momento, precisa de ajuda para realizar alguma tarefa. Para isso, é preciso saber procurar o que se deseja saber, independente do nível de conhecimento que se tenha a seu respeito. Existem algumas dicas, apresentadas por Oliveira, Guerra e McDonnell (2018), listadas a seguir: (1) Apesar de existir bastante ajuda em português, os autores recomendam sempre buscar em língua inglesa, pois, segundo os mesmos, é nesse idioma que se encontram os maiores fóruns de ajuda e comunidades que lidam com o R; (2) Antes de buscar ajuda externa, deve-se explorar o máximo possível o *help* do próprio programa R ou RStudio; (3) Quando solicitar ajuda em alguma comunidade ou fórum, como, por exemplo, o *Stack Overflow*, os autores recomendam que seja apresentado, pelo menos, um exemplo replicável, informando a versão do R que está sendo utilizada e o sistema operacional do computador onde o programa está sendo executado.

Embora não esteja mencionado na obra de referência (OLIVEIRA; GUERRA; McDONNEL, 2018), seria interessante que antes de postar/ solicitar ajuda em uma comunidade (ou fórum), realize-se uma inspeção das postagens para verificar se a sua dúvida já não foi sanada anteriormente.

2.3 Console

Como dito anteriormente, o console do RStudio é uma das quatro principais partes de sua interface gráfica. No console, também, é possível digitar linhas de comando em linguagem R e, é nele também que, o R apresenta a maioria dos resultados da execução dos comandos.

É possível observar no console, sempre no início de cada linha, a existência do símbolo “>”, o que indica a linha onde devem ser inseridos os comandos. Ao clicar nessa linha você está posicionando o cursor para digitar seu primeiro comando em R. Para exemplificar, digite 2*3 e aperte *enter*. O resultado obtido será apresentado abaixo igual a [1] 6, que é o resultado da operação de multiplicação entre 2 e 3, que é 6.

A linguagem R utiliza alguns símbolos como operadores aritméticos para realização de contas, por isso, muitos usuários reconhecem o R como uma grande calculadora científica. Os operadores aritméticos reconhecidos pelo R são apresentados no Quadro 1, conforme apresentado por Ribeiro (2012).

Quadro 1 – Símbolos de operações matemáticas

Símbolo	Operação
+	Adição
-	Subtração
*	Multiplicação
/	Divisão
^	Potenciação
%%	Mod (resto de divisão)

Fonte: RIBEIRO (2012).

É possível utilizar os operadores aritméticos para realizar cálculos sobre quaisquer valores, que podem ser digitados diretamente no console ou no editor de *scripts*. Veja os exemplos apresentados no Quadro 2.

Quadro 2 – Exemplos de operações matemáticas no R

7*9+2*6
2.5*4
(50+7)/(8*(3-5/2))
3^4

Fonte: OLIVEIRA, GUERRA e MCDONNELL (2018).

Observe que, se digitados no console e após apertar a tecla *enter* ao final de cada linha de comando, o R executa imediatamente o comando.

2.4 Editor de códigos

Quando se deseja criar uma série de linhas de comando e executá-las posteriormente, indica-se utilizar a parte 1 (editor de códigos) do RStudio. O conjunto de linhas de comandos é denominado *script* e, além de poder ser executado em conjunto, é possível armazenar em arquivo a programação elaborada em extensão (terminação) .R.

Outra diferença no uso do editor de códigos em relação ao console diz respeito à ordem de execução, que não é mais apertar a tecla *enter*. Para executar um *script* criado no editor de códigos, pode-se utilizar a tecla "Run" no canto superior direito do editor de códigos, mas, para isso, é necessário selecionar com o *mouse* as linhas de comando que deseja executar ou fazer uso da tecla *Source*, também no canto superior direito do editor de códigos, se não desejar selecionar o *script* ou parte dele antes da execução. Detalhes sobre o uso da tecla *Source* podem ser encontrados em Oliveira, Guerra e McDonnell (2018).

Ainda no editor de códigos é possível salvar os *scripts* criados. Para isso, basta clicar no desenho do disquete, na parte superior do RStudio, ou ir em *File* > *Save*, escolher uma pasta em seu computador e dar um nome apropriado para o arquivo.

Uma das vantagens do uso do RStudio para editar códigos do R é que com ele é possível ter aberto ao mesmo tempo várias janelas de códigos que são as chamadas guias. Isso pode ser feito clicando em *File -> New File -> R Script*.

PARA SABER MAIS

Além de operações aritméticas básicas, como soma, subtração, multiplicação e divisão, várias outras operações matemáticas também podem ser realizadas no RStudio. Por exemplo, é possível usar a função sqrt(x) para extrair a raiz quadrada de um valor armazenado em uma variável x, ou até mesmo inserir o próprio valor numérico no lugar da variável x, por exemplo, sqrt(25) e obter o valor 5 como resultado.

2.5 Comandos e argumentos importantes

Existem muitos comandos ou *scripts* agrupados em conjuntos, os quais são conhecidos como pacotes. Os pacotes recebem nomes específicos, os quais são utilizados quando se deseja utilizar o(s) *script(s)* contido(s) neles.

O R é composto por muitos pacotes (*packages*), também conhecidos como bibliotecas (*libraries*). Para exemplificar, os comandos básicos do R são agrupados em um pacote denominado base.

É possível instalar pacotes que façam parte da instalação padrão do R. Para isso, é necessário estar conectado à internet e clicar na aba *Packages* da parte 4 do RStudio e clicar no botão *Install*. Abrirá uma caixa onde há um campo para ser digitado o nome do pacote que se deseja instalar.

Muitos dos pacotes do R recebem atualização continuamente. Portanto, é importante que seja verificada a existência de atualizações com certa regularidade. Para isso, ainda na parte 4 do RStudio, na aba *Packages*, clique no botão *Update*. Logo após, será mostrada uma caixa com uma lista de pacotes que podem ser atualizados.

Sempre que se desejar utilizar um pacote específico, quando estiver instalado no seu computador, basta carregá-lo na memória com o uso do comando library (nome do pacote).

Dentro dos pacotes existem funções que também são conjuntos de comandos agrupados e que possuem argumentos, os quais devem ser declarados sempre que forem utilizadas. Os argumentos de uma função do R devem ser informados dentro de parênteses ( ) e são definidos com o uso do sinal de igualdade (=) e separados entre si por vírgula (,).

Para exemplificar, considere a função seq( ), utilizada para gerar uma sequência de números, possui argumentos chamados from, to e by. Portanto, o comando para ser executado corretamente precisa ser digitado como seq(from=1, to=10, by=3), por exemplo, e gerará o resultado [1] 1 4 7 10, pois a função pediu para gerar uma sequência de números iniciando em 1 e saltando de três em três até o número 10.

No entanto, é possível omitir os nomes dos argumentos dentro das funções, desde que sejam digitados na ordem correta em que foram criados. Considerando, ainda, o uso da função seq( ), pode-se obter o mesmo resultado anterior digitando apenas seq(1,10,3).

Oliveira, Guerra e McDonnell (2018, p. 15) afirmam que função é “uma sequência de comandos preparados para serem usados de forma simples e, assim, facilitar sua vida”.

As funções do R são utilizadas para trabalhos diversos como cálculos simples e, mais complexos, estatísticas gerais, geração de relatórios e elaboração de gráficos, etc. Durante a instalação do R, diversas funções também são instaladas e ficam prontas para usar.

2.5.1 Objetos (variáveis)

Para que o R não seja utilizado apenas como uma calculadora, faz-se necessário compreender o conceito de objeto ou variável. Oliveira, Guerra e McDonnell (2018, p. 15) definem objeto como "uma estrutura pré-definida que 'recebe' algum valor". De forma mais simples, é um espaço na memória do computador, que será utilizado pelo R para armazenar um valor ou resultado de um comando.

Para um objeto ser criado, pode-se utilizar a operação de atribuição, denotada com o símbolo "<-". Também é possível fazer a atribuição com o sinal de igualdade "=", porém não é muito utilizada pelos usuários.

Um cuidado que se deve ter ao criar um objeto diz respeito ao seu nome, o qual deve sempre ser iniciado com uma letra qualquer, maiúscula ou minúscula (o R faz diferenciação), seguida de letras ou caracteres especiais. O Quadro 3 apresenta alguns exemplos de criação de objetos.

Quadro 3 – Atribuição de objetos

Códigos	Comentários (não é obrigatório digitar)
x <- 10	# o objeto x receberá o valor 10.
15 -> y.A	# o objeto y.A receberá o valor 15.
X <- 16	# o objeto X receberá o valor 6.
Y1 = 13	# o objeto Y1 receberá o valor 13.

Fonte: adaptado de MELLO e PETERNELLI (2013).

Vale observar que existe diferença entre letras minúsculas e maiúsculas para o R, independente do sistema operacional utilizado. Outra observação é que todo objeto criado fica disponível na aba *Environment* da parte 3 do RStudio.

O Quadro 4 apresenta alguns símbolos e comandos importantes e bastante utilizados pelos usuários do R na criação de *scripts*.

Quadro 4 – Símbolos e comandos importantes

Ação	Comando
Faz com que o R ignore o que será digitado após*	#
Separa dois comandos numa mesma linha	;
Dado ausente	NA
Sai do programa R	q( )
Lista todos os objetos na sessão atual do R	ls( )
Remove um objeto x	rm(x)
Remove os objetos x e y	rm(x,y)
Concatenar valores**	c( )

* Tudo que for digitado após o símbolo # é ignorado pelo R, torna-se um comentário.
** Cria vetores de dados.

Fonte: MELLO e PETERNELLI (2013).

Os dados inseridos ou importados para o R são armazenados como objetos e podem ser armazenados de diversas maneiras ou estruturas. De forma mais genérica, podemos utilizar a classificação de dados utilizada por Oliveira, Guerra e McDonnell (2018): dados estruturados, semiestruturados e não estruturados.

Dados estruturados é o tipo mais amigável ou mais fácil de se trabalhar no R, pois são dados com um bom nível de organização de informações em colunas (atributos, variáveis, etc.) e linhas (registros, itens, observações, etc.). Em geral, esse tipo de dado encontra-se diretamente em bancos de dados ou em arquivos com algum tipo de separação entre suas colunas, como, por exemplo, uma planilha Excel.

Dados não estruturados são dados sem uma estrutura previsível, ou seja, são conjuntos de dados não uniformizados ou com forma única. No geral, são arquivos com conteúdo de textos. Por isso, não se pode dizer que são dados “desorganizados”, mas que são dados com

organizações específicas, como, por exemplo, mensagens de e-mail, arquivos PDF, imagens, vídeos, etc. Analisar dados não estruturados é muito mais complexo em comparação a uma análise de dados estruturados, os quais requerem análises mais avançadas como conhecimento avançado em mineração de dados. Em contrapartida, é o tipo de dado mais comumente disponível.

E por último, dados semiestruturados, assim como os estruturados, também possuem organização fixa, no entanto, nem sempre seguem o padrão linha/coluna, seguem estrutura mais complexa e, geralmente, hierarquizada, estruturada através de tags ou de marcadores de campos. Como exemplo, JSON, XML, HTML, YAML, etc. Este tipo de estrutura é mais comumente utilizado em troca de dados pela internet. Sua vantagem é que são mais facilmente transformados para dados estruturados.

No R, os dados armazenados em objetos podem ser organizados em diferentes estruturas. O Quadro 5 apresenta as estruturas mais utilizadas para armazenamento de dados.

Quadro 5 – Estruturas de dados mais utilizadas no R

Estrutura	Descrição
vector	Vetor com um ou mais elementos, array com uma dimensão.
matrix	Matriz, array com duas dimensões.
Array	Pode conter uma (vetor), duas (matriz) ou mais dimensões.
factor	Vetor de dados categóricos.
Data.frame	Parecido com a estrutura de matriz, mas permite colunas de diferentes tipos em um mesmo objeto.
list	Objeto que permite combinar diferentes estruturas de dados num único objeto.

Fonte: MELLO e PETERNELLI (2013)

Para exemplificar, saiba que a sequência de números criadas anteriormente com a função seq( ), que resultou em [1] 1 4 7 10, é armazenada em um objeto do tipo vetor.

Quanto ao tipo, que são diversos reconhecidos pelo R, distinguem-se pelo armazenamento de conteúdos diferentes, que vão desde tabelas de dados, textos, números ou valor lógico do tipo verdadeiro ou falso (Booleano). O Quadro 6 apresenta os principais tipos de objetos reconhecidos pelo R.

Quadro 6 – Principais tipos de objetos do R

Tipo	Descrição
character	Texto ou caracteres.
numeric	Números inteiros ou reais.
logical	Verdadeiro ou falso (TRUE/FALSE).
complex	Números complexos.
function	Comandos.

Fonte: MELLO e PETERNELLI (2013)

Objetos mais simples como vetor, matriz, array ou fator podem possuir somente um tipo de dado. Já estruturas mais complexas como list ou data.frame podem possuir mais de um tipo.

ASSIMILE

Aleatorização: em muitas pesquisas existe a necessidade de alocar sujeitos a grupos de estudo de forma aleatória. Em situações como essa é possível encontrar um exemplo com comandos do R para o caso de aleatorização em dois grupos, em Siqueira e Tibúrcio (2011). No exemplo, as autoras utilizam comandos como rep( ), runif( ), order( ), cbind( ) e data.frame( ). Vale a pena investigar sobre esses comandos!

2.6 Manipulação de dados

Quando se deseja manipular dados, em qualquer programa que seja, é importante, primeiramente, que sejam dados organizados e confiáveis. Muitas das vezes, ou quase sempre, os dados não estão prontos para um tratamento analítico mais profundo. Portanto, um primeiro passo a ser realizado é a sua limpeza. Por isso, é fundamental ter o domínio de técnicas de manipulação de dados.

Entende-se como manipulação de dados, de forma geral, "o ato de transformar, reestruturar, limpar, agregar e juntar dados" (OLIVEIRA; GUERRA; MCDONNELL, 2018, p. 25). É uma etapa de extrema importância no tratamento de dados a ponto de muitos estudiosos afirmarem que a manipulação representa cerca de 80% de um trabalho de análise de dados.

É possível armazenar dados em vetor com a função combine( ) ou apenas c( ), a qual combina todos os valores em um vetor. Considere que você insere um conjunto de dados no console (parte 2) do RStudio utilizando a função c( ) e aperta a tecla *enter*, sempre ao final de cada linha, para apresentar o que foi armazenado. O Quadro 7 apresenta a inserção e o resultado obtido.

Quadro 7 – Exemplo de armazenamento de dados em um vetor do R

```
x<-c(2,3,5,7,11)     # os cinco primeiros números primos
x                    # exibe o conteúdo do objeto x
[1] 2 3 5 7 11
```

Fonte: MELLO e PETERNELLI (2013).

Uma maneira de inserir dados mais complexos é através de um data. frame. Cada coluna de um data.frame é considerado um vetor e, é por isso, que este tipo de armazenamento pode ter mais de um tipo de dado por coluna. O Quadro 8 apresenta um exemplo realizado no console do RStudio, onde se criam vetores de diferentes tipos e depois

os combinam ou os unem em um data.frame. Utiliza-se no exemplo a função str( ) para saber o tipo de cada variável ou coluna.

Quadro 8 – Construção de um data frame no R

```
nome <- c("João", "José", "Maria", "Joana")
idade <- c(45, 12, 28, 31)
adulto <- c(TRUE, FALSE, TRUE, TRUE)
uf <- c("DF", "SP", "RJ", "MG")
clientes <- data.frame(nome, idade, adulto, uf)
clientes
   nome idade adulto uf
1  João    45   TRUE DF
2  José    12  FALSE SP
3 Maria    28   TRUE RJ
4 Joana    31   TRUE MG
str(clientes)   # mostra a estrutura do data.frame
'data.frame':    4 obs. of  4 variables:
 $ nome  : Factor w/ 4 levels "Joana","João",..: 2 3 4 1
 $ idade : num  45 12 28 31
 $ adulto: logi  TRUE FALSE TRUE TRUE
 $ uf    : Factor w/ 4 levels "DF","MG","RJ",..: 1 4 3 2
```

Fonte: adaptado de OLIVEIRA, GUERRA e MCDONNELL (2018).

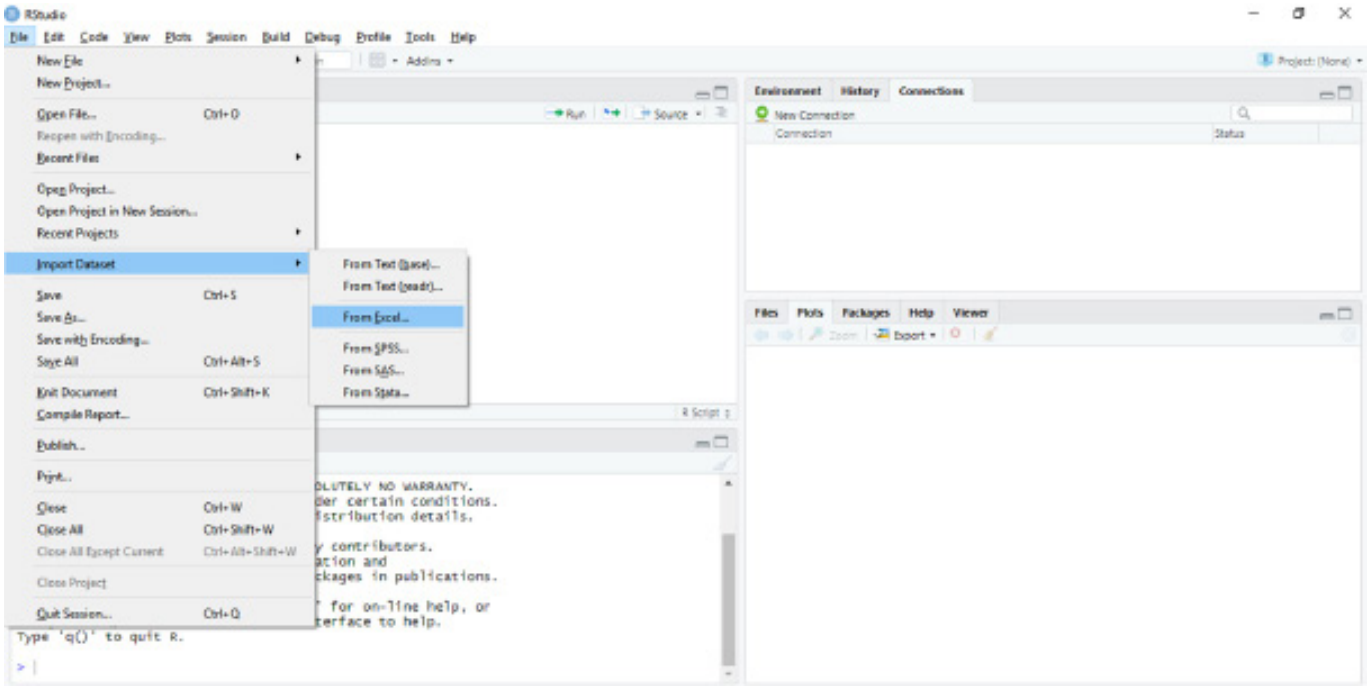
2.6.1 Leitura de dados de uma planilha Excel

É bastante comum realizar o armazenamento de dados de uma pesquisa em planilhas eletrônicas. A planilha Excel é bastante utilizada para este tipo de atividade.

No entanto, por se tratar de uma planilha de dados e não um programa de análise de dados, faz-se necessária a importação dos dados armazenados para um programa que possa realizar uma análise estatística mais complexa.

Como primeiro passo, realiza-se a importação para o RStudio, a qual pode ser feita de uma maneira bastante simples acessando o menu *File -> Import Dataset -> From Excel[...]*. Além do Excel, é possível importar dados em formato csv, sav, dta e sas. A Figura 2 mostra o caminho para realizar a importação de dados do Excel.

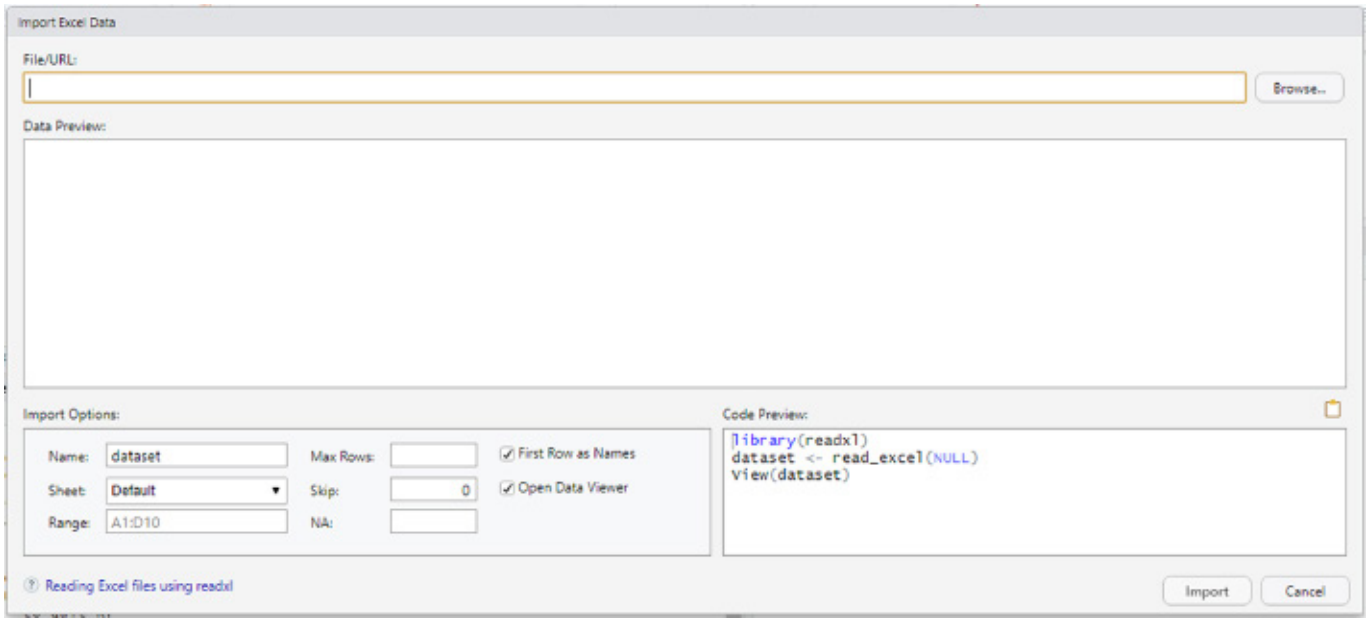
Figura 2 – Importação de dados do Excel para RStudio via menu



Fonte: elaborada pelo autor.

Feito o caminho apresentado na Figura 2, aparecerá uma caixa de diálogos como apresentado na Figura 3.

Figura 3 – Caixa de diálogos para importação de planilha Excel para o RStudio



Fonte: elaborada pelo autor.

É preciso preencher alguns campos da caixa de diálogos para que ela encontre o arquivo Excel que desejamos importar. Para iniciar a importação, clique no campo do canto superior direito, chamado *Browse...,* indique o diretório de localização da planilha. Em seguida, a planilha aparecerá no campo *Data Preview*. Com os dados pré-visualizados é possível mudar o tipo de dado clicando na primeira linha de cada coluna. Também é possível pular linhas indicando no campo *skip* e indicar os valores ausentes no campo NA. Para finalizar, basta clicar no botão *Import*, no canto inferior direito.

Existem outras maneiras de importação de dados para o RStudio. No entanto, não serão tratadas neste texto. Mais detalhes podem ser encontrados no suporte *online* do RStudio em Luraschi (2019).

Assim como para importar dados há mais de uma maneira, uma das características do R ou RStudio é que quase tudo pode ser realizado de mais de uma maneira. Isso pode ser considerado como uma vantagem da linguagem, porque permite ao usuário utilizar o que for do seu agrado e interesse.

TEORIA EM PRÁTICA

O exemplo aqui apresentado foi extraído de Mello e Peternelli (2013) e diz o seguinte: suponha que você esteja realizando um experimento para avaliar o desempenho escolar de uma classe de alunos. Para cada aluno (unidade) são registrados o nome, a idade, o sexo e a nota final. Os dados são organizados conforme a Tabela 1.

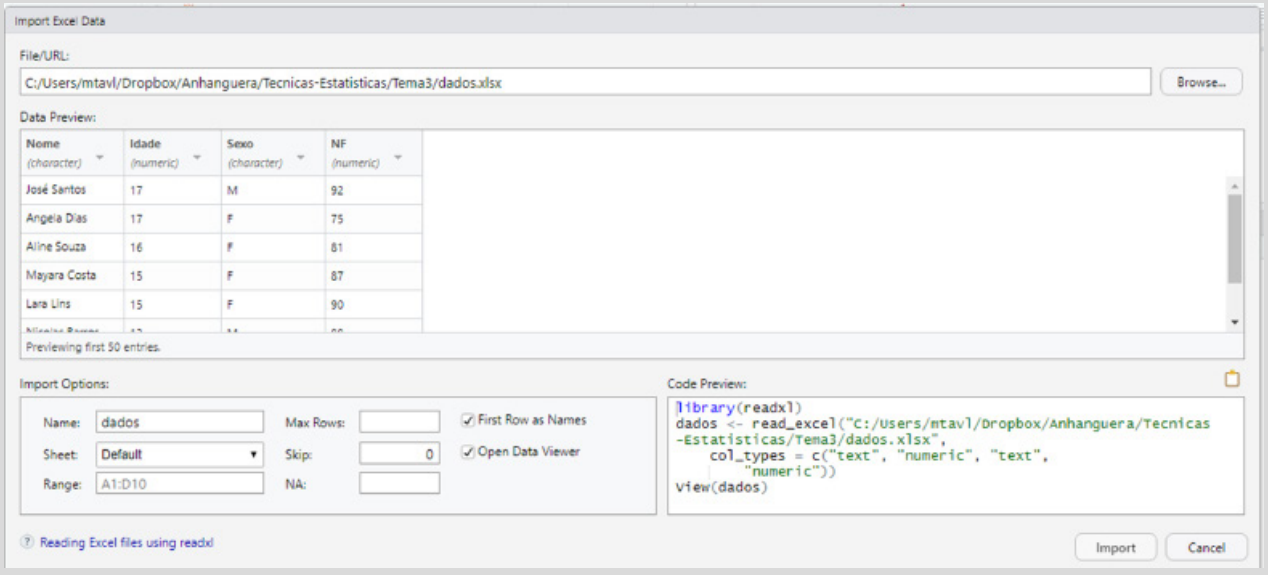
Tabela 1 – Exemplo de dados que podem ser representados como um data.frame.

Nome	Idade	Sexo	NF
José Santos	17	M	92
Angela Dias	17	F	75
Aline Souza	16	F	81
Mayara Costa	15	F	87
Lara Lins	15	F	90
Nicolas Barros	13	M	88

Fonte: MELLO e PETERNELLI (2013).

Os dados foram digitados em uma planilha Excel e em seguida foram importados para o R via RStudio, segundo procedimento apresentado anteriormente e que são mostrados para este caso particular.

Figura 4 – Importação de dados de planilha Excel

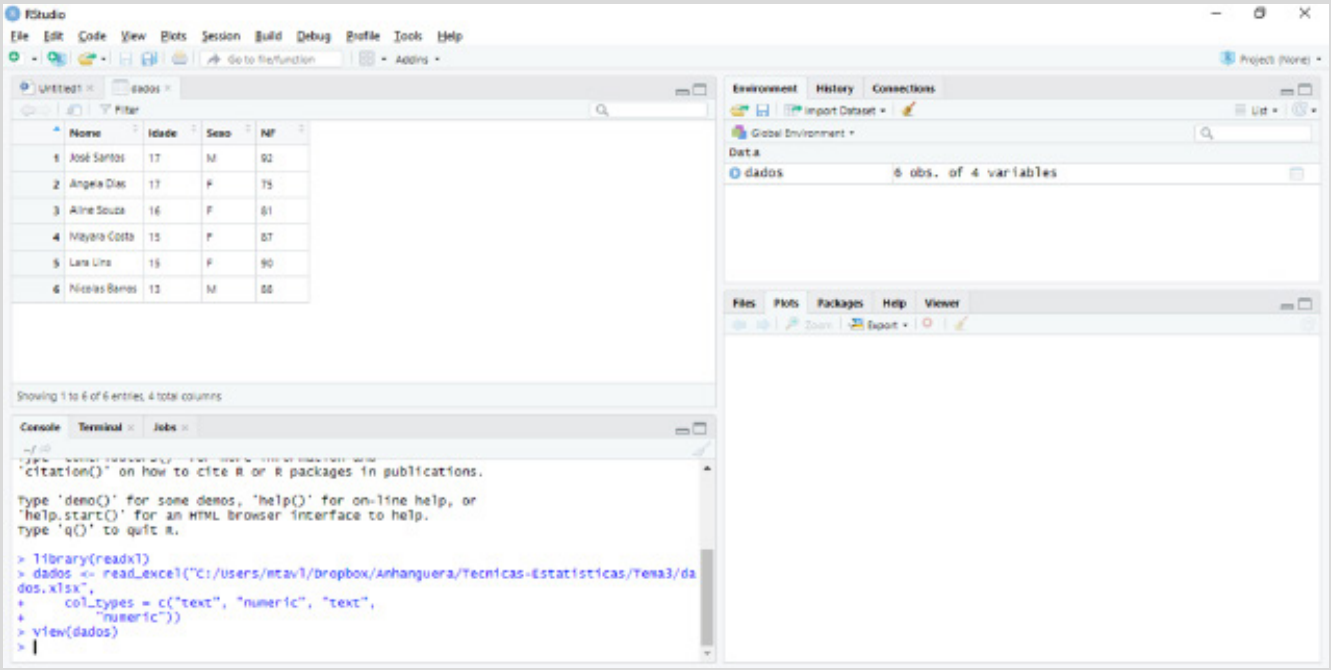


Fonte: elaborada pelo autor.

Observe que as colunas “Idade” e “NF” estão classificadas como numéricas (numeric). Não estava assim no primeiro momento em que a pré-visualização apareceu, estavam como caracter (character). Para modificar foi necessário apenas clicar na seta que se encontra junto ao nome da coluna e escolher o tipo de variável para a qual deseja modificar e clicar no botão Import.

A tela do RStudio ficará mostrando o conjunto de dados na parte 1 da interface gráfica do programa, como mostrado na Figura 5.

Figura 5 – Interface do RStudio após importação de dados de planilha Excel



Fonte: elaborada pelo autor.

Perceba que a planilha importada recebeu o nome de “dados”, que era o nome da planilha Excel “dados.xlsx”.

Ao digitar no console a palavra “dados”, o RStudio apresenta o seguinte resultado:

```
> dados
# A tibble: 6 x 4
  Nome          Idade Sexo     NF
  <chr>         <dbl> <chr> <dbl>
1 José Santos       17 M        92
2 Angela Dias       17 F        75
3 Aline Souza       16 F        81
4 Mayara Costa      15 F        87
5 Lara Lins         15 F        90
6 Nicolas Barros    13 M        88
```

Observe que uma numeração automática apareceu para cada uma das linhas. Isso ocorre porque o objeto “dados” foi armazenado como um data.frame, o qual atribui “nomes” para as linhas automaticamente, onde, por padrão, esses nomes são números em ordem crescente de cima para baixo.

Para finalizar, é possível verificar os atributos do objeto “dados”, que é um data.frame com o seguinte comando e com a obtenção de resultado.

```
> attributes(dados)
$names
[1] “Nome”  “Idade” “Sexo”  “NF”
$row.names
[1] 1 2 3 4 5 6
$class
[1] “tbl_df”     “tbl”        “data.frame”
```

O objeto dados possui três tipos de atributos, names, row.names e class. É possível realizar uma série de manipulações com o conjunto de dados, inclusive, realizar análises estatísticas que vão desde simples medidas descritivas até técnicas mais complexas.

VERIFICAÇÃO DE LEITURA

1. A interface gráfica RStudio é uma das interfaces existentes para leitura de linguagem R. Atualmente é a mais utilizada pelos usuários desse ambiente. É uma interface mais amigável e com algumas funcionalidades extras em relação ao programa de execução original. Com respeito ao RStudio, em quantas partes principais está dividida a sua interface gráfica?
   Assinale a alternativa CORRETA.

   a. Duas.

   b. Quatro.

   c. Seis.

   d. Uma.

   e. Três.

2. A interface gráfica RStudio é dividida em algumas partes principais, as quais têm funcionalidades distintas em grande parte da inserção e manipulação da linguagem R em seu ambiente. Em qual das partes principais do RStudio são apresentados a maioria dos resultados da execução das linhas de comando da linguagem R?
   Assinale a alternativa CORRETA.

   a. Editor de códigos.

   b. *Environment*.

   c. Console.

d. History.

e. Packages.

3. No editor de códigos do RStudio, quando se digita uma série de linhas de comandos para executá-los posteriormente, dá-se um nome específico para esse conjunto de comandos. Qual o nome dado ao conjunto de comandos digitados no editor de códigos do RStudio? Assinale a alternativa CORRETA.

   a. Objeto.

   b. Comandos.

   a. Pacote.

   a. Script.

   a. Variável.

Referências bibliográficas

LURASCHI, J. **Importing Data with RStudio.** 2019. Disponível em: https://support.rstudio.com/hc/en-us/articles/218611977-Importing-Data-with-RStudio. Acesso em: 28 jun. 2019.

MELLO, M. P.; PETERNELLI, L. A. **Conhecendo o R:** uma visão mais que estatística. Viçosa, MG: UFV. 2013.

OLIVEIRA, P.F.; GUERRA, S.; McDONNELL, R. **Ciência de dados com R:** introdução. Brasília: IBPAD. 2018. Disponível em: https://www.ibpad.com.br/o-que-fazemos/publicacoes/introducao-ciencia-de-dados-com-r#download. Acesso em: 28 jun. 2019.

RIBEIRO, R. **Linguagem R:** guia prático para iniciantes. Amazon Servicos de Varejo do Brasil Ltda. Arquivo Kindle. 2012.

RSTUDIO. **Download RStudio**. 2019. Disponível em: https://www.rstudio.com/products/rstudio/download/. Acesso em: 28 jun. 2019.

SIQUEIRA, Arminda. L., TIBÚRCIO, Jacqueline. D. **Estatística na área da saúde**: conceitos, metodologia, aplicações e prática computacional. Belo Horizonte: Coopmed, 2011. 520 p.

The R Foundation. **The R Project for Statistical Computing**. 2019. Disponível em: https://cloud.r-project.org/. Acesso em: 28 jun. 2019.

Gabarito

Questão 1 – Resposta: B

Resolução: O RStudio é uma interface gráfica para manipulação e execução de linguagem R e está dividida em quatro partes principais visuais.

Feedback de reforço: Lembre-se da interface do RStudio e de como ela está dividida.

Questão 2 – Resposta: D

Resolução: O RStudio é uma interface gráfica do ambiente R e apresenta quatro partes principais. Na parte chamada console do RStudio são apresentadas a maioria dos resultados da execução de comandos da linguagem R.

Feedback de reforço: Lembre-se da interface do RStudio e de como ela está dividida.

Questão 3 – Resposta: D

Resolução: O conjunto de comandos digitados no editor de códigos do RStudio para posterior execução recebe o nome de script.

Feedback de reforço: Lembre-se dos conceitos básicos associados à linguagem de programa R e como são praticados na interface RStudio.

Análise de dados com a linguagem R

Autor: Marcelo Tavares de Lima

Objetivos

- Apresentar métodos de análise de dados.
- Descrever a programação em linguagem R que realiza análise de dados.
- Apresentar aplicações de análise de dados em R.

1. Introdução

Para conhecer os dados que temos disponíveis, é necessário realizar uma exploração. Para tanto, precisamos buscar as ferramentas metodológicas apropriadas para a sua manipulação.

Para um bom tratamento exploratório e analítico de dados é preciso conhecer bem os dados que se tem disponível. Quando se tem esse conhecimento, conseguimos selecionar as ferramentas apropriadas, inclusive o programa computacional apropriado.

Neste texto, serão apresentados os principais comandos da linguagem R que fazem análise de dados. Desejamos que você possa se apropriar do conteúdo e que ele possa trazer a você um conhecimento significativo. Bons estudos!

2. Tipos de análise de dados

A análise estatística de dados pode ser realizada de várias maneiras e com vários métodos científicos existentes. Segundo Siqueira e Tibúrcio (2011, p. 5), "existe mais de uma ênfase de análise estatística: a estatística clássica, também chamada de frequentista, e a estatística bayesiana". Muitas ferramentas de mineração de dados (*data mining*) e de inteligência artificial fazem uso de técnicas estatísticas clássicas e bayesianas.

Os métodos clássicos estatísticos são métodos de análise descritiva de dados, com elaboração de tabelas de distribuição de frequências, análise de correlação e associação, elaboração de gráficos, etc. Já os métodos bayesianos são métodos que fazem parte da inferência bayesiana onde "evidências ou observações são usadas para calcular probabilidades de que uma hipótese seja verdadeira ou mesmo para atualizar uma probabilidade previamente calculada" (SIQUEIRA; TIBÚRCIO, 2011, p. 6).

Como etapas de uma análise de dados, podemos listar as seguintes tarefas, segundo Siqueira e Tibúrcio (2011): inspeção cuidadosa dos dados, a formulação de um modelo apropriado para os dados, o ajuste do modelo aos dados, a verificação do ajuste do modelo e a apresentação dos achados e a realização de conclusões. Ainda é possível considerar que a análise pode ser realizada em duas etapas: uma análise preliminar, onde são feitas as verificações dos dados, e a análise definitiva.

Como análise preliminar, podemos considerar as seguintes etapas: a) processamento dos dados de forma conveniente para análises posteriores; b) verificação da qualidade dos dados, como busca por inconsistências, erros, observações atípicas, dados faltantes, dentre outros problemas relacionados aos dados. Ainda nesta etapa do trabalho, é importante avaliar a necessidade de realização de algum tipo de transformação nos dados e, por último; c) análise descritiva dos dados.

A análise definitiva de um conjunto de dados, para ser realizada, exige do analista que tenha em mente, previamente, a determinação do método apropriado, a qual depende diretamente de vários fatores. Chatfield (1995 *apud* Siqueira e Tibúrcio, 2011, p. 6) apresenta como sugestão o levantamento de alguns questionamentos, como:

> existe alguma publicação sobre o assunto? Você, alguém que conhece ou algum centro de pesquisa já enfrentou problema parecido? Há alguma informação *a priori*? É possível reformular o problema de maneira que o torne mais simples para ser resolvido? É possível dividir o problema em partes disjuntas e resolver cada uma delas por vez? Quais são os objetivos do estudo? Qual é a estrutura dos dados? Dentre outros questionamentos, para dar um direcionamento certo para a escolha do método apropriado.

A prática se inicia com a disposição de um banco de dados brutos, o qual deve ser organizado de maneira apropriada para ser consolidado. A partir disto, parte-se para as etapas de análise, como já descrito, são

duas etapas: análise descritiva (por onde sempre se deve começar) e análise inferencial ou inferência estatística.

A análise descritiva é composta, basicamente, por elaboração de tabelas, gráficos e algumas medidas estatísticas, que também são conhecidas como estatísticas, como, por exemplo, valores de médias, medianas, variâncias, desvios padrão, coeficientes de variação, proporções, etc. É basicamente uma etapa de descrição do conjunto de dados.

Apesar da possibilidade de se obter muitos resultados com a análise descritiva, ela pode não ser o suficiente, pois, se a análise se encerrasse por aí, os resultados encontrados valeriam apenas para o conjunto ou amostra analisada (SIQUEIRA; TIBÚRCIO, 2011). Por isso, é necessário ir além desta etapa de análise para realizar análise que possa extrapolar o resultado encontrado com a amostra trabalhada. Falar em extrapolar o resultado significa que os resultados encontrados em um conjunto de dados específico (amostra) podem ser considerados resultados válidos para a população de onde foram retirados.

A inferência estatística é a parte da estatística composta por um conjunto de métodos que ajudam a extrapolar resultados para uma população. Tem-se, de forma simplificada, dois principais métodos básicos, que são os testes de hipóteses (TH) e a estimação (pontual ou intervalar) (SIQUEIRA; TIBÚRCIO, 2011). Neste texto serão apresentados métodos de análise descritiva de dados com suporte do programa computacional R através da IDE (interface) RStudio. Desejamos que você possa aproveitar e aprender bastante com esta leitura. Bons estudos!

3. Análise de dados com o R

Neste item iremos apresentar exemplos de aplicação da linguagem R para tratar dados, tanto quantitativos quanto qualitativos. Também serão apresentados os comandos utilizados para cada análise realizada.

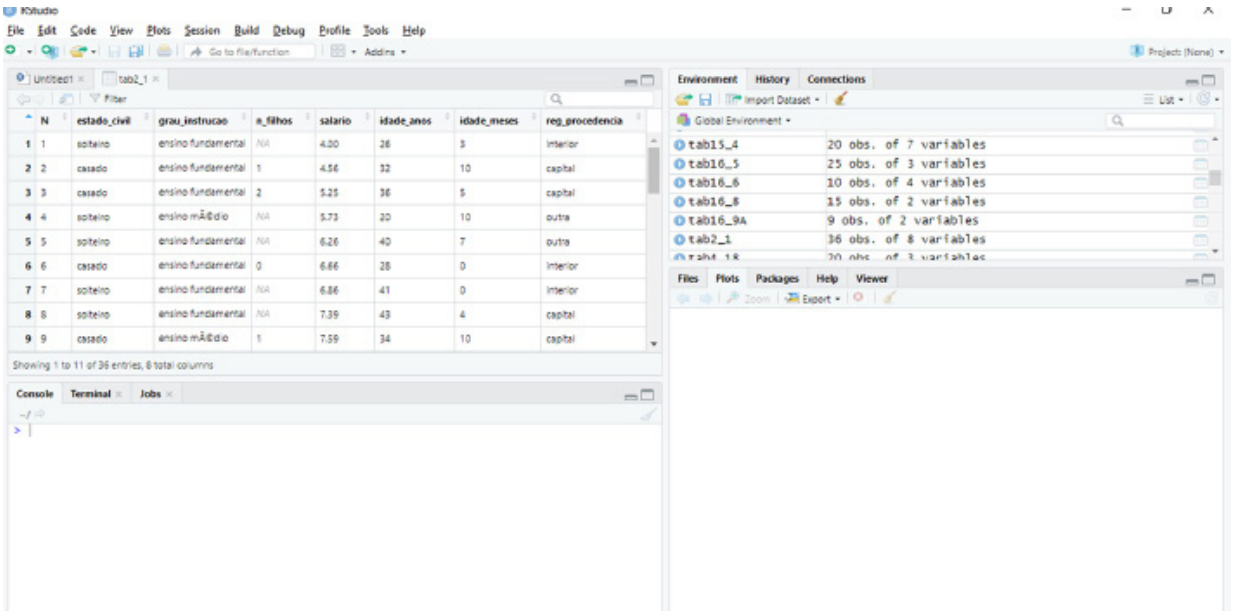
Os dados utilizados são originários do professor Pedro Morettin (2019), professor titular do Departamento de Estatística da Universidade de São Paulo, os quais estão disponibilizados na página de internet dos trabalhos do professor. No portal do professor são disponibilizados os dados em planilha Excel e, também, em R Workspace, formato de dados para a linguagem R.

Será utilizada a IDE RStudio por ser uma interface que facilita e traz muitas vantagens na utilização de linguagem R para análise de dados.

Será utilizada a base "tab2_1" dos dados disponibilizados. Para acessar a base de dados, basta fazer *download* do arquivo "dados" em formato R Workspace, disponibilizado na página do Professor Morettin, e abrir o RStudio para abrir o arquivo.

Para abrir um conjunto de dados no RStudio basta utilizar os menus de comandos que se encontram na parte superior da tela a partir do menu "File". Depois de aberto o banco de dados no RStudio, o seu conteúdo aparecerá na tela superior esquerda, conforme a Figura 1. Os dados também são apresentados em Bussab e Morettin (2017).

Figura 1 – Interface do RStudio com a abertura do banco de dados.



Fonte: elaborada pelo autor.

Vale lembrar que, por se tratar de um conjunto de dados pronto para análise, não serão realizadas etapas de organização (limpeza, transformação, etc.), etapa extremamente importante sempre que se deseja realizar uma análise de dados.

O conteúdo do banco de dados se refere a

> informações sobre estado civil, grau de instrução, número de filhos, salário (expresso como fração de salário mínimo), idade (medida em anos e meses) e procedência de 36 empregados da seção de orçamentos da Companhia MB (BUSSAB; MORETTIN, 2017, p. 14).

Para iniciar a análise de dados, é importante obter um resumo das variáveis que compõem o conjunto de dados. Para isto, é possível utilizar o comando “summary( )”, conforme mostra o Quadro 1, juntamente com o resultado da execução do comando.

Quadro 1 – Obtendo um resumo das variáveis que compõem um conjunto de dados com o comando “summary( )”.

Comando:

```
summary(tab2_1)
```

Resultado:

```
      N          estado_civil          grau instrução    n._filhos
 Min. : 1.00  casado :20 ensino fundamental:12     Min.   :0.00
 1st Qu.: 9.75 solteiro:16    ensino médio      :18     1st Qu.:1.00
 Median: 18.50                superior : 6 Median: 2.00
 Mean: 18.50                                  Mean: 1.65
 3rd Qu.: 27.25                                  3rd Qu.: 2.00
 Max. : 36.00                                 Max.: 5.00
                                          NA’s: 16
   salário        idade_anos     idade_meses     reg._procedência
 Min.: 4.000   Min. : 20.00 Min.: 0.000 capital: 11
 1st Qu.: 7.553   1st Qu.:30.00   1st Qu.: 3.750   interior:12
 Median: 10.165   Median: 34.50   Median: 6.000 outra: 13
 Mean: 11.122   Mean: 34.58 Mean: 5.611
 3rd Qu.: 14.060   3rd Qu.:40.00   3rd Qu.: 8.000
 Max.: 23.300   Max.: 48.00 Max.: 11.000
```

Fonte: elaborado pelo autor.

É possível observar que as medidas apresentadas com o uso do comando “summary”(tab2_1)” são valor mínimo (Min.), primeiro quartil (1st. Qu.), mediana (Median), média aritmética simples (Mean), terceiro quartil (3rd. Qu.) e valor máximo (Max) para variáveis quantitativas e frequências das categorias para variáveis qualitativas. O comando “summary(tab2_1)” pode ser digitado tanto na parte 1 (editor de códigos) no canto superior esquerdo do RStudio quanto na parte 2 (console) no canto inferior esquerdo.

O primeiro resultado (da esquerda para a direita) mostrado no Quadro 1 é referente a uma coluna existente no conjunto de dados que se refere apenas a enumeração dos sujeitos (N) que compõem o banco de dados, portanto pode ser ignorada. Os demais resultados podem ser utilizados para avaliação dos valores que compõem as variáveis.

Considerando, ainda, o resultado apresentado no Quadro 1, observe que a variável “estado_civil” contém duas categorias, solteiro e casado, as quais têm frequência igual a 20 e 16, respectivamente, ou seja, do total de 36 sujeitos que compõem o banco de dados, 20 são solteiros e 16 casados.

A variável “n._filhos” contém o número de filhos de cada um dos 36 sujeitos que compõem o banco de dados. Trata-se de uma variável quantitativa, no entanto, vale a pena chamar a atenção para uma informação que é dada pelo resumo, a quantidade de ausência de informação (NA’s). Existem 16 registros sem a informação da quantidade de filhos.

Suponha que se deseja obter o total, a média aritmética simples e o desvio padrão da idade (em anos), segundo o estado civil dos sujeitos. Para isso, podemos utilizar o pacote “descr”, que fornece alguns valores de medidas descritivas de variáveis quantitativas estratificadas por categorias de uma variável qualitativa.

Para instalar o pacote “descr”, basta digitar o comando “install.packages(“descr”)” e, para carregar o pacote digite o comando

“library(descr)”, sem as aspas. Suponha que desejamos obter medidas descritivas para a idade (em anos) segundo o estado civil dos sujeitos. O Quadro 2 apresenta o comando utilizado e o resultado obtido.

Quadro 2 – Medidas descritivas para a idade (em anos) segundo estado civil

Comando:

```
Options(digits = 4) # apresenta resultados com duas decimais
compmeans(tab2_1$idade_anos, tab2_1$estado_civil)
```

Resultado:

Valor médio de “tab2_1$idade_anos” segundo “tab2_1$estado_civil”

	Média	N	Desv. Pd.
casado	35.15	20	5.89
solteiro	33.87	16	7.80
Total	34.58	36	6.73

Fonte: elaborado pelo autor.

Os resultados apresentados pelo R são descritos com o uso de notação internacional para delimitador decimal, ou seja, os dígitos decimais de um número são separados da sua parte inteira por ponto (.) e, valores acima de 999 são descritos com separação de vírgula (,) para a casa de milhares, milhões, etc.

Pode-se observar os resultados para os casados, a idade média é de 35,15, sendo que são 20 sujeitos e, desvio padrão para a idade igual a 5,89 anos. Os solteiros apresentam idade média de 33,875 anos, sendo um total de 16 sujeitos no banco de dados e desvio padrão igual a 6,7374 anos. Há também uma linha de total, a qual apresenta a idade média, a frequência e o desvio padrão para todos os sujeitos do banco de dados.

A função “compmeans” fornece o resultado apresentado. Ela possui dois argumentos, “tab2_1$idade_anos” e “tab2_1$estado_civil”. O primeiro argumento se refere à variável numérica que desejamos obter medidas descritivas e a segunda variável é a variável que estratifica os resultados

segundo as suas categorias (variável qualitativa). Perceba que na frente dos nomes das variáveis aparece o nome do banco de dados seguido de “$” para indicar para o programa o banco de dados em que estão armazenadas. É possível evitar escrever os argumentos sem o nome do banco de dados. Para isso, basta executar o comando “attach(tab2_1)” e testar o comando novamente como “compmeans(idade_anos, estado_civil)”. Você verificará que obterá o mesmo resultado do Quadro 2.

Quando se deseja obter uma tabela de frequências que envolva duas variáveis qualitativas, pode-se utilizar comando do R conforme apresentado no Quadro 3, utilizando as variáveis “grau instrução” e “reg_procedência”, as quais são argumentos da função “table”.
O primeiro argumento, “grau instrução” será descrito nas linhas do resultado e, o segundo argumento “reg_procedência” apresentará os resultados nas colunas.

Quadro 3 – Obtenção de tabela de frequências cruzando duas variáveis qualitativas.

Comando:
Table (grau instrução, reg_procedência)
Resultado:

	reg_procedência		
grau_instrução	capital	interior	outra
ensino fundamental	4	3	5
ensino médio	5	7	6
superior	2	2	2

Fonte: elaborado pelo autor.

O resultado obtido com uso do comando “table( )” são as frequências segundo classificação cruzada para as variáveis de grau de instrução e região de procedência. Por exemplo, a frequência 4 apresentada na primeira linha e primeira coluna do resultado do corpo da tabela indica que no banco de dados existem 4 sujeitos com ensino fundamental e procedentes da capital. A interpretação dos demais valores internos da tabela resultante é feita de forma semelhante.

PARA SABER MAIS

O Professor Pedro Alberto Morettin é professor titular aposentado do Departamento de Estatística da Universidade de São Paulo (USP). Ele é doutor e mestre em Estatística pela University of California (Berkeley) e é graduado em Matemática pela USP. Tem várias produções na área de estatística, em especial na área de séries temporais. Contribui muito com o desenvolvimento da estatística no Brasil e no exterior (MORETTIN, 2019).

Uma outra maneira de obter resultados do cruzamento entre duas variáveis qualitativas é através da função “crosstab”, que pertence ao pacote “descr”, o qual já está carregado no R, portanto não precisa ser carregado novamente. Se por um acaso você fechar o programa, precisará carregar novamente o(s) pacote(s) necessário(s) para realizar análises de seu interesse. O comando completo com argumentos e o resultado é apresentado no Quadro 4.

Quadro 4 – Tabela de frequências cruzadas entre variáveis qualitativas com o RStudio

Comando:

crosstab(grau_instrucao,reg_procedencia)

Resultado:

Conteúdo das células

|-------------------------|

| Contagem |

|-------------------------|

==

reg_procedência

grau_instrução	capital	interior	outra	Total
ensino fundamental	4	3	5	12
ensino médio	5	7	6	18
superior	2	2	2	6
Total	11	12	13	36

Fonte: elaborado pelo autor.

É possível perceber que a diferença nos resultados entre os comandos “table” e “crosstab” é que este último apresenta os totais por linhas e colunas, além de apresentar resultados de forma mais elegante.

ASSIMILE

As análises de dados realizadas com a linguagem R são ricas no sentido de que o mesmo tipo de análise pode ser realizado com comandos distintos, produzindo resultados com pequenas diferenças entre si. Portanto, o analista precisa conhecer as possibilidades de comando para realizar uma análise específica e escolher a que lhe for mais conveniente.

É possível verificar se existe algum tipo de associação entre as variáveis de grau de instrução e região de procedência pelo teste estatístico de qui-quadrado ou teste exato de Fisher. Maiores detalhes sobre este teste podem ser encontrados nas referências deste material.

O comando utilizado para a obtenção de valores esperados e o teste estatístico de qui-quadrado continua sendo “crosstab”, a diferença é que são acrescentados mais dois argumentos, conforme apresentado no Quadro 5.

Quadro 5 – Tabela cruzada com valores de frequências absolutas e valores esperados, com teste de associação de qui-quadrado e teste exato de Fisher.

Comando:
crosstab(grau_instrucao,reg_procedencia,expected = TRUE,chisq =TRUE)
Resultado:

Conteúdo das células
Contagem
Valores esperados

grau_instrução	reg_procedência capital	interior	outra	Total
ensino fundamental	4	3	5	12
	3.7	4.0	4.3	
ensino médio	5	7	6	18
	5.5	6.0	6.5	
superior	2	2	2	6
	1.8	2.0	2.2	
Total	11	12	13	36

Estatísticas para todos os fatores da tabela
Pearson’s Chi-squared test

Qui² = 0.6614219 g.l. = 4 p = 0.956
Frequência esperada mínima: 1.833333
Células com frequências esperada < 5: 6 de 9 (66.66667%)

Warning message:
In chisq.test(tab, correct = FALSE, ...) :
  Chi-squared approximation may be incorrect
Comando:
fisher.test (grau_instrução,reg_procedência)
Resultado:
Fisher's Exact Test for Count Data
data: grau_instrucao and reg_procedencia
p-value = 0.9716
alternative hypothesis: two.sided

Fonte: elaborado pelo autor.

Com a inclusão de dois argumentos na função "crosstab" foi possível obter dois resultados importantes, os valores esperados para os valores da tabela e o teste de associação entre as variáveis, o teste qui-quadrado e o teste exato de Fisher. Para ser considerado válido, o teste qui-quadrado exige que a maioria dos valores esperados $\left(\frac{\text{total da coluna} \times \text{total da linha}}{\text{total geral}}\right)$ tenham valores maiores ou iguais a 5.

tenham Se isso não for garantido, o teste qui-quadrado não é considerado um teste válido e, para tanto, o teste exato de Fisher passa a substituí-lo (SIQUEIRA; TIBÚRCIO, 2011).

Observado o primeiro resultado do Quadro 5, é possível perceber que a maioria dos valores esperados são menor que 5, por isso, no mesmo resultado é mostrada uma mensagem de aviso (*Warning message*:) informando que o teste qui-quadrado pode ter um resultado inválido ou não confiável. Para isso, um novo comando é elaborado com a função "fisher.test (grau_instrucao, reg_procedencia)" para obtermos o resultado do teste exato de Fisher.

Para avaliar a possível existência de associação entre as variáveis (dependência), basta olhar o resultado do valor p (p-value). É usual considerar evidências de associação entre as variáveis quando o valor

p é menor que 0,05. Caso contrário, conclui-se que não há evidências de associação entre as variáveis. Considerando este critério, pode-se concluir pela inexistência de evidência de associação entre grau de instrução e região de procedência (p-value = 0.9716).

Para avaliar associação entre duas variáveis quantitativas é possível construir uma matriz de correlação para as variáveis quantitativas com valores de Coeficiente de correlação de Pearson, conforme comandos apresentados no Quadro 6.

Quadro 6 – Verificando associação entre duas variáveis quantitativas pelo Coeficiente de correlação de Pearson

Comando para separar as variáveis quantitativas do banco em um vetor:

x<-c("n_filhos","salario","idade_anos","idade_meses")

Comando para obter uma matriz de correlação de Pearson para as variáveis quantitativas:

cor(tab2_1[x])

Resultado:

	n._filhos	salário	idade_anos	idade_meses
n._filhos	1	NA	NA	NA
salário	NA	1.0000000000	0.3633622	0.0007217399
idade_anos	NA	0.3633621809	1.0000000	-0.1377571614
idade_meses	NA	0.0007217399	-0.1377572	1.0000000000

Fonte: elaborado pelo autor.

Uma medida de correlação pode variar de –1 até +1 e é considerada associação forte entre duas variáveis quanto mais próximo desses valores extremos for a medida de correlação. Por exemplo, a medida de correlação entre as variáveis idade_anos e salário é igual a 0,3633, que

pode ser considerada uma correlação fraca positiva. Uma correlação positiva indica que o crescimento ou decrescimento das duas variáveis envolvidas na obtenção da estatística tem o mesmo sentido, ou seja, quando uma cresce a outra também cresce. O mesmo ocorre para o decrescimento (SIQUEIRA. TIBÚRCIO, 2011).

Este texto não tem a pretensão em esgotar as possibilidades de análise de dados com a linguagem R. Existe uma infinidade de comando e funções que podem ser utilizadas em uma análise. Portanto, sugere-se que você não se limite a este texto para dar continuidade nos seus estudos de linguagem de programação R e de conhecimento de métodos estatísticos.

TEORIA EM PRÁTICA

Considere que você trabalha em uma empresa de pesquisa de mercado e precisa analisar o perfil de potenciais clientes de um determinado produto. Para isso, você realiza uma coleta de dados para buscar informações que possam subsidiar seu trabalho.

O conjunto de dados utilizado foi retirado de Costa (2012) e se refere a uma pesquisa de perfil demográfico feito a 20 consumidores adultos do produto X. O banco de dados é apresentado na Tabela 1. Perceba que não são utilizados acentos nas palavras, pois eles causam problemas na hora da importação para o ambiente R, como, por exemplo, para a escolaridade relativa ao ensino médio, o dado na tabela é apresentado como “Ens. Medio”.

Tabela 1 – Conjunto de dados

Sexo	Idade	Escolaridade	N_filhos	Classe_social
M	35	Ens. Medio	2	B
M	25	Ens. Medio	1	B
F	40	Ens. Superior	1	C
M	25	Ens. Medio	3	B
M	32	Ens. Medio	2	C
F	22	Ens. Medio	0	C
M	37	Ens. Superior	2	B
M	28	Ens. Medio	0	B
F	25	Ens. Medio	1	B
F	39	Ens. Superior	2	C
M	35	Ens. Fundamental	1	B
F	21	Ens. Fundamental	0	A
F	27	NA	0	A
F	45	Ens. Medio	2	C
M	57	Pos-Graduacao	4	C
F	33	Ens. Medio	2	A
M	36	Ens. Fundamental	0	B
M	35	Ens. Medio	2	C
M	33	Ens. Medio	2	B
F	22	Ens. Superior	0	C

Fonte: adaptada de Costa (2011).

Sexo: M – masculino; F – feminino.
Idade: em anos com dois dígitos.
Escolaridade: NA – sem informação.
Classe social: A – alta; B – média; C – baixa.

O conjunto de dados está disponibilizado em planilha MS Excel e você deseja exportar para o RStudio para fazer a análise dos dados com a linguagem R. Portanto, para fazer isso, utilizar a seguinte programação em R.

```
library(readxl)
> dados <- read_excel("dados.xlsx",
+     col_types = c("text", "numeric", "text",
+         "numeric", "text"), na = "NA")
> View(dados)
```

Também é possível realizar a importação com os menus disponíveis no RStudio. A planilha se chama “dados” e está armazenada em alguma pasta no seu computador. Para fazer a exportação, precisa carregar o pacote “readxl” para utilizar a função “read_excel”.

Agora, você precisa utilizar os conhecimentos que tem de análise de dados com a linguagem R para produzir um relatório e apresentar para o seu grupo de trabalho. Bom trabalho!

VERIFICAÇÃO DE LEITURA

1. Qual das medidas estatísticas avalia a associação entre duas variáveis quantitativas?
   Assinale a alternativa CORRETA.

   a. Média aritmética.

   b. Qui-quadrado.

   c. Coeficiente de correlação.

   d. Valor esperado.

   e. Desvio padrão.

2. Se o conjunto de dados a ser analisado estiver armazenado em planilha eletrônica e se pretende realizar a análise com a linguagem R, será necessário exportar os dados. Qual o nome de um dos pacotes criados para esse fim? Ou seja, para importação de dados para o R.

Assinale a alternativa CORRETA.

a. descr.

b. readxl.

c. table.

d. crosstab.

e. summary.

3. Para que a associação entre duas variáveis qualitativas seja considerada significativa é preciso observar o resultado numérico de uma determinada medida. De qual medida estatística estamos falando?
Assinale a alternativa CORRETA.

a. Valor p.

b. Valor esperado.

c. Frequência observada.

d. Proporção observada.

e. Correlação.

Referências bibliográficas

BUSSAB, Wilton.; MORETTIN, Pedro A. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017. 554p.

COSTA, G. G. de O. **Curso de estatística inferencial e probabilidades:** teoria e prática. São Paulo: Atlas, 2012.

SIQUEIRA, A. L.; TIBÚRCIO, J. D. **Estatística na área da saúde**: conceitos, metodologia, aplicações e prática computacional. Belo Horizonte: Coopmed, 2011.

MORETTIN, P. A. **Estatística básica**. Disponível em: https://www.ime.usp.br/~pam/EstBas.html. Acesso em: 24 ago. 2019.

Gabarito

Questão 1 – Resposta: C

Resolução: O Coeficiente de correlação é uma das medidas estatísticas que verifica a existência de correlação entre duas variáveis quantitativas.

Feedback de reforço: A matriz de correlação é uma matriz que apresenta dados de coeficientes de correlação para variáveis quantitativas de um conjunto de dados. Esta mede a relação linear entre duas variáveis quantitativas.

Questão 2 – Resposta: B

Resolução: O pacote readxl foi elaborado para realizar a importação de dados de planilhas elestrônicas para o R.

Feedback de reforço: Em geral, as planilhas utilizadas para armazenamento de conjuntos de dados são elaboradas em planilha MS Excel. Por isso, o pacote foi criado com o nome readxl.

Questão 3 – Resposta: A

Resolução: O valor p é um resultado que acompanha os testes estatísticos. Ele serve para avaliar a constatação de evidências de associação entre variáveis qualitativas.

Feedback de reforço: Todo resultado de teste estatístico vem acompanhado com um valor p, para constatar ou refutar uma hipótese que esteja sendo testada.

Elaborando gráficos estatísticos com o R

Autor: Marcelo Tavares de Lima

Objetivos

- Apresentar os principais gráficos estatísticos.
- Apresentar os principais comandos da linguagem R para elaboração de gráficos estatísticos.
- Desenvolver exemplos de aplicação para elaboração de gráficos na linguagem R.

1. Introdução

Os recursos visuais estatísticos, ou seja, os gráficos, são ferramentas de extrema importância para apresentar resultados de pesquisas, resultados de trabalhos corporativos, dentre outros. São recursos bastante intuitivos e de amplo alcance. Por amplo alcance, deseja-se afirmar que são recursos que ajudam a entender os resultados até mesmo pelo mais leigo no assunto ali tratado.

Bussab e Morettin (2017, p. 6) afirmam que "os métodos gráficos têm encontrado um uso cada vez maior devido ao seu forte apelo visual". O uso desse recurso, em geral, é mais fácil de ser compreendido quando comparado, por exemplo, a informações contidas em tabelas ou resumos numéricos.

Chambers *et al*. (1983 *apud* Bussab e Morettin, 2017, p. 6) afirmam que gráficos são utilizados para diversos fins, tais como:

> (a) buscar padrões e relações; (b) confirmar (ou não) certas expectativas que se tinha sobre os dados; (c) descobrir novos fenômenos; confirmar (ou não) suposições feitas sobre os procedimentos estatísticos usados; (e) apresentar resultados de modo mais rápido e fácil.

Com o avanço dos recursos computacionais e tecnológicos, o uso de métodos gráficos tem se tornado cada vez mais frequente no processo de análise de dados e na tomada de decisão.

Nesta leitura você será apresentado aos principais comandos da linguagem R elaborados para construir gráficos diversos, desde gráficos mais simples unidimensionais até gráficos mais complexos e multidimensionais. Desejamos que você tenha um excelente momento de estudo!

2. Gráficos estatísticos com a linguagem R

A produção de gráficos depende diretamente do tipo de variável utilizada, isto é, existem gráficos apropriados para cada tipo de variável existente, como, por exemplo, existem gráficos apropriados para variáveis categóricas e quantitativas. Detalhando um pouco mais, existem gráficos apropriados para variáveis categóricas nominais e categóricas ordinais e, também, existem gráficos apropriados para variáveis quantitativas discretas e quantitativas contínuas.

A produção de gráficos também depende diretamente da quantidade de variáveis envolvidas na sua elaboração. Quanto maior a quantidade de variáveis envolvidas no processo, mais complexo será o gráfico produto final. Por exemplo, um histograma é um gráfico apropriado para representar uma variável quantitativa contínua, enquanto um gráfico de colunas pode representar uma variável qualitativa (categórica) ou mais de uma variável categórica combinada com uma variável quantitativa contínua. Gráficos de linha são apropriados para representar variáveis quantitativas contínuas ao longo do tempo, ou seja, uma série temporal.

A produção de gráficos com a linguagem R tem vantagens em relação ao uso de outros tipos de programas computacionais porque, além de, ser uma linguagem de código aberto, ou seja, é possível modificar a programação, também tem recursos que ajudam a elaborar gráficos complexos e elegantes.

O uso da informação visual (gráfica) é importante porque facilita a transmissão da informação e abrange um público maior, ou seja, é mais fácil e de maior alcance. Falar em facilidade de transmissão de informação é falar de simplificar a divulgação de resultados encontrados em análise de dados.

A facilitação de apresentação de resultados pode ser considerada um atributo de grande interesse para quem trabalha com análise de dados, pois isso aumenta a atratividade em relação ao público-alvo, além de melhorar a comunicação.

O propósito deste texto é apresentar os principais comandos da linguagem R que possuem como produto recursos visuais. Serão apresentadas linhas de comando, assim como os resultados obtidos com a sua execução.

Para iniciar, começaremos produzindo gráficos com apenas uma variável e daremos início com a elaboração de gráficos apropriados para uma variável qualitativa. Considere os dados de Siqueira e Tibúrcio (2011), disponibilizados no portal da editora cujo endereço de internet se encontra nas referências deste texto, referentes ao estudo de prevalência de toxoplasmose, doença infecciosa, congênita ou adquirida, causada pelo protozoário *Toxoplasma gondii*. A prevalência dessa doença varia ao redor do mundo, dependendo diretamente do clima e dos hábitos da população, mas é muito frequente em regiões tropicais.

O banco de dados de Siqueira e Tibúrcio (2011) é composto por 278 crianças, compreendendo diversas variáveis, dentre elas a variável qualitativa que se refere ao motivo da realização de exame de sangue (qualitativa nominal).

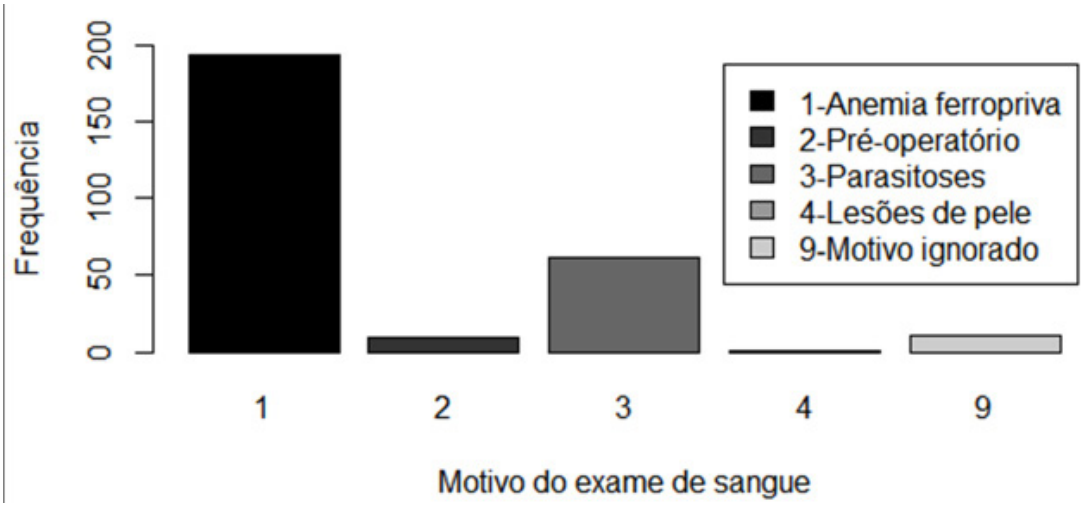
Para a execução dos comandos é necessário baixar os dados no seu computador e, em seguida, fazer a importação para o R através da IDE RStudio via menu de opções *File −> Import Dataset −> From Text (base)...* Selecionar a pasta onde os dados estão armazenados −> Import.

O banco de dados “Toxo” tem inúmeras variáveis, no entanto, iremos utilizar a variável referente ao motivo da realização de exame de sangue para produzir um gráfico de colunas. A programação em linguagem R para obter um gráfico de colunas verticais (gráficos de colunas) representando a frequência absoluta deverá ser executada conforme os comandos a seguir (apresentado com letras de fonte distinta para destacar que são linhas de programação).

```
barplot(table(Exame),xlab="Motivo do exame de sangue", ylab="Frequência", ylim=c(0,200),
legend.text=c("1-Anemia ferropriva", "2-Pré-operatório", "3-Parasitoses","4-Lesões de pele",
"9-Motivo ignorado"),col=gray(0:5/5))
```

O resultado da programação apresentada está mostrado na Figura 1.

Figura 1 – Gráfico de barras elaborado com linguagem R



Fonte: adaptada de SIQUEIRA e TIBÚRCIO (2011).

O gráfico de composição de setores, também conhecido como gráfico de pizza, é apropriado para representar a composição de uma variável qualitativa em porcentagem. Sua composição consiste em um círculo de raio arbitrário, representado o total da distribuição de valores, dividido em setores, correspondendo às partes de forma proporcional. Para exemplificar, considere os dados de escolaridade apresentados por Bussab e Morettin (2017) descritos na Tabela 1 e na base de dados "tab2_1", disponibilizada pelo professor Pedro Morettin (MORETTIN, 2019).

Tabela 1 – Escolaridade de pessoas de uma amostra

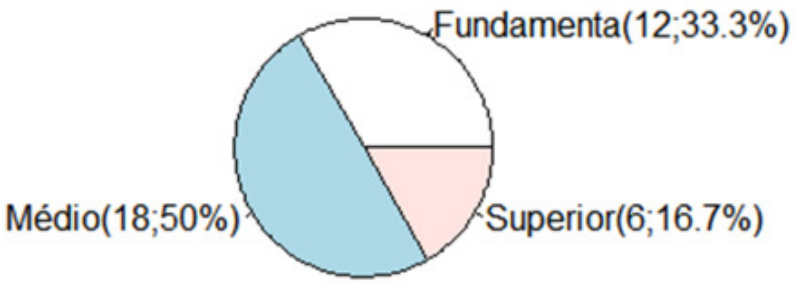
Grau de instrução	Frequência	Porcentagem
Fundamental	12	33,33
Médio	18	50,00
Superior	6	16,67
Total	36	100,00

Fonte: adaptada de BUSSAB e MORETTIN (2017).

Os comandos de linguagem R utilizados são apresentados a seguir e o gráfico gerado apresentado na Figura 2. Lembre-se que tudo que estiver após o sinal de cerquilha (#) não é executado pelo R, pois é apenas comentário para documentar a programação. Vale lembrar que os resultados do R são impressos com número em notação internacional, ou seja, faz uso de ponto (.) para separar dígitos decimais da parte inteira de um número e utiliza vírgula (,) para separar casa de milhares, milhões, etc.

```
# é separada a variável escolaridade do banco de dados tab2_1 para facilitar a execução
dos comandos
escolaridade<-tab2_1$grau_instrução
# é criada a tabela para a variável escolaridade
x<-table(escolaridade)
# são criados rótulos para as categorias da variável escolaridade
rotulos<-paste(c("Fundamental","Médio","Superior"),"(",c(12,18,6),";",round(c(33.33,50,16.67),
1),"%)",sep="")
# é criado o gráfico de pizzas com os argumentos de rotulação e legenda
pie(x,labels=rotulos)
```

Figura 2 – Gráfico de setores para uma variável qualitativa (Escolaridade, n = 36 pessoas)



Fonte: adaptada de BUSSAB e MORETTIN (2017).

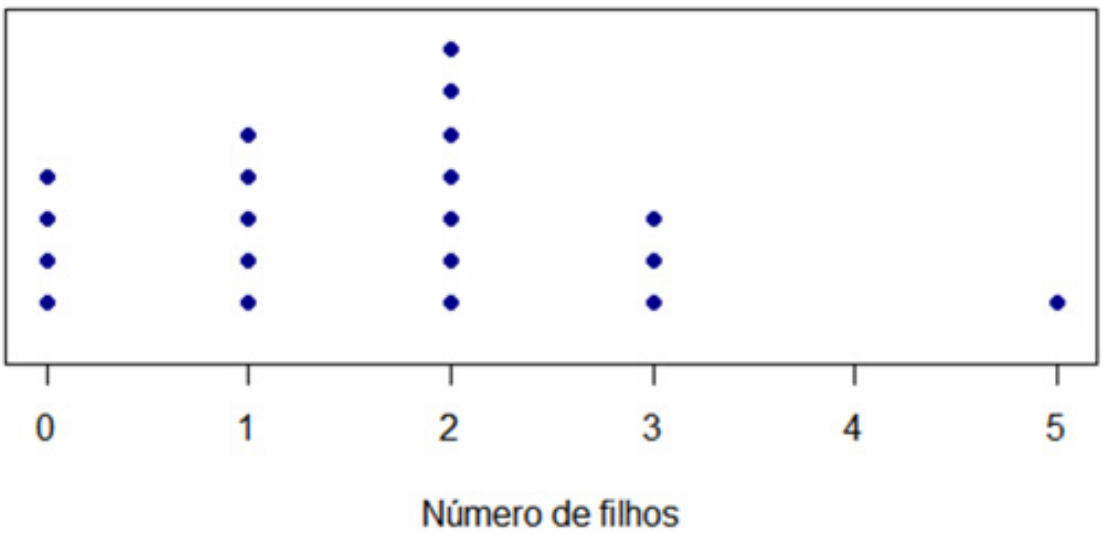
Para representar uma variável quantitativa discreta existe um gráfico chamado de dispersão unidimensional. É um gráfico representado por pontos ao longo de uma reta (com uma escala apropriada).

Os valores observados da variável são representados por pontos ao longo da reta e, quando repetidos, são empilhados em cima do valor que se repete. A programação a seguir produz um gráfico de dispersão unidimensional e sua visualização é apresentada na Figura 3.

```
# tabela de frequência de valores da variável para verificar as frequências
table(tab2_1$n_filhos)
#resultado
0 1 2 3 5 # valores da variável
4 5 7 3 1 # frequência dos valores da variável

stripchart(tab2_1$n_filhos, # base de dados
        method = "stack", # elementos empilhados
        offset = 1, # espaçamento entre os elementos
        pch = 19, # formato do elemento (19 = circular)
        col="darkblue", # cor do elemento
        ylim=c(0,7), # eixo vertical
        xlab="Número de filhos", # rótulo horizontal
        cex=1 # tamanho dos elementos.
        )
```

Figura 3 – Gráfico de dispersão unidimensional para variável quantitativa discreta (número de filhos, n = 36)



Fonte: adaptada de BUSSAB e MORETTIN (2017).

Para variável quantitativa contínua, como, por exemplo, a idade das pessoas, um possível gráfico apropriado trata-se do histograma, que é um gráfico de colunas contíguas em que as bases das colunas representam faixas de valores para a variável que está sendo representada, e os valores do eixo vertical ou de ordenadas representam a frequência, absoluta ou relativa, ou ainda a densidade de frequência (BUSSAB; MORETTIN, 2017).

Para exemplificar a construção de um histograma com a linguagem R, considere a variável idade (em anos) contida no banco de dados "tab2_1" disponibilizado por Bussab e Morettin (2017). A programação a seguir ajuda a produzir um histograma com valores de idade (em anos) agrupados em faixas de valores com comprimento de 5 anos, iniciando no menor valor de idade encontrado no banco de dados, que é 20 anos, e a partir daí acrescentando 5 anos até o valor máximo da idade.

```
# Para verificar um resumo da variável para descobrir o valor mínimo e máximo da idade.

summary(tab2_1$idade_anos)
  Min.   1st Qu. Median   Mean   3rd Qu. Max.
  20.00   30.00   34.50   34.58   40.00   48.00

# Para construir um histograma com a variável idade (em

# anos) do banco de dados tab2_1, com alguns parâmetros

# determinados, como título do eixo horizontal e vertical,

# valores de idade divididos em classe de valores de

# amplitude igual a cinco, sem título principal no gráfico

# com valores de frequência absoluta no eixo vertical e

# variando de 0 a 15 e rótulos em cada coluna e

# colunas de cor azul claro.

hist(tab2_1$idade_anos,xlab='Idade (em anos)', ylab='Frequência', main='', col='lightblue', breaks=seq(20,50, by=5), labels=TRUE,ylim=c(0,15))
```

O gráfico resultante está apresentado na Figura 4.

Figura 4 – Histograma para a idade em anos (n=36)

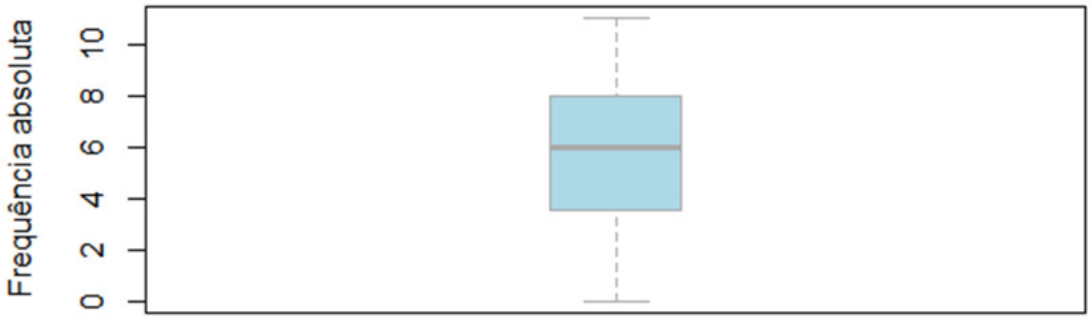
Fonte: adaptada de BUSSAB e MORETTIN (2017).

Continuando a elaboração de gráficos para variáveis quantitativas contínuas, existe um gráfico que resume uma série de medidas estatísticas em sua visualização, como valor mínimo e máximo da variável, valor da mediana, valor do primeiro quartil e terceiro quartil e identificação de valores atípicos, conhecidos como *outliers* (Bussab e Morettin, 2017). Trata-se do diagrama de caixa ou *boxplot*.

Ainda utilizando o banco de dados “tab2_1”, vamos utilizar a idade em meses para produzir um *boxplot* com os seus valores. A programação elaborada está apresentada a seguir, assim como o resultado gráfico da sua execução na Figura 5.

```
boxplot(tab2_1$idade_meses,
    pch="*",  # tipo de marcador dos outliers (só aparece se tiver um valor identificado como outlier
    col="lightblue", # cor do preenchimento do boxplot
    border="darkgrey", # cor da linha do box plot
    boxwex=0.3, # Tamanho da caixa
    ylab="Frequência absoluta" # rótulo eixo vertical
    )
```

Figura 5 – boxplot para a variável idade em meses (n = 36)



Fonte: adaptada de BUSSAB e MORETTIN (2017).

A caixa apresentada no *boxplot* representa o valor do primeiro quartil na sua base, o valor da mediana com a linha central e o valor do terceiro quartil na sua parte superior. As linhas que saem da caixa representam a distribuição dos valores da variável até os pontos $LI = q_1 - 1{,}5d_q$ para baixo e $LS = q_3 - 1{,}5d_q$ para cima. Os valores compreendidos entre esses dois limites são conhecidos como valores adjacentes e os valores que estiverem abaixo de LI e acima de LS são os chamados valores atípicos ou *outliers*. Os elementos q_1 e q_3 representam o primeiro e o terceiro quartil, respectivamente. Já o elemento d_q representa a distância interquartil, obtida por $d_q = q_3 - q_1$.

A utilidade de um *boxplot* é de dar uma ideia de distribuição dos valores da variável, como posição, dispersão, assimetria e valores atípicos. A posição central da distribuição é dada pela mediana e a dispersão pela distância interquartil (d_q).

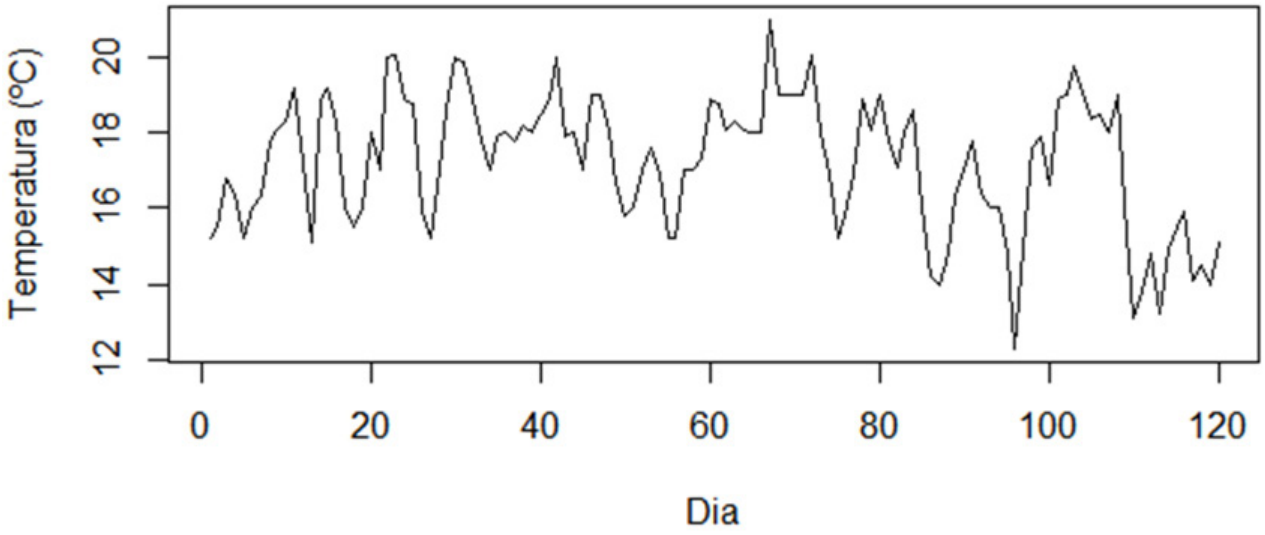
Quando os dados de uma variável são coletados ao longo do tempo, seja variável quantitativa, tem-se o que se chama de série temporal, ou seja, dados observados em instantes ordenados do tempo. Segundo Bussab e Morettin (2017, p. 27), “espera-se que exista relação entre as observações em instantes de tempo diferentes”.

Para exemplificar, considere a variável temperatura do banco de dados “cd_poluição” disponibilizado pelo Professor Pedro Morettin e, também, no seu livro publicado com o Professor Wilton Bussab (MORETTIN;

BUSSAB, 2017). Segundo os autores, os dados são referentes a temperaturas observadas na cidade de São Paulo, no período de 1º de janeiro a 30 de abril de 1991 (n = 120). O gráfico da série temporal referente a essas temperaturas é elaborado conforme *script* em linguagem R apresentado a seguir e seu produto gráfico é apresentado na Figura 6.

```
plot.ts(cd_poluicao$temp, xlab="Dia",ylab="Temperatura")
```

Figura 6 – Série temporal para os dados de temperatura da cidade de São Paulo, janeiro a abril de 1991 (n = 120)



Fonte: adaptada de BUSSAB e MORETTIN (2017).

PARA SABER MAIS

Quando se trabalha com poucos dados, é possível inseri-los diretamente no R com a criação de vetores, matrizes ou data frames de dados, conforme apresentado na Leitura Fundamental 3. Portanto, é importante conhecer como se faz a inserção de dados via esses objetos, pois, muitas vezes, a inserção de dados dessa forma é mais rápida do que qualquer outra.

Os gráficos até aqui apresentados foram produtos do tratamento de uma única variável. No entanto, há situações em que se deseja produzir gráficos com mais de uma variável: duas, três ou mais. É claro que é preciso ter um certo cuidado com isso, pois, senão, em vez de uma imagem facilitar a comunicação de resultados, pode complicar e tornar mais dificultosa.

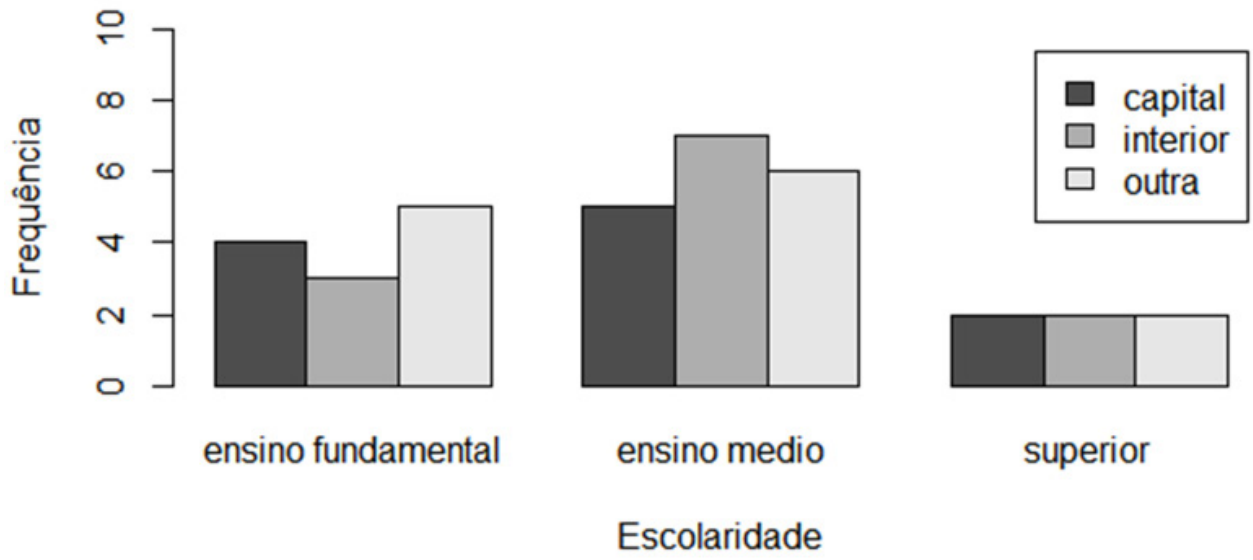
Iremos considerar, então, a produção de gráficos com mais de uma variável e, para dar início, vamos apresentar visualizações com a informação de duas variáveis qualitativas.

Continuando a utilizar a base de dados “tab2_1” e utilizando as variáveis de escolaridade e região de procedência, duas variáveis qualitativas, a primeira é ordinal e a segunda é nominal, é possível elaborar um gráfico de colunas conforme mostram as linhas de comando a seguir.

```
attach(tab2_1) # anexar o banco de dados
tabela<-table(reg_procedencia,grau_instrucao)
tabela # verificar as frequências da tabela
                grau_instrucao
reg_procedencia ensino fundamental ensino mÃ©dio superior
      capital                 4          5       2
      interior                3          7       2
      outra                   5          6       2
# construção do gráfico
barplot(tabela, xlab='Escolaridade', ylab='Frequência', ylim=c(0,10),legend=TRUE,
beside=TRUE)
```

O gráfico resultante é apresentado na Figura 7.

Figura 7 – Gráfico de barras verticais para duas variáveis qualitativas (n=36)

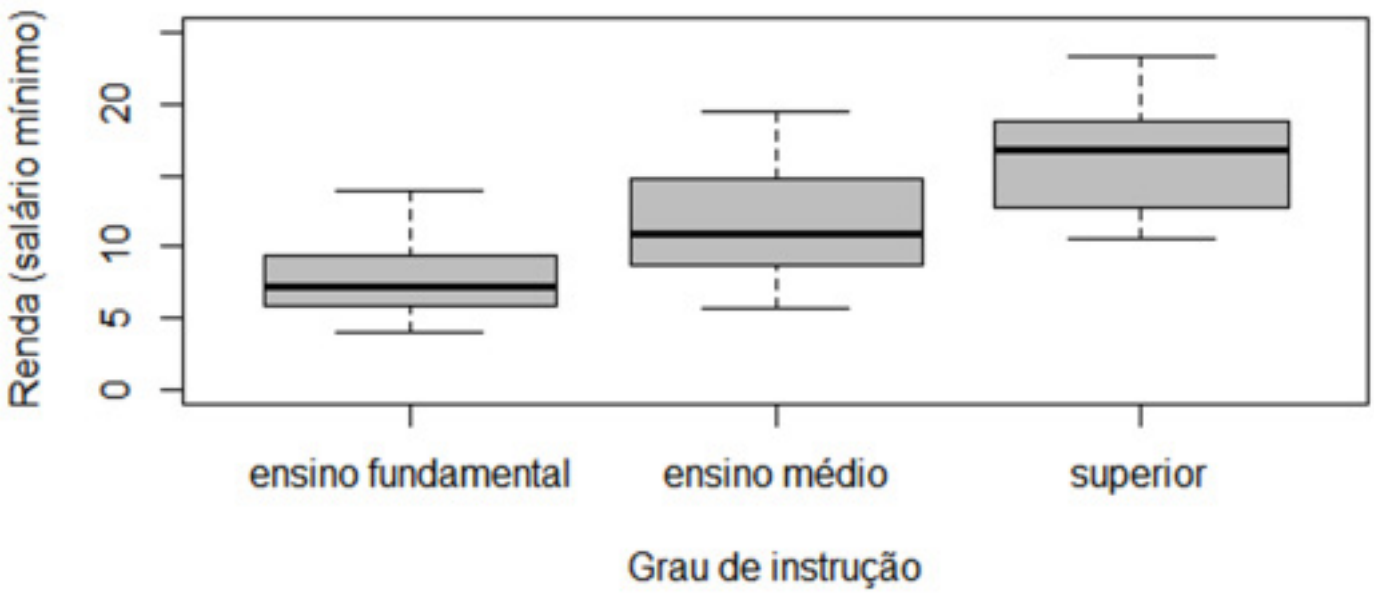


Fonte: adaptada de BUSSAB e MORETTIN (2017)

Quando se tem informações de duas variáveis, onde uma é qualitativa e a outra é quantitativa, um gráfico apropriado para representar a relação entre as duas variáveis é um diagrama de caixas ou *boxplot*. A programação em linguagem R apresentada a seguir produz um gráfico desse tipo com as variáveis salário (em salários mínimos) e o grau de instrução dos dados "tab2_1". A Figura 8 apresenta o gráfico resultante.

```
boxplot(tab2_1$salario~tab2_1$grau_instrucao,
    xlab="Grau de instrução", # título eixo horizontal
    ylab="Renda (salário mínimo)", # título eixo vertical
    col="grey", # cor das caixas
    ylim=c(0,25)) # faixa de valores do eixo vertical
```

Figura 8 – Diagrama de caixas (*boxplot*)



Fonte: elaborada pelo autor.

A utilidade do uso do um *boxplot* para analisar relação entre duas variáveis distintas em sua natureza, uma quantitativa e outra qualitativa, é a possibilidade de identificar padrões da relação, como, por exemplo, o Gráfico 8. Com ele é possível observar nitidamente que quanto maior o grau de instrução, maior o salário.

Quando se deseja avaliar a relação entre duas variáveis quantitativas, pode-se utilizar o diagrama de dispersão, gráfico apropriado para identificar padrões e tendências que possam existir entre variáveis dessa natureza. A programação em linguagem R, que produz de forma simples um gráfico de dispersão entre duas variáveis, é apresentada a seguir. O produto gráfico é mostrado na Figura 9.

```
plot(tab2_1$idade_anos, # variável do eixo horizontal
    tab2_1$salario, # variável do eixo vertical
    ylim = c(0,20), # faixa de valores do eixo y
    xlim=c(0,50), # faixa de valores do eixo x
    pch=16, # tipo de elemento gráfico para (x,y)
    col="darkblue", # cor dos elementos gráficos
    xlab = "Idade em anos", # rótulo do eixo x
    ylab = "Salário" # rótulo do eixo y
    )
```

Figura 9 – Diagrama de dispersão (n = 36)

Fonte: elaborada pelo autor.

Para os dados utilizados é possível perceber uma leve tendência ao crescimento do salário quando a idade aumenta. Essa observação indica que existe uma correlação positiva entre essas duas variáveis.

ASSIMILE

A elaboração de gráficos está diretamente relacionada com o fato de se desejar contar algo, uma história, por exemplo, para um potencial observador ou público-alvo. Este recurso metodológico, além de estar atrelado aos métodos estatísticos, também está relacionado com a ciência de dados, pois ambos, quando fazem uso de recursos visuais, têm como propósito o resumo de resultados e a comunicação dos mesmos.

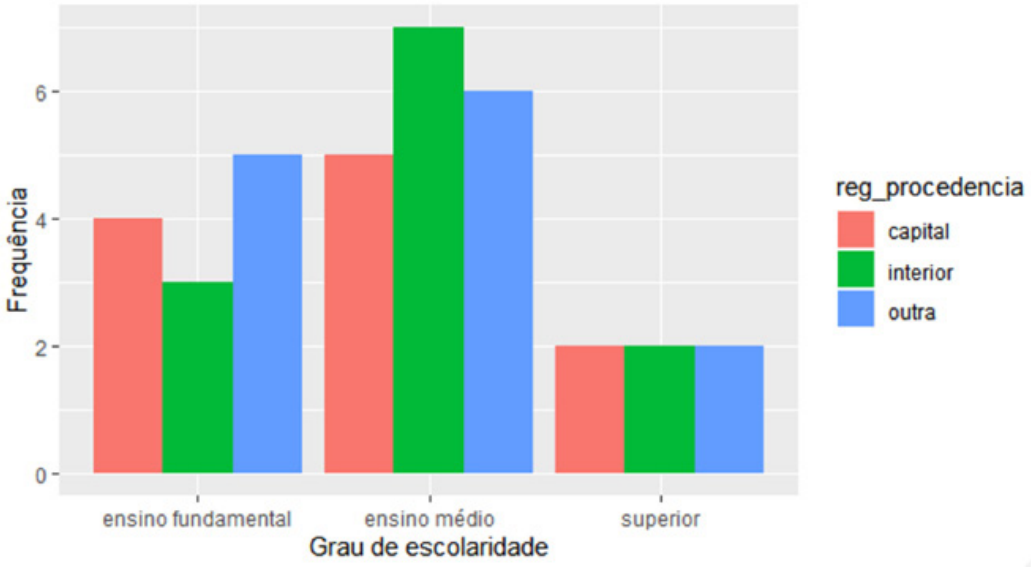
Existem gráficos mais complexos, gráficos com três ou mais variáveis, que apresentam no mesmo plano mais de um tipo de visualização. São combinações de gráficos, matrizes de correlação, dentre outros.

Existe uma biblioteca no R chamada ggplot2, que possui uma infinidade de recursos para elaboração de gráficos de diversos tipos. A vantagem que ela possui em relação aos comandos básicos de produção de gráficos é a produção de gráficos complexos e visualmente mais elegantes. Observe a programação em linguagem R a seguir para a elaboração de um gráfico de barras com informações de duas variáveis qualitativas.

```
library(ggplot2)
g<-ggplot(data = tab2_1)
g<-ggplot(data = tab2_1) +
  geom_bar(
    mapping = aes(x = grau_instrucao, # variável
                  fill = reg_procedencia), # preenchimento
    position = “dodge” # colunas lado a lado
  )
# Rótulos(títulos)
g <- g +
  labs(y = ‘Frequência’, # rótulo vertical
       x = ‘Grau de escolaridade’ # rótulo horizontal
  )
g # impressão do gráfico
```

O resultado gráfico é apresentado na Figura 10.

Figura 10 – Gráfico de barras com duas informações de variáveis qualitativas (grau de instrução e região de procedência) (n = 36)



Fonte: elaborada pelo autor.

É possível perceber que com o uso da biblioteca surgem mais recursos para produção visual, melhores gráficos, mais elegantes e, muitas vezes, mais informativos. Portanto, vale a pena buscar conhecimento sobre a biblioteca. Mais detalhes sobre a biblioteca ggplot2 podem ser encontrados em Oliveira, Guerra e Mcdonnell (2018).

Este texto apresentou recursos de programação em linguagem R para produção de gráficos estatísticos. É certo que o assunto não se esgota aqui, no entanto, pode ser um início do estudo da produção de visualização de dados, pois muito ainda existe a ser conhecido e explorado.

TEORIA EM PRÁTICA

Considere que você trabalha em uma empresa de pesquisa de mercado e precisa analisar o perfil de potenciais clientes de um determinado produto. Para isso, você realiza uma coleta de dados para buscar informações que possam subsidiar seu trabalho.

O conjunto de dados utilizado foi retirado de Costa (2012) e se refere a uma pesquisa de perfil demográfico feito a 20 consumidores adultos do produto X. O banco de dados é apresentado na Tabela 1.

Tabela 1 – Conjunto de dados

Sexo	Idade	Escolaridade	N_filhos	Classe_social
M	35	Ens. Medio	2	B
M	25	Ens. Medio	1	B
F	40	Ens. Superior	1	C
M	25	Ens. Medio	3	B
M	32	Ens. Medio	2	C
F	22	Ens. Medio	0	C
M	37	Ens. Superior	2	B
M	28	Ens. Medio	0	B
F	25	Ens. Medio	1	B
F	39	Ens. Superior	2	C
M	35	Ens. Fundamental	1	B
F	21	Ens. Fundamental	0	A
F	27	NA	0	A
F	45	Ens. Medio	2	C
M	57	Pos-Graduacao	4	C
F	33	Ens. Medio	2	A
M	36	Ens. Fundamental	0	B
M	35	Ens. Medio	2	C
M	33	Ens. Medio	2	B
F	22	Ens. Superior	0	C

Fonte: adaptada de Costa (2011).

Sexo: M: masculino; F: feminino.
Idade: em anos com dois dígitos.
Escolaridade: NA: sem informação.
Classe social: A: alta; B: média; C: baixa.
O conjunto de dados está disponibilizado em planilha MS Excel e você deseja exportar para o RStudio para fazer a análise dos dados com a linguagem R. Portanto, para fazer isso, deve utilizar a seguinte programação em R.

```
library(readxl)
> dados <- read_excel("dados.xlsx",
+     col_types = c("text", "numeric", "text",
+         "numeric", "text"), na = "NA")
> View(dados)
```

Também é possível realizar a importação com os menus disponíveis no RStudio. A planilha se chama "dados" e está armazenada em alguma pasta no seu computador. Para fazer a exportação, precisa carregar o pacote "readxl" para utilizar a função "read_excel".

Agora, você precisa utilizar os conhecimentos que tem em produção de gráficos estatísticos com a linguagem R para produzir visualizações que vão compor um relatório e apresentar para o seu grupo de trabalho. Bom trabalho!

VERIFICAÇÃO DE LEITURA

1. Quando se deseja avaliar a relação entre duas variáveis quantitativas, pode-se recorrer a um recurso gráfico. Qual o nome do recurso gráfico apropriado para realizar essa verificação?
   Assinale a alternativa CORRETA.

   a. Gráfico de barras.

   b. Gráfico de linhas.

c. Boxplot.

d. Diagrama de dispersão.

e. Gráfico de colunas.

2. Histograma é um tipo de gráfico estatístico apropriado para qual tipo de variável?
Assinale a alternativa CORRETA.

a. Variável quantitativa discreta.

b. Variável quantitativa contínua.

c. Variável qualitativa nominal.

d. Variável qualitativa ordinal.

e. Qualquer tipo de variável.

3. A elaboração de gráficos mais simples com linguagem R faz uso de funções básicas, no entanto, para elaborar gráficos com mais recursos e mais bem elaborados, existe uma biblioteca (pacote) apropriada. Qual o nome desta biblioteca? Assinale a alternativa CORRETA.

a. barplot.

b. boxplot.

c. ggplot2.

d. plot.ts.

e. table.

Referências bibliográficas

BUSSAB, Wilton.; MORETTIN, Pedro A. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017. 554p.

COSTA, G. G. de O. **Curso de estatística inferencial e probabilidades:** teoria e prática. São Paulo: Atlas, 2012.

MORETTIN, P. A. **Estatística básica**. Disponível em: https://www.ime.usp.br/~pam/EstBas.html. Acesso em: 24 ago. 2019.

OLIVEIRA, P. F.; GUERRA, S.; McDONNELL, R. **Ciência de dados com R:** introdução. Brasília: IBPAD. 2018. Disponível em: https://www.ibpad.com.br/o-que-fazemos/publicacoes/introducao-ciencia-de-dados-com-r#download. Acesso em: 5 set. 2019.

SIQUEIRA, A. L., TIBÚRCIO, J. D. **Estatística na área da saúde**: conceitos, metodologia, aplicações e prática computacional. Belo Horizonte: Coopmed, 2011. Disponível em: https://www.coopmed.com.br/index.php/estatistica-na-area-da-saude.html. Acesso em: 5 set. 2019.

Gabarito

Questão 1 – Resposta: D

Resolução: Quando se deseja avaliar a relação entre duas variáveis quantitativas com um recurso visual, pode-se utilizar um diagrama de dispersão.

Feedback de reforço: Lembre-se dos gráficos apropriados para variáveis quantitativas.

Questão 2 – Resposta: B

Resolução: O histograma é um recurso gráfico apropriado para representar uma variável quantitativa contínua.

Feedback de reforço: Histograma é um gráfico de colunas justapostas para representar variáveis quantitativas contínuas.

Questão 3 – Resposta: C

Resolução: Para elaborar gráficos com mais recursos visuais, pode-se utilizar a biblioteca ggplot2, a qual possui muito mais possibilidade que as funções básicas da linguagem R para elaboração de gráficos.

Feedback de reforço: Lembre-se do nome da biblioteca apresentada e exemplificado o seu uso na leitura fundamental para elaboração de gráficos mais sofisticados.

Junção de bancos de dados e sumarização estatística usando R

Autor: Marcelo Tavares de Lima

Objetivos

- Descrever sobre bancos de dados em linguagem R.
- Apresentar formas de junção de bancos de dados.
- Apresentar como se faz uma sumarização estatística de bancos de dados.

1. Introdução

Em um processo de análise de dados, a primeira coisa que se deve fazer é criar um conjunto de dados com informações que serão utilizadas no processo de análise e geração de visualização no formato apropriado do programa computacional que será utilizado para a realização deste trabalho.

Em linguagem R é possível realizar a criação de um conjunto de dados de diversas maneiras. Esta é uma das principais características desta linguagem de programação, a versatilidade em realizar a mesma tarefa de diversas maneiras.

Neste texto você aprenderá a criar e a manipular bases de dados em linguagem R, assim como realizar resumo ou sumarização de variáveis contidas nos bancos de dados em tratamento estatístico. Desejamos que você possa ter um excelente momento de aprendizagem!

2. Bancos de dados no R

Como tarefa inicial, a criação de um banco de dados em linguagem R, segundo Kabacoff (2015), está relacionada diretamente com a escolha da estrutura de dados para organizá-los e com a importação dos mesmos para este ambiente.

A inserção de dados no ambiente R pode ser feita de forma manual ou importada de uma fonte externa. Tais fontes externas podem incluir arquivos de texto, planilhas, outros programas estatísticos e dados de outros sistemas de gerenciamento de dados.

Um conjunto de dados é um arranjo retangular com linhas representando observações ou registros e colunas representando variáveis. A Tabela 1, adaptada de Kabacoff (2015), apresenta um simples exemplo de um conjunto de dados.

Tabela 1 – Conjunto de dados de pacientes (fictício)

IDPaciente	DataAdmissao	Idade	Diabetes	Classe
1	15/10/2014	25	Tipo1	Pobre
2	01/11/2014	34	Tipo2	Media
3	21/10/2014	28	Tipo1	Alta
4	28/10/2014	52	Tipo1	Pobre

Fonte: adaptada de KABACOFF (2015).

As diferentes áreas de estudo (computação, estatística, ciência de dados) denominam os campos de um conjunto de dados, muitas vezes, de nomes diferentes. Por exemplo, os profissionais da estatística se referem a observações e variáveis, enquanto que profissionais da computação se referem a registros e campos. Neste texto, vamos nos referir, principalmente, aos termos utilizados pelos estatísticos, por se tratar de uma proposta de sumarização estatística.

O conteúdo de um conjunto de dados pode ser muito diversificado. Ele pode incluir tanto variáveis quantitativas quanto qualitativas, como, por exemplo, a variável idade do conjunto de dados de pacientes (Tabela 1) é uma variável quantitativa, enquanto que a variável diabetes é uma variável qualitativa.

A linguagem R possui uma variedade de estruturas de dados, incluindo escalares, vetores, *arrays*, *data frames* e listas. A Tabela 1, por exemplo, é um exemplo de *data frame* em linguagem R. Como dito antes, a diversidade de estruturas de dados é uma das principais características da linguagem R, o que proporciona flexibilidade em lidar com dados.

Um *data frame* é uma estrutura em ambiente R que suporta conjuntos de dados de forma semelhante aos conjuntos de dados encontrados em programas estatísticos como o SAS (*Statistical Analysis System*), SPSS (*Statistical Package for the Social Sciences*) e STATA, por exemplo.

Sua vantagem é permitir diferentes tipos de variáveis no mesmo conjunto de dados.

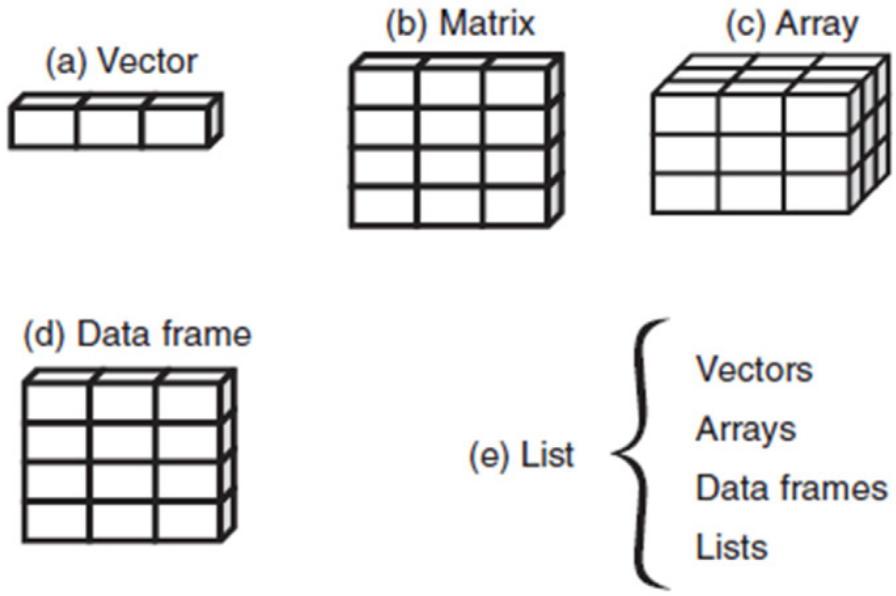
Os tipos de dados reconhecidos pela linguagem R são numérico, caracter, lógico (Verdadeiro/Falso, complexos (números imaginários) e *bytes* (KABACOFF, 2015). Por exemplo, os dados da Tabela 1, IDPaciente, DataAdmissao e Idade são variáveis numéricas, enquanto que Diabetes e Classe são variáveis caracter.

Para complementar, é importante dizer à linguagem R, por exemplo, no caso da Tabela 1, que IDPaciente é um identificador, que DataAdmissao contém datas e que, Diabetes e Classe são variáveis categóricas nominal e ordinal, respectivamente. Variáveis identificadoras, na linguagem R, são denominadas *rownames*, enquanto que variáveis categóricas (nominal e ordinal) são denominadas *factors*.

Na linguagem R, denomina-se objeto qualquer elemento que possa ser atribuído a uma variável (KABACOFF, 2015). Isto inclui constantes, estruturas de dados, funções e gráficos. Um objeto tem duas principais características, o modo (descrição de como os objetos são armazenados) e uma classe (designa como o seu conteúdo pode ser manipulado pelas funções).

A diferença nos tipos objetos reconhecidos pela linguagem R ocorre em termos dos tipos de dados que podem armazenar, na maneira como são criados, a complexidade estrutural e a notação utilizada para identificar os elementos. A Figura 1 apresenta os esquemas de estruturas de dados reconhecidos pelo R.

Figura 1 – Estruturas de dados reconhecidas no R



Fonte: adaptada de KABACOFF (2015).

Vetores são arranjos unidimensionais que podem ser compostos por dados numéricos, caracter ou string e dados lógicos. A função c( ) é utilizada para criar um vetor em linguagem R. Para exemplificar, considere os comandos apresentados a seguir para a construção de vetores em linguagem R que podem ser digitados tanto no editor de códigos quanto no console do RStudio.

```
a <− c(1, 2, 5, 3, 6, -2, 4)
b <− c(“um”, “dois”, “três”)
c <− c(TRUE, TRUE, TRUE, FALSE)
```

O vetor a é um vetor numérico, o vetor b é um vetor caracter e o vetor c é um vetor lógico. Observe que vetores só podem ter um tipo de dado no mesmo objeto. Vale lembrar que escalares são vetores de um elemento. Por exemplo, podemos criar o objeto f <− 3, que é um vetor numérico f composto pelo número três, g <− “BR” e h<− TRUE.

Matrizes são arranjos bidimensionais onde cada elemento é do mesmo tipo, ou seja, é numérico, caracter ou lógico. Matrizes são criadas com o comando matrix( ). A função genérica que cria uma matriz é apresentada a seguir.

matriz <− matrix(vetor, nrow=número de linhas, ncol=número de colunas, byrow=valor lógico, dimnames=list(nome das linhas, nome das colunas))onde, vetor contém os elementos da matriz, nrow e ncol especifica as dimensões das linhas e das colunas, dimnames contém os rótulos (opcionais) das linhas e colunas da matriz. A opção byrow indica se a matriz deve ser preenchida por linha (byrow=TRUE), o padrão é ser preenchida por coluna.

Arrays (arranjos) são como as matrizes, no entanto, possuem mais de duas dimensões. São criados com a função array e com os seguintes argumentos.

arranjo <− array(vetor, dimensões, dimnames)

onde vetor contém o conjunto de dados do array, dimensões é um vetor numérico que indexa cada dimensão e dimnames é uma lista opcional de rótulos das dimensões.

Data frame é um objeto mais genérico que uma matriz, de forma que diferentes colunas podem conter diferentes tipos de dados (numéricos, caracter, etc.). Como dito, é similar aos conjuntos de dados encontrados no SAS, SPSS e STATA. Os *data frame* são as estruturas de dados mais comumente trabalhadas em linguagem R. A programação em linguagem R que cria um *data frame* é apresentada a seguir.

dados <− data.frame(coluna1, coluna2, ...)

onde, coluna1, coluna2, ..., são vetores colunas de qualquer tipo (numérico, caracter, lógico).

Para exemplificar a construção de um *data frame*, consideremos os dados da Tabela 1. Suponha que estavam separados em vetores, conforme mostrado na programação a seguir.

```
IDPaciente <− c(1, 2, 3, 4)
Idade <− c(25, 34, 28, 52)
Diabetes <− c(“Tipo1”, “Tipo2”, “Tipo1”, “Tipo1”)
Classe <− c(“Pobre”, “Media”, “Alta”, “Pobre”)
```

Deseja-se criar um banco único com as variáveis. Para isso, basta utilizar a função data.frame, conforme apresentado a seguir.

```
pacientes <− data.frame(IDPaciente, Idade, Diabetes, Classe)
```

Vale lembrar que a linguagem R é sensível ao uso de letras maiúsculas e minúsculas (*case sensitive*). O resultado da união dos vetores será o banco de dados conforme mostrado a seguir.

```
pacientes # imprime o banco de dados criado
  IDpaciente Idade Diabetes Classe
1          1    25    Tipo1  Pobre
2          2    34    Tipo2  Media
3          3    28    Tipo1   Alta
4          4    52    Tipo1  Pobre
```

Cada coluna dever ser de um único tipo (numérica, caracter ou string), no entanto, um objeto do tipo *data frame* tem a vantagem de ter em seu conteúdo tipos variados de variáveis, como um banco de dados usual. Por conta dessa característica, as colunas de um *data frame* serão chamadas de variáveis, como sinônimo.

É possível manipular um *data frame* de muitas maneiras. Uma das mais simples está relacionada com a seleção de uma ou um subconjunto de variáveis do banco. Suponha que você deseja imprimir apenas as variáveis de identificação do paciente (IDPaciente) e a idade (Idade). Para isso, você precisa selecionar as colunas apropriadas do banco de dados que possuem as variáveis que deseja conforme linhas de programação a seguir.

```
paciente[1:2] # seleciona a primeira e a segunda colunas do data frame
# resultado
  IDPaciente Idade
1          1    25
2          2    34
3          3    28
4          4    52
```

Outra maneira de selecionar colunas/variáveis do *data frame*.

```
paciente[c("Diabetes", "Classe")] # seleciona as variáveis através de um vetor
  Diabetes Classe
1    Tipo1  Pobre
2    Tipo2  Media
3    Tipo1   Alta
4    Tipo1  Pobre
```

Para selecionar uma única variável de um *data frame*, pode-se utilizar o seguinte comando.

```
paciente$idade
paciente$Idade # seleciona a variável idade do data frame paciente
# Resultado
[1] 25 34 28 52
```

PARA SABER MAIS

Com os dados da Tabela 1 é possível criar um data frame no ambiente R para elaborar um conjunto de dados único para ser analisado com as funções estatísticas apropriadas de análise de dados.

3. Junção de bancos de dados e sumarização no R

No item anterior foram apresentados os principais tipos de objetos que a linguagem R reconhece em seu ambiente de trabalho. É possível observar que o objeto *data frame* é aquele que mais se assemelha com o que se conhece de conjunto de dados. Por isso, será apresentado com maiores detalhes nesta seção.

Antes de apresentarmos um *data frame* com maiores detalhes, iremos apresentar as funções cbind( ) e rbind( ), as quais realizam as tarefas de

juntar colunas e linhas, respectivamente. Segundo Aquino (2014, p. 32), elas “juntam vetores formando matrizes, ou seja, uma forma retangular de representação dos dados em que eles estão distribuídos em linhas e colunas”.

Para exemplificar, de forma muito simples, considere as linhas de comando apresentadas a seguir, extraídas de Aquino (2014).

```
x <− c(7, 9, 8, 10, 1)
y <− c(9, 8, 10, 9, 3)
z <− c(10, 9, 9, 9, 2)
matriz <− cbind(x, y, z) # junta os vetores x, y e z.
```

Como resultados dos comandos acima, tem-se a seguinte matriz.

```
     x  y  z
[1,]  7  9 10
[2,]  9  8  9
[3,]  8 10  9
[4,] 10  9  9
[5,]  1  3  2
```

Perceba que os vetores foram juntados como colunas de uma matriz. No entanto, se for utilizada a função rbind( ), o resultado seria distinto, conforme apresentado a seguir.

```
rbind(x, y, z)
  [,1] [,2] [,3] [,4] [,5]
x    7    9    8   10    1
y    9    8   10    9    3
z   10    9    9    9    2
```

Agora, os vetores x, y e z foram juntados como vetores linhas.

É possível associar rótulos, tanto às linhas quanto às colunas de uma matriz. Para exemplificar, considere a matriz criada com a função cbind(x, y, z) anteriormente onde iremos associar os seguintes rótulos, conforme linhas de comando a seguir.

```
m <- cbind(x, y, z) # atribui a matriz ao objeto m
# atribui nome às colunas da matriz
colnames(m) <- c("Matemática", "Português", "História")
#atribui nome às linhas da matriz
rownames(m) <- c("Helena", "José", "Maria", "Francisco", "Macunaíma")
m # imprime o objeto m
```

O resultado da impressão do objeto "m" é apresentado a seguir.

```
          Matemática Português História
Helena             7         9       10
José               9         8        9
Maria              8        10        9
Francisco         10         9        9
Macunaíma          1         3        2
```

O resumo de um conjunto de dados (objeto *data frame* no R) pode ser realizado de muitas maneiras. Essa é uma das vantagens da linguagem R. Para exemplificar, considere o uso da função summary( ), pertencente ao pacote básico de funções do R. Vamos utilizar o conjunto de dados apresentado na Tabela 1, que foi denominado de "paciente". Com as linhas de comando a seguir é possível obter um resumo do conjunto de dados.

```
# importação dos dados para o ambiente R
library(readxl) # carrega a biblioteca que importa
# importa o banco de dados para o objeto "paciente"
> paciente <- read_excel("C:/Tecnicas-Estatisticas/Tema6-TE/paciente.xlsx", col_types =
c("numeric", "date", "numeric", "text", "text"))
# solicita o sumário das variáveis
summary(paciente)
# resultado! Ops! Deu um problema!
  IDPaciente    DataAdmissao                  Idade        Diabetes
 Min.   :1.00   Min.   :2014-10-15 00:00:00   Min.: 25.00   Length: 4
 1st Qu.: 1.75   1st Qu.: 2014-10-19 12:00:00   1st Qu.: 27.25   Class: character
 Median: 2.50   Median: 2014-10-24 12:00:00   Median: 31.00   Mode: character
 Mean: 2.50   Mean: 2014-10-24 00:00:00   Mean: 34.75
 3rd Qu.: 3.25   3rd Qu.: 2014-10-29 00:00:00   3rd Qu.: 38.50
 Max.: 4.00   Max.: 2014-11-01 00:00:00   Max.: 52.00
   Classe
 Length: 4
 Class: character
 Mode: character
```

Observe que ao pedir um sumário das variáveis do *data frame* "paciente", as variáveis caracter "Diabetes" e "Classe", quando importadas com o pacote readxl para a IDE RStudio foram caracterizadas como caracter. No entanto, para que o sumário apresente suas categorias é necessário convertê-las para factor, um outro tipo de objeto que é apropriado para armazenar variáveis qualitativas (categóricas), tanto nominal quanto ordinal. As variáveis "Diabetes" e "Classe" são variáveis categóricas nominal e ordinal, respectivamente. Portanto, para serem corretamente classificadas, devemos executar a seguinte programação para convertê-las em objetos corretos a fim de obtermos seus resultados no sumário.

```
paciente$Diabetes<-as.factor(paciente$Diabetes)
paciente$Classe<-factor(paciente$Classe, order=TRUE)
```

Após a conversão das variáveis em objetos corretamente classificados, vamos solicitar novamente o sumário do banco de dados, conforme programação a seguir, apresentada juntamente com o resultado.

```
# solicita o sumário das variáveis
summary(paciente)
```

```
  IDPaciente    DataAdmissao                  Idade       Diabetes   Classe
 Min.   :1.00   Min.   :2014-10-15 00:00:00   Min.   :25.00   Tipo1:3   Alta :1
 1st Qu.:1.75   1st Qu.:2014-10-19 12:00:00   1st Qu.:27.25   Tipo2:1   Media:1
 Median :2.50   Median :2014-10-24 12:00:00   Median :31.00             Pobre:2
 Mean   :2.50   Mean   :2014-10-24 00:00:00   Mean   :34.75
 3rd Qu.:3.25   3rd Qu.:2014-10-29 00:00:00   3rd Qu.:38.50
 Max.   :4.00   Max.   :2014-11-01 00:00:00   Max.   :52.00
```

Perceba que após transformar as variáveis "Diabetes" e "Classe" em objetos do tipo factor, o sumário apresenta as frequências de suas categorias corretamente.

ASSIMILE

Quando um conjunto de dados possuir um número grande de variáveis ou de registros (linhas), é melhor organizá-lo, antes de manipulá-lo em ambiente R, em outro programa computacional como uma planilha eletrônica, por exemplo, pois, é mais fácil a organização. Quando estiver pronto, basta importá-lo para o ambiente R e convertê-lo em um objeto do tipo *data frame*.

Como já dito anteriormente, existem inúmeras formas de manipular *data frames* no R. Vimos que uma forma de obter um resumo (sumário) das variáveis de um banco de dados, basta utilizar a função summary( ), desde que todas as variáveis estejam corretamente classificadas.

Uma forma de manipulação interessante trata-se da agregação de bancos de dados segundo alguma variável do *data frame*. Para exemplificar uma situação deste tipo, considere o banco de dados tab2_1, disponibilizado no livro dos professores Bussab e Morettin (2015) e, também, na página do Professor Morettin (MORETTIN, 2019).

O banco de dados contém as variáveis “N”, “estado civil”, “grau_instrução”, “n._filhos”, “salário”, “Idade_anos”, “idade_meses”, “reg_procedência” referentes a 36 empregados de uma empresa. Suponha que desejamos agregar o banco de dados para a idade média segundo a região de procedência (reg_procedencia) e o grau de escolaridade (grau_instruçãoo).

Para iniciar o trabalho, iremos visualizar parte do banco de dados, importado como um *data frame* do ambiente R, segundo os comandos apresentados a seguir.

```
head(tab2_1) # faz a leitura das cinco primeiras linhas do banco
# resultados
  N estado_civil     grau_instrução n._filhos salário idade_anos idade_meses
1 1     solteiro ensino fundamental       NA    4.00        26          3
2 2      casado ensino fundamental        1    4.56        32         10
3 3      casado ensino fundamental        2    5.25        36          5
4 4     solteiro      ensino médio       NA    5.73        20         10
5 5     solteiro ensino fundamental       NA    6.26        40          7
6 6      casado ensino fundamental        0    6.66        28          0
  reg_procedência
1        interior
2         capital
3         capital
4           outra
5           outra
6        interior
```

Agora iremos agregar para obter a idade média segundo algumas variáveis do banco de dados, no caso, grau de instrução e região de procedência, conforme linhas de comando a seguir.

```
idademedia <-aggregate(tab2_1$idade_anos, by=list(grau_instrucao,reg_
procedencia), FUN=mean)
options(digits = 4) # para apresentar o resultado com 4 dígitos
idademedia # imprime o objeto com o resultado
           Group.1 Group.2     x
1 ensino fundamental capital 34.50
2 ensino médio capital 36.40
3 superior capital 38.00
4 ensino fundamental interior 31.67
5 ensino médio interior 33.86
6 superior interior 33.50
7 ensino fundamental outra 42.00
8 ensino médio outra 30.50
9 superior outra 28.50
```

Observe que a idade média, em anos, é agregada segundo as duas variáveis escolhidas previamente. Perceba que, ao fazer uma agregação, está sendo criado um novo *data frame* com as estruturas definidas, o que seria diferente se obtivéssemos a idade média segundo as variáveis

“grau_instrução” e “reg_procedência” pela função summary( ) por exemplo, a qual criaria um objeto de outra classe e seria uma espécie de resultado de tratamento estatístico.

É possível perceber, com o pouco apresentado neste texto, que há riqueza de possibilidades de tratamento de dados que a linguagem R pode proporcionar. Sugerimos que o seu estudo não se restrinja a este material, mas que ele possa ser um estímulo para você ir em busca de mais conhecimento sobre a ferramenta.

TEORIA EM PRÁTICA

Você é responsável pela equipe de analytics de uma empresa e está elaborando um treinamento com a linguagem R para todos da equipe. Você pretende deixar toda a sua equipe habilitada para realizar tratamento de dados para a gestão dos negócios de sua empresa.

O conjunto de dados utilizado para o treinamento foi retirado de Costa (2012) e se refere a uma pesquisa de perfil demográfico feito com 20 consumidores adultos do produto X. O banco de dados é apresentado na Tabela 2.

Tabela 2 – Conjunto de dados

Sexo	Idade	Escolaridade	N_filhos	Classe_social
M	35	Ens. Medio	2	B
M	25	Ens. Medio	1	B
F	40	Ens. Superior	1	C
M	25	Ens. Medio	3	B
M	32	Ens. Medio	2	C
F	22	Ens. Medio	0	C
M	37	Ens. Superior	2	B
M	28	Ens. Medio	0	B
F	25	Ens. Medio	1	B
F	39	Ens. Superior	2	C

M	35	Ens. Fundamental	1	B
F	21	Ens. Fundamental	0	A
F	27	NA	0	A
F	45	Ens. Medio	2	C
M	57	Pos-Graduacao	4	C
F	33	Ens. Medio	2	A
M	36	Ens. Fundamental	0	B
M	35	Ens. Medio	2	C
M	33	Ens. Medio	2	B
F	22	Ens. Superior	0	C

Fonte: adaptada de COSTA (2011).

Sexo: M: masculino; F: feminino.
Idade: em anos com dois dígitos.
Escolaridade: NA: sem informação.
Classe social: A: alta; B: média; C: baixa.

O conjunto de dados está disponibilizado em planilha MS Excel e você deseja exportar para o RStudio para treinar sua equipe com manipulação e tratamento de dados para fazer a análise dos dados com a linguagem R.

Para iniciar, você precisará mostrar para a sua equipe como importar dados de planilha para o RStudio. Precisará mostrar as formas de objetos reconhecidos por ele e como, principalmente, o RStudio armazena um conjunto de dados.

Nessa etapa inicial, é importante mostrar como manipular dados para realizar, por exemplo, agregações e sumarizações como etapas de exploração de dados.

Com o desafio em mãos, pergunta-se: como você fará o planejamento deste treinamento? Quais comandos pretende apresentar no primeiro encontro? Lembre-se de que, por se tratar de um ambiente que exige conhecimento de linguagem de programação, pode ser que alguns tenham uma certa dificuldade. Portanto, prepare um material bastante didático!

VERIFICAÇÃO DE LEITURA

1. Sabemos que um conjunto de dados é um arranjo retangular com linhas e colunas de informações importantes. Em sua estruturação, o que significam as linhas de um conjunto de dados?
   Assinale a alternativa CORRETA.

   a. Variáveis.

   b. Números.

   c. Caracteres.

   d. Registros.

   e. Estatísticas.

2. A linguagem R possui ampla variedade de estruturação de dados, o que é considerado uma grande vantagem. No entanto, a estrutura que mais se aproxima do que se conhece como banco de dados é apenas uma delas. Qual o seu nome?
   Assinale a alternativa CORRETA.

   a. Escalar.

   b. Vetor.

   c. Lista.

   d. Array.

   e. Data frame.

3. Para um tratamento adequado e correto é preciso classificar corretamente as variáveis de um *data frame.* Variáveis categóricas são inicialmente classificadas como caracter pela linguagem R. No entanto, precisam ser convertidas para uma outra estrutura. Qual o nome dessa estrutura? Assinale a alternativa CORRETA.

   a. Array.

   b. Factor.

   c. Caracter.

   d. Vector.

   e. Matrix.

Referências bibliográficas

AQUINO, J. A. **R para cientistas sociais.** Ilhéus, BA: EDITUS, 2014.

BUSSAB, Wilton.; MORETTIN, Pedro A. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017. 554p.

COSTA, G.G. de O. **Curso de estatística inferencial e probabilidades:** teoria e prática. São Paulo: Atlas, 2012.

KABACOFF, R.I. **R in action:** data analysis and graphics with R. 2nd. Shelter Island/ NY: Manning Publications Co., 2015.

MORETTIN, P. A. **Estatística básica**. 2019. Disponível em: https://www.ime.usp.br/~pam/EstBas.html. Acesso em: 24 ago. 2019.

Gabarito

Questão 1 – Resposta: D

Resolução: As linhas de um conjunto de dados são conhecidas genericamente como registros do conjunto de dados.

Feedback de reforço: Pense na estrutura de um conjunto de dados, imagine o que são as colunas e o que devem ser as linhas.

Questão 2 – Resposta: E

Resolução: Um conjunto de dados, como é conhecido e estruturado em programas de tratamento e armazenamento de informações, é estruturado em linguagem R como um data frame.

Feedback de reforço: Lembre-se das estruturas de objetos reconhecidas pela linguagem R.

Questão 3 – Resposta: B

Resolução: Para receberem o correto tratamento analítico, as variáveis categóricas precisam ser classificadas como factor.

Feedback de reforço: Lembre-se que as variáveis categóricas não são números. São atributos de classificação.

Modelos preditivos com R

Autor: Marcelo Tavares de Lima

Objetivos

- Apresentar os principais modelos preditivos.
- Descrever sobre alguns modelos preditivos.
- Implementar modelos preditivos em linguagem R.

1. Introdução

Prezado aluno, é sabido que ao longo de nossas vidas temos, continuamente, que tomar decisões para resolver problemas. E é claro que sempre queremos tomar a melhor decisão possível, seja no âmbito pessoal ou profissional. No entanto, tomar boas decisões é uma tarefa complexa. Como diz Ragsdale (2014, p. 1), "os problemas enfrentados pelos tomadores de decisão no ambiente comercial competitivo, de ritmo frenético e com uso intenso de dados de hoje, são geralmente de extrema complexidade e podem ser resolvidos por vários cursos de ação possíveis". A escolha da melhor alternativa representa a essência da tomada de decisão.

Um processo de modelagem de dados é construído a partir de um conjunto de relacionamentos matemáticos e de suposições, os quais representam um problema ou fenômeno a ser analisado com o intuito de se obter um resultado. Este processo é construído da maneira mais parcimoniosa possível, ou seja, é uma forma simplificada do real problema, o qual, quando implementado em computador, também é conhecido como modelo em computador.

Um termo bastante utilizado e difundido na área de modelagem é o termo em inglês *business analytics*, o qual representa um campo de estudos que faz uso de dados, computadores, estatística e matemática, com o propósito de obtenção da melhor resolução para problemas diversos (RAGSDALE, 2014). A *business analytics* utiliza métodos e ferramentas científicas para elaborar processos de tomada de decisão.

Neste texto você será apresentado aos modelos preditivos, conhecer suas especificidades e aprender a implementá-los em linguagem de programação R. Que você possa ter um excelente momento de aprendizagem!

2. Modelos preditivos de classificação no R

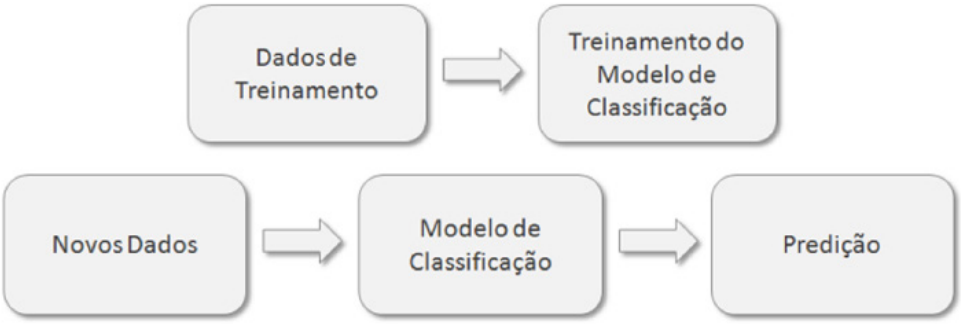
São uma classe de modelos matemáticos ou modelos simbólicos que utilizam relações matemáticas para descrever ou representar um problema para tomada de decisão através de previsão de resultados (KABACOFF, 2015).

Os modelos preditivos podem ser utilizados quando se deseja modelar dados com variáveis respostas quantitativas ou categóricas, como, por exemplo, no caso de resposta categórica, avaliar se o cliente de um banco pagará ou não um empréstimo realizado (KABACOFF, 2015). São, em geral, quando se fala de modelo preditivo com variável resposta categórica – que serão tratados neste texto, modelos com variáveis categóricas binárias, ou seja, modelos com duas categorias de resultados, cuja intenção é obter uma função de classificação para prever novos casos a serem classificados em uma ou outra categoria da variável resposta segundo um conjunto de outras variáveis chamadas preditoras ou explicativas.

Dentre os métodos de classificação existentes, os quais também são conhecidos como aprendizagem supervisionada, podemos citar a regressão logística, árvores de decisão, florestas aleatórias (*random forests*), *support vector machines*, redes neurais, etc.

A análise por modelos de classificação se inicia com a divisão do conjunto de dados em duas partes: uma parte para treinamento (aprendizado) e outra para validação - e, com a amostra para treinamento, desenvolve-se um modelo preditivo, o qual é testado em termos de sua acurácia com os dados de validação. Caetano (2015) apresentou uma ilustração que exemplifica bem o que está descrito. A Figura 1 replica a ilustração do autor.

Figura 1 – Diagrama de modelo de classificação

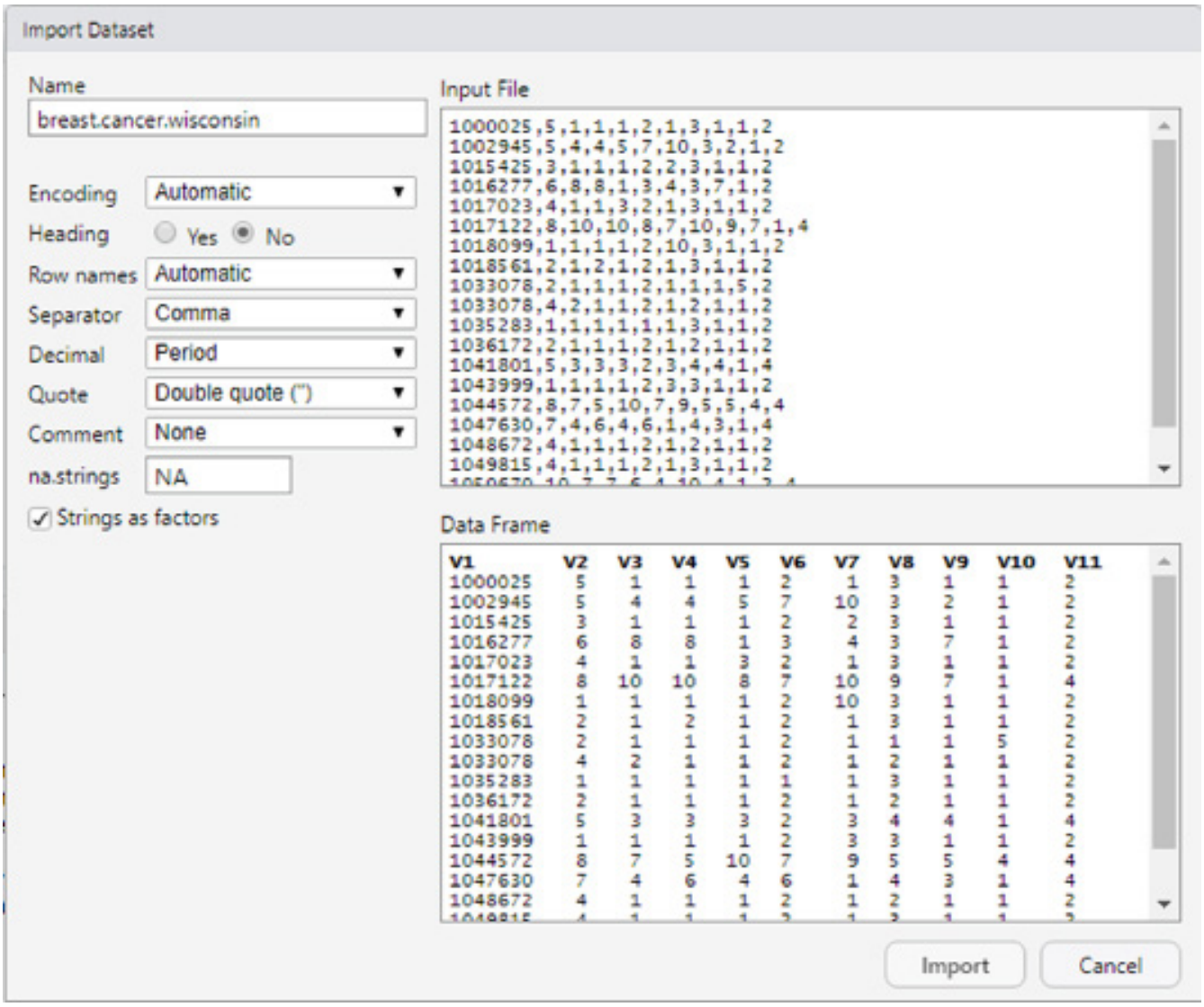


Fonte: Caetano (2015).

Para praticarmos o uso de comandos da linguagem R na construção de um modelo de classificação serão utilizados dados do *Wisconsin Breast Cancer*, disponibilizados pela *Donald Bren School of Information & Computer Sciences da University of California, Irvine (UCI)* em seu repositório de aprendizagem de máquina (UCI, 1987). O objetivo será desenvolver um modelo para prever se uma paciente tem câncer de mama segundo características da amostra de tecido abaixo da pele, retirada com agulha fina.

O conjunto de dados está disponibilizado no formato de texto delimitado com vírgulas. Para fazer download, basta acessar a página do servidor de *machine learning* da UCI (UCI, 1987). Após baixar o conjunto de dados em sua máquina, utilize o RStudio para importá-lo para o ambiente R, utilizando os menus File → Import Dataset → From Text (base). Será aberta uma caixa de busca do arquivo. Procure a pasta onde salvou a base de dados e clique nela, o RStudio mostrará uma caixa de diálogos conforme mostra a Figura 2. Para completar o processo, basta clicar no botão “Import” que a importação será finalizada e o conjunto de dados fica totalmente disponibilizado no RStudio.

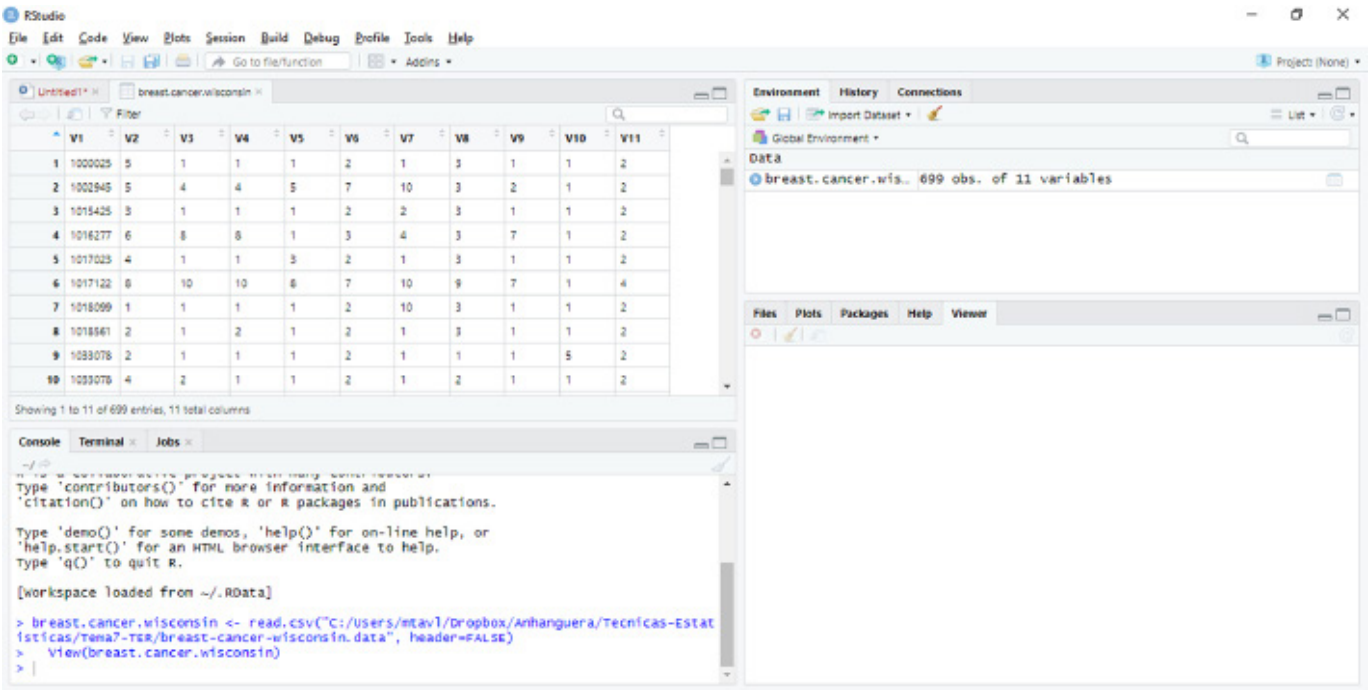
Figura 2 – Caixa de importação de dados do RStudio



Fonte: RStudio Versão 1.2.1335 para Windows (2019).

O conjunto de dados possui 699 amostras de tecidos de mama. Com a importação realizada, a tela do RStudio deverá estar apresentada conforme mostra a Figura 3.

Figura 3 – Interface do RStudio com os dados importados do Winsconsin Breast Cancer



Fonte: RStudio Versão 1.2.1335 para Windows (2019).

É possível observar na Figura 3 que as variáveis importadas apresentam nomes genéricos com V1 até V11. No entanto, na página onde os dados são disponibilizados, também, há um documento com a descrição das variáveis (coluna) do conjunto de dados. O documento tem nome "*breast-cancer-wisconsin*", é um arquivo em formato texto e é do tipo NAMES.
O Quadro 1 apresenta os nomes das variáveis do banco de dados.

Quadro 1 – Lista de variáveis do conjunto de dados do *Winsconsin Breast Cancer* (características citológicas)

V1: ID – variável de identificação da amostra.
V2: Espessura da massa (valores de 1 a 10).
V3: Uniformidade do tamanho da célula (valores de 1 a 10).
V4: Formato da célula (valores de 1 a 10).
V5: Adesão marginal (valores de 1 a 10).
V6: Tamanho da célula epitelial única (valores de 1 a 10).
V7: Apenas núcleos (valores de 1 a 10).
V8: Cromatina branda (valores de 1 a 10).
V9: Nucléolos normais (valores de 1 a 10).
V10: Mitose (valores de 1 a 10).
V11: Classe (2 = benigno; 4 = maligno).
Quanto mais próximo de 1: mais próximo de benigno.
Quanto mais próximo de 10: mais próximo de maligno.

Fonte: elaborado pelo autor.

Para não deixar o nome do banco de dados muito extenso, é possível simplificar o seu nome com a seguinte linha de programação “dados←breast.cancer.wisconsin”. Assim, o banco passa a ser chamado de “dados”. A variável que classifica a amostra como benigna ou maligna precisa ser convertida para fator, pois é uma variável categórica. O comando utilizado deverá ser elaborado conforme apresentado a seguir, juntamente com um resumo da classificação.

```
dados$Classe<-factor(dados$V11, levels=c(2,4), labels=c(“Benigno”,”Maligno”))
summary(dados$Classe)
# Resultado
Benigno Maligno
   458     241
```

É possível observar que das 699 amostras, 458 são classificadas como benignas e 241 como malignas, representando em percentual, respectivamente, 65,5 e 34,5% do total da amostra.

O desafio é descobrir um conjunto de regras de classificação que possam ser utilizadas para predizer de forma acurada a malignidade de algumas combinações das características citológicas.

Para dar início à construção da modelagem, o conjunto de dados é dividido em amostra de treinamento (70%) e amostra de validação (30%), conforme a programação a seguir.

```
set.seed(1234) # cria uma semente
treino<-sample(nrow(dados), 0.7*nrow(dados))
dados.treino<-dados[treino,]
dados.valid<-dados[-treino,]
# Resultado
table(dados.treino$Classe)
Benigno Maligno
   319     170
table(dados.valid$Classe)
Benigno Maligno
   139      71
```

A amostra de treinamento tem 489 casos (319 benignos, 170 malignos) e a amostra de validação tem 210 casos (139 benignos, 71 malignos). Como dito anteriormente, a amostra de treinamento será utilizada para criar esquemas de classificação com as técnicas listadas anteriormente, como regressão logística, árvore de decisão, árvore de decisão condicional, floresta aleatória, etc. A amostra de validação será usada para avaliar a eficiência desses esquemas.

2.1 Regressão logística

É um tipo de modelo linear generalizado útil para predizer um resultado binário (variável resposta dicotômica) a partir de um conjunto de dados (KABACOFF, 2015). A função glm( ) implementada no pacote básico do R é apropriada para ajustar um modelo de regressão logística binária. Variáveis preditoras categóricas (fatores) são automaticamente implementadas através de um conjunto de variáveis *dummy*.

PARA SABER MAIS

Uma variável binária (também denominada variável *dummy*) é aquela que só tem dois valores distintos, geralmente zero e um. Em um modelo de classificação, a variável dependente também pode ser influenciada por variáveis de natureza qualitativa, onde, em geral, significam a presença ou ausência de uma "qualidade" ou atributo, como ser homem ou mulher, ser católico ou não, etc.

Todas as variáveis independentes ou preditoras do conjunto de dados *Wisconsin Breast Cancer* são numéricas, de tal forma que a codificação para variável *dummy* torna-se desnecessária. A programação em linguagem R para o conjunto de treinamento é apresentada a seguir.

```
dados.treino<-dados.treino[-1] # exclui V1 (ID)
dados.treino<-dados.treino[-10] # exclui V11 (classe original)
# transforma V7 em variável numérica
dados.treino$V7 <-as.numeric(dados.treino$V7)
# Recodificação para identificar valor ausente
dados.treino$V7[dados.treino$V7==11] <- NA
# Ajuste do modelo
modelo <-glm(Classe~.,data=dados.treino,family=binomial())
# visualização dos resultados do modelo
summary(modelo)
```

Um modelo de regressão logística é ajustado aos dados utilizando como variável dependente a "Classe" e as demais variáveis como variáveis independentes do modelo. O modelo está baseado nos casos inclusos no data frame "dados.treino". O resultado da programação é apresentado a seguir:

```
Call:
glm(formula = Classe ~ ., family = binomial(), data = dados.treino)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1639  -0.1030  -0.0486   0.0125   2.1703

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.21807    1.68576  -7.248 4.24e-13 ***
V2            0.63712    0.18155   3.509 0.000449 ***
V3           -0.04355    0.25163  -0.173 0.862582
V4            0.54020    0.26961   2.004 0.045108 *
V5            0.30750    0.13360   2.302 0.021359 *
V6            0.05277    0.21309   0.248 0.804398
V7            0.50693    0.16980   2.986 0.002831 **
V8            0.75279    0.22032   3.417 0.000634 ***
V9            0.09499    0.14487   0.656 0.512013
V10           0.79333    0.31869   2.489 0.012798 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As estimativas dos parâmetros do modelo são apresentadas na coluna “Estimate” e os valores de probabilidades que avaliam a significância estatística são apresentados na última coluna (Pr>|z|). Depois de ajustado o modelo, é possível classificar novos casos utilizando o conjunto de dados de validação, conforme programação a seguir.

```
dados.valid<-dados.valid[-1] # exclui V1 (ID)
dados.valid<-dados.valid[-10] # exclui V11
# Transforma V7 em numérica
dados.valid$V7 <-as.numeric(dados.valid$V7)
# Recodifica os valores ausentes (missing)
dados.valid$V7[dados.valid$V7==11] <- NA
# Utiliza o modelo pra fazer predições
prob <- predict(modelo, dados.valid, type=”response”)
logito.pred <- factor(prob>0.5, levels=c(FALSE, TRUE), labels=c(“Benigno”,”Maligno”))
logito.perf <- table(dados.valid$Classe, logito.pred, dnn=c(“Atual”, “Predito”))
logito.perf
# resultado (matriz de confusão)
        Predito
Atual     Benigno Maligno)
  Benigno     132        7
  Maligno       4       62
```

Depois do modelo ajustado com dados.treino , utiliza-se dados.valid para classificar casos no conjunto de dados. Por padrão, a função predict( ) faz a previsão do logaritmo da *odds ratio* (razão de chances) de se obter um resultado de câncer maligno, no entanto, ao se utilizar o comando “type=response”, o resultado obtido é a probabilidade.

Os casos com probabilidade acima de 0,5 (50%) são classificados no grupo de “Maligno” e, o caso contrário, no grupo de “Benigno”. Para finalizar, é obtida uma tabela para comparar a classificação inicial e a classificação obtida com o modelo ajustado (predita) – chamada matriz de confusão.

O resultado apresentado pela matriz de confusão mostra que 132 casos classificados previamente como benignos tiveram a mesma classificação dada pelo modelo ajustado, e 62 casos que foram classificados previamente como malignos, também o foram pelo modelo ajustado. Cinco casos no *data frame* “dados.valid” tiveram previsão “missing” e não foram incluídos na avaliação a acurácia. O total de casos corretamente classificados (acurácia) é de ou, aproximadamente, igual a 95% na amostra de validação.

Para finalizar, observe que as variáveis V3, V6 e V9, correspondendo, respectivamente, segundo o Quadro 1, a “uniformidade do tamanho”, “tamanho da célula” e “nucléolos normais”, apresentaram valores de probabilidade muito alto, representando que não têm significância estatística para o ajuste do modelo. O que fazer, se for possível fazer algo, com variáveis preditoras que resultaram em coeficientes não significativos?

Em um contexto de previsão, torna-se útil remover as variáveis que não apresentaram coeficientes significativos do modelo final. Para isso, pode-se utilizar a regressão logística *stepwise* para gerar um modelo menor, com menos variáveis. Na prática, variáveis preditoras são adicionadas ou removidas até se obter o menor valor da estatística Critério de Informação de Akaike (AIC).

ASSIMILE

O critério de informação de Akaike (AIC) possibilita realizar comparação entre modelos de regressão logística ajustados. Esta estatística leva em conta o ajuste do modelo e o número de parâmetros necessários para o ajuste. Modelos com menores valores de AIC sinalizam ajuste adequado, preferencialmente com poucos parâmetros.

Para obter um modelo mais parcimonioso, podemos utilizar a seguinte programação em linguagem R.

```
modelo.reduzido <-step(modelo)
summary(modelo.reduzido)
# Resultado do modelo reduzido
Call:
glm(formula = Classe ~ V2 + V4 + V5 + V7 + V8 + V10, family = binomial(), data =
dados.treino)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.08207  -0.10511  -0.04929   0.01393   2.13590

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.1350     1.5972  -7.598 3.01e-14 ***
V2            0.6532     0.1802   3.625 0.000289 ***
V4            0.5971     0.2136   2.795 0.005187 **
V5            0.3076     0.1249   2.463 0.013788 *
V7            0.5025     0.1663   3.021 0.002519 **
V8            0.7584     0.2115   3.586 0.000335 ***
V10           0.7947     0.3108   2.557 0.010554 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 623.253  on 484  degrees of freedom
Residual deviance:  75.369  on 478  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 89.369

Number of Fisher Scoring iterations: 8
```

É possível observar que as variáveis citadas anteriormente como não significativas foram excluídas automaticamente pelo comando executado, obtendo-se, assim, um modelo com menos variáveis.

2.2 Árvore de decisão

É um procedimento de classificação bastante popular em mineração de dados. Consiste em criar um conjunto de regras que dividem as variáveis preditoras em duas (binárias) para criar uma árvore que possa ser utilizada para classificar novas observações em um dos dois grupos. Existem dois tipos de árvore de decisão, a clássica e a condicionada. Neste texto iremos apresentar apenas o método de classificação de árvore de decisão clássica (KABACOFF, 2015).

O processo se inicia considerando uma variável dependente binária ou dicotômica e um conjunto de variáveis preditoras. O algoritmo para se construir uma árvore de decisão clássica, segundo Kabacoff (2015), é o seguinte:

1. Escolhe-se uma variável independente (preditora) que melhor particione os dados em dois grupos, tal que a homogeneidade da resposta em dois grupos possa ser maximizada. Se a variável independente for quantitativa contínua, escolhe-se um ponto de corte que maximize a homogeneidade para os dois grupos que serão criados.

2. Separa-se os dados nos dois grupos criados e continua-se o processo para cada subgrupo.

3. Repete-se os passos 1 e 2 até que um subgrupo contenha menos que o número mínimo de observações ou nenhuma divisão diminua a heterogeneidade além de um valor específico. Os subgrupos finais são chamados “nós terminais”. Cada nó terminal é classificado como uma categoria da resposta ou baseada no valor mais frequente para a amostra desse nó.

4. Para classificar um caso, executa-se a árvore até o nó terminal e designa-o para o valor assinalada da resposta como no passo 3.

Kabacoff (2015, p. 394, tradução nossa) afirma que, “infelizmente, este processo tende a produzir árvores muito extensas e com problemas de sobreajuste. Como consequência, novos casos não são bem classificados”. A compensação para este problema, segundo o mesmo autor, é realizar uma poda na árvore, selecionando erros de predição dez vezes menores.

Em linguagem R, uma árvore de decisão pode ser construída com a função rpart( ) e prune( ), ambas implementadas no pacote rpart. Para exemplificar a criação de uma árvore de decisão com linguagem R, utilizaremos o banco de dados Winsconsin Breast Cancer, separados em dados para treinamento e dados para validação, conforme programação a seguir.

```
library(rpart) # carrega o pacote (biblioteca)
set.seed(1234) # seleciona uma semente aleatória
# cria a árvore de decisão segundo alguns parâmetros
arvore <- rpart(Classe~., #indica a variável dependente
        data=dados.treino, # indica os dados
        method="class", # indica o método
        parms=list(split="information")) # indica como
                    será feita a construção dos ramos
arvore$cptable # avalia o sobreajuste para podar (se
            # necessário a árvore definitiva.
# resultado do cptable
        CP nsplit  rel error    xerror       xstd
1 0.81764706      0 1.00000000 1.0000000 0.06194645
2 0.03823529      1 0.18235294 0.2176471 0.03440066
3 0.01764706      3 0.10588235 0.2000000 0.03308581
4 0.01000000      4 0.08823529 0.1823529 0.03169642
```

A componente cptable da função rpart( ) contém informações sobre o erro de predição para vários tamanhos de árvores de decisão.
O parâmetro de complexidade (cp) é usado para penalizar árvores grandes. O tamanho da árvore é definido pelo número de divisões dos ramos (nsplit). Uma árvore com n divisões tem $n + 1$ nós terminais.

A coluna rel error do cptable contém a taxa do erro para uma árvore de um dado tamanho (nsplit). O erro cross-validated (xerror) é calculado com base na validação cruzado 10-fold. A coluna xstd contém o erro padrão do erro cross-validated.

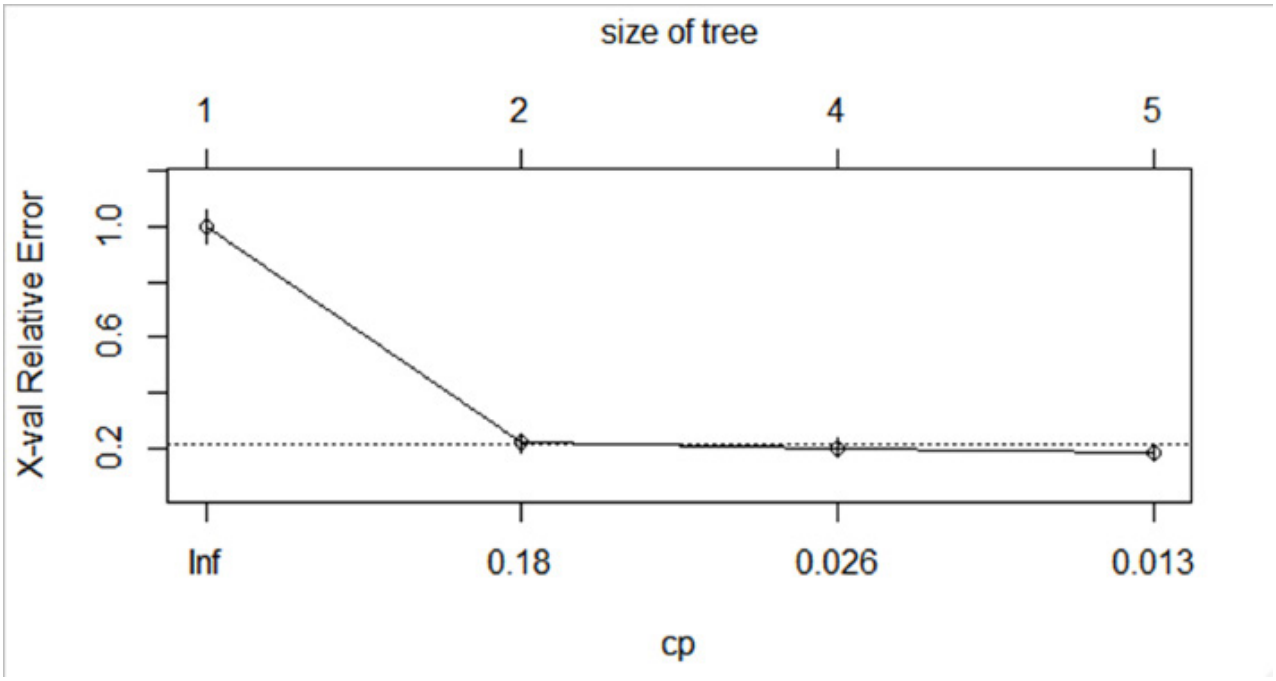
A escolha da árvore definitiva pode ser feita com base no menor erro cross-validated que esteja dentro do intervalo de 1 erro padrão, ou seja $x_{error} \pm x_{std}$. O menor erro cross-validated aparece na última linha do resultado e é igual a 0,1823529 e tem erro padrão igual a 0,03169642. Neste caso, a menor árvore possível com erro cross-validated dentro do intervalo 0,1823529 ± 0,03169642 (ou seja, 0,1506 e 0,2140) é selecionada. Olhando para o resultado do cptable, é possível identificar que uma árvore com três divisões (erro cross-validated = 0,20) é a que melhor se ajusta a essas condições.

De maneira equivalente ao procedimento de escolha apresentado no parágrafo anterior, é possível selecionar o tamanho de árvore definitivo associado ao parâmetro de complexidade (cp) através de uma visualização gráfica conforme programação a seguir.

```
plot(arvore) # plota um gráfico de linhas do erro cross-
        # validated versus o cp.
```

O resultado da programação é apresentado na Figura 4.

Figura 4 – Erro cross-validated *versus* cp



Fonte: elaborada pelo autor.

A Figura 4 indica que o número de divisões da árvore definitiva está entre 2 e 4, pois é onde parece tocar a linha tracejada. Duas divisões não são a melhor escolha porque estão acima da linha tracejada, e quatro divisões estão abaixo. Portanto, conclui-se que três divisões são o número ideal, o que resultará em uma árvore com quatro nós terminais.

A função prune( ) utiliza o parâmetro de complexidade (cp) para cortar a árvore no tamanho desejado. Ela utiliza a árvore total e corta as divisões menos importantes com base no cp. Os resultados apresentados de cp indicam que uma árvore com três divisões tem cp igual a 0,01764706. A programação a seguir com o uso do valor do cp citado realiza a poda desejada na árvore total.

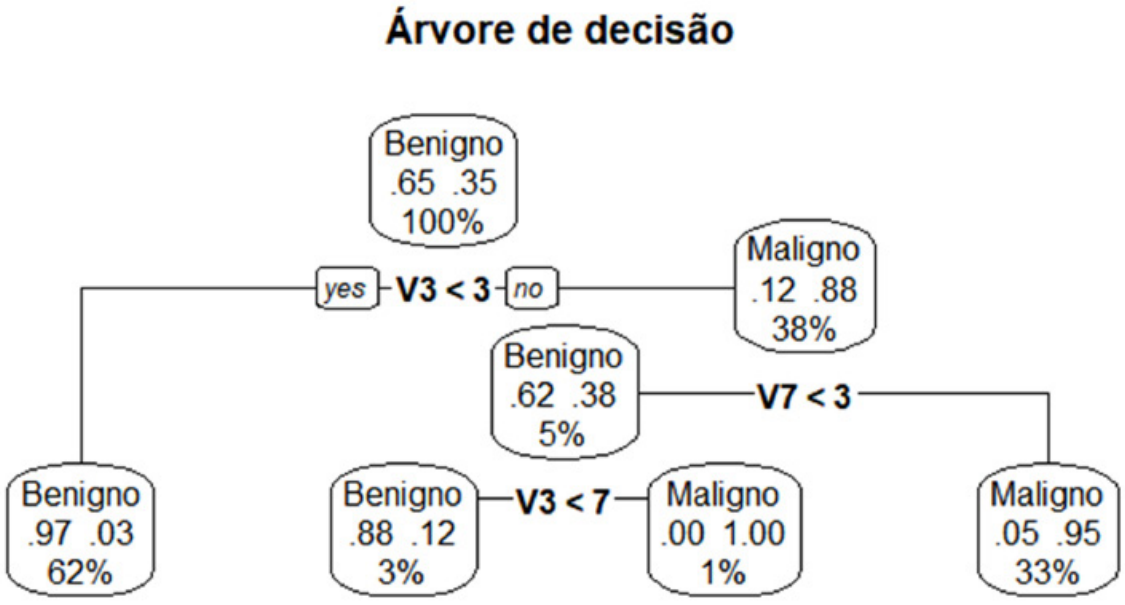
```
arvore.poda<-prune(arvore,cp=0.01764706)
```

A função prp( ) pertencente ao pacote rpart.plot é utilizada para criar um gráfico de árvore de decisão. Ela funciona com algumas opções, como, por exemplo, a opção type=2 apresenta no gráfico os rótulos abaixo de cada nó da árvore. O parâmetro extra=104 inclui os valores

de probabilidades para cada classe e, também, a porcentagem de observações em cada nó. A opção fallen.leaves=TRUE mostra os nós terminais na parte de baixo do gráfico. A programação em linguagem R para a produção do gráfico da árvore decisão é apresentada a seguir. O gráfico resultante apresentado na Figura 5 mostra os pontos de corte nas variáveis que classificam a amostra utilizada como benigna ou maligna até um determinado nível. É possível observar que existem valores dentro de cada nó apresentado, onde, por exemplo, no primeiro nó da árvore, classificado como “Benigno”, os valores “.65” e “.35” representam as probabilidades de classificação segundo a variável V3 com ponto de corte “<3”. Já o valor 100% indica que neste nó, para a realização da classificação, foi utilizada 100% da amostra.

```
library(rpart.plot) # carrega a biblioteca
prp(arvore.poda, # chama a árvore podada
   type=2, # rótulos embaixo de cada nó
   extra=104, # inclui probabilidades e percentual
   fallen.leaves = TRUE, # nós finais na parte de baixo
   main="Árvore de decisão") # título do gráfico
# gráfico resultante
```

Figura 5 – Árvore de decisão para dados da *Wisconsin Breast Cancer*



Fonte: elaborada pelo autor.

Para classificar uma observação, deve-se iniciar no topo da árvore e mover-se para a esquerda se uma condição for verdadeira ou para a direita no caso contrário. Continua se movendo para baixo da árvore até alcançar o nó terminal. Classifica-se a observação através do rótulo do nó referente.

Para finalizar, utilizamos a função predict( ) para classificar cada observação do banco de validação, conforme programação apresentada a seguir.

```
arvore.pred <- predict(arvore.poda, #chama a árvore podada
                  dados.valid, # dados validação
                  type="class") # classificação
# tabela com a classificação final
classifica <- table(dados.valid$Classe, # classificação
                                  # prévia
                arvore.pred, # classificação atual
                dnn=c("Atual", "Predito")) # rótulos
classifica # imprime a tabela de comparação
# resultado
        Predito
Atual     Benigno Maligno
  Benigno     130       9
  Maligno       8      63
```

A tabela comparativa entre a classificação inicial e a classificação final, após a modelagem por árvore de decisão, apresenta acurácia igual a (130 + 63)/210 = 0,9190, ou seja, 91,90% na amostra de validação. Diferente da modelagem realizada pela regressão logística, todos os 210 casos modelados com árvore de decisão tiveram classificação na árvore definitiva.

3. Conclusão

Existe uma infinidade de modelos preditivos. Este texto apresentou apenas dois tipos. No entanto, para exemplificar, podemos citar os seguintes modelos de classificação: árvore de decisão condicional, florestas aleatórias, *support vector machines*, etc.

O que pretendemos apresentar aqui foi apenas uma degustação deste vasto campo de procedimentos quantitativos. Desejamos que você possa buscar outros tipos existentes e escolher aquele que mais lhe ajudar a solucionar seus problemas, sejam corporativos ou acadêmicos. Que você possa se empolgar com o mundo da análise de dados. Desejamos todo o sucesso a você!

TEORIA EM PRÁTICA

Você é responsável pela equipe de analytics de uma concessionária e está avaliando o perfil de seus clientes. Para isso, busca uma base de dados que sua empresa possui com algumas informações de seus clientes. Você pretende utilizar um modelo preditivo para classificar os clientes como adimplentes ou inadimplentes. Será utilizada linguagem R para a elaboração de um modelo de classificação (preditivo) para realizar esse trabalho.

O conjunto de dados utilizado para a realização do trabalho foi retirado de USP (2019) e se tem como variáveis o status do cliente (adimplente, inadimplente), renda mensal (em salários mínimos), número de dependentes e vínculo empregatício (0 = sem vínculo, 1 = com vínculo). O banco de dados é apresentado na Tabela 1.

Tabela 1 – Dados sobre clientes da concessionária

ST	R	ND	VE
0	2,5	3	1
1	1,7	3	1
0	4	2	1
1	2,3	2	1
1	3,7	4	0
0	4,8	1	0
1	1,9	3	0
0	5,3	2	1
1	3,1	4	1
1	1,9	3	1
1	2,3	4	1
0	3,6	1	0
0	4,7	2	1
0	5,8	2	0
0	6	4	0
0	3,9	3	1
1	2,4	4	1
1	1,7	4	1
0	3,7	2	0
0	4,8	1	0
0	3,2	2	1
1	2,7	3	1
1	1,2	3	1

ST	R	ND	VE
0	8,2	5	0
1	1,8	1	1
1	2,5	1	1
1	2,2	3	1
0	4	1	0
0	4,2	1	0
0	3,7	1	0
1	2,4	2	1
1	1,6	3	1
1	2	1	1
1	2,5	3	1
0	3,8	1	0
0	4,3	2	0
1	2	2	1
0	5,2	2	0
1	2,4	3	0
0	2,6	4	0
0	1,3	2	1
0	3,8	1	1
0	4,5	0	1
0	3	0	1
1	2,1	2	1
1	1,9	2	1

ST	R	ND	VE
0	1,7	4	0
1	1,7	2	1
1	1,3	3	1
0	2,5	1	1
0	3,5	2	0
0	5,6	3	0
0	3,8	2	0
0	4	0	0
1	2,5	1	1
1	1,2	2	0
0	3	1	0
0	3	1	0
1	2,1	2	1
0	2,5	1	0
0	2,9	1	0
0	4	3	0
0	3,2	3	0
1	1,2	2	1
0	3,5	3	0
0	4	1	0
1	2,3	3	1
0	2,9	4	0
1	2,4	2	1

ST	R	ND	VE
0	5	3	0
1	2,2	3	0
1	1,3	3	1
1	1,7	3	1
0	3	2	0
0	3	2	1
0	3,5	2	1
0	5,8	2	1
0	4,8	1	0
1	2,3	3	1
1	2,6	2	1
1	1,8	2	1
1	2,9	2	1
0	3,2	1	0
0	4,2	1	0
0	2,6	1	0
0	6	1	0
1	4,5	3	1
1	1,3	2	1
1	2,4	2	1
0	4,3	2	0
1	1,8	0	1
0	2,4	2	0

Fonte: USP (2019)

ST: status.
R: renda em salários mínimos.
ND: número de dependentes.
VE: vínculo empregatício.

O conjunto de dados pode ser transportado para uma planilha MS Excel e, depois, para o RStudio, para treinar sua equipe com manipulação e tratamento de dados e fazer a análise dos dados com a linguagem R.

Com o desafio em mãos, pergunta-se: como você fará para elaborar o modelo preditivo? Quais tipo de modelagem pretende utilizar para classificar os clientes da concessionária? Pense sobre como vai abordar esse desafio!

VERIFICAÇÃO DE LEITURA

1. Suponha que um modelo preditivo deseja classificar as pessoas frequentadoras de um parque de diversões segundo as categorias de frequentador assíduo ou não assíduo. Em qual tipo de variável que a resposta do modelo pode ser classificada?
   Assinale a alternativa CORRETA.

   a. Contínua.

   b. Quantitativa.

   c. Categórica.

   d. Parâmetro.

   e. Estatística.

2. Para se fazer uma análise de dados com modelos de classificação, divide-se o conjunto de dados em duas partes, onde uma delas avalia a acurácia do modelo ajustado. Como se chama a parte do conjunto de dados utilizadas para isso?
   Assinale a alternativa CORRETA.

   a. Treinamento.

   b. Validação.

   c. Classificação.

   d. Acurácia.

   e. Ajustamento.

3. Com a divisão do conjunto de dados para elaborar modelos de predição, a primeira parte utilizada para aprendizado é constituída por qual percentual da amostra?
Assinale a alternativa CORRETA.

a. 30%.

b. 50%.

c. 80%.

d. 70%.

e. 60%.

Referências bibliográficas

CAETANO, M. **Modelos de classificação:** aplicações no setor bancário. 2015. 94f. Dissertação (Mestrado em matemática aplicada) – Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Campinas/SP, 2015. Disponível em: http://repositorio.unicamp.br/jspui/bitstream/REPOSIP/306286/1/Caetano_Mateus_M.pdf. Acesso em: 2 out. 2019.

KABACOFF, R.I. **R in action:** data analysis and graphics with R. 2nd. Shelter Island/NY: Manning Publications Co., 2015.

RAGSDALE, C.T. **Modelagem de planilha e análise de decisão:** uma introdução prática a business analytics. São Paulo: Cengage Learning, 2014. 594p.

RStudio Team. **RStudio:** Integrated Development for R, Version 1.2.1335 © 2009-2019, RStudio, Inc. RStudio, Inc., Boston, MA URL. 2019. Disponível em: http://www.rstudio.com/. Acesso em: 9 out. 2019.

UCI. **Machine learning repository**. Center for Machine Learning and Intelligent Systems. Irvine, CA. Disponível em: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/. Acesso em: 9 out 2019.

USP**. Regressão logística.** Notas de aula. Universidade de São Paulo. 2019. Disponível em: https://edisciplinas.usp.br/pluginfile.php/3769787/mod_resource/content/1/09_RegressaoLogistica.pdf. Acesso em: 9 out. 2019.

Gabarito

Questão 1 – Resposta: C

Resolução: A variável resposta é uma classificação de frequentador assíduo ou não, então pode ser considerada como categórica.

Feedback de reforço: Perceba que a variável resposta não é numérica.

Questão 2 – Resposta: B

Resolução: Para elaborar um modelo de classificação, o conjunto de dados é dividido em duas partes, onde a parte chamada de dados de validação é utilizada para avaliar a acurácia do modelo ajustado.

Feedback de reforço: Lembre-se que a segunda parte da divisão dos dados é quem avalia a predição.

Questão 3 – Resposta: D

Resolução: O conjunto é dividido em duas partes, onde os dados para treinamento, ou seja, para a elaboração do modelo representa 70% dos dados totais.

Feedback de reforço: Lembre-se que a maior parte dos dados ficam alocados na primeira parte, a de aprendizagem.

Bons estudos!