ISSN: 2317-0840

LECA: Pacote educacional com interface gráfica de usuário para estatística descritiva e probabilidade no R

Ana Carolina A. Barbosa[1†], Deyse M. P. Gebert[2], Airton Kist[3]

[1]*DEINFO/UEPG.*
[2]*DEMAT/UEPG. E-mail: dmpgebert@uepg.br.*
[3]*DEMAT/UEPG. E-mail: kist@uepg.br.*

Resumo: *A acessibilidade ao desenvolvimento proporcionada pelo software gratuito e de código aberto R permite que pacotes sejam constantemente disponibilizados e atualizados em seu repositório. Inúmeros desses pacotes se destacam na área da estatística, porém exigindo um conhecimento prévio do usuário quanto a linguagem R para a utilização de suas funções e interpretação de seus resultados, não abordando o tópico de forma educacional. Pacotes que oferecem uma interface gráfica ao usuário costumam ser mais acessíveis nesse cenário, uma vez que exigem pouco conhecimento adicional para sua utilização. O pacote em R LECA oferece funções de distribuição de frequências para variáveis aleatórias – gerando a tabela de frequências e a opção de cálculos de medidas descritivas e geração de gráficos sobre ela –, construção de diagramas de árvore a partir de cenários fornecidos pelo usuário e o acesso ao software estatístico PQRS©. Essas funções são acessadas dentro de uma interface gráfica de usuário, em português brasileiro, implementada em linguagem de programação Tcl. O resultado final facilita a manipulação e leitura de dados por seus usuários, mesmo os inexperientes com o software R. Essa abordagem visa estimular a utilização do R e do LECA para aprendizagem de estatística e probabilidade no Brasil e a contribuição em futuras atualizações do pacote, que possibilitarão sua utilização como material de apoio em um maior número de áreas.*

Palavras-chave: Software R; Estatística descritiva; Probabilidade; Interface gráfica de usuário.

Abstract: *The development accessibility provided by the free and open source software R allows packages to be constantly made available and updated on it's repository. A great number of these packages are known for being statistical focused, but they require a previous knowledge of the R language for their functions to be used and their results to be read, which isn't an educational approach for the matter. Packages that provide a graphical user interface tend to be more accessible in this scenario, since they require little additional knowledge for their use. The R package LECA offers frequency distribution functions for aleatory variables – creating the frequency table and the options of computing descriptive measures and generating graphs on it –, the construction of tree diagrams from scenarios chosen by the user and allows access to the statistical software PQRS©. These functions are accessed through a brazilian portuguese graphical user interface developed in Tcl programming language. The final result makes the manipulation and reading of data easier for it's user, even the ones that have little to no experience with R. This approach has the goal to encourage the use of R and LECA in the learning of statistics and probability in Brazil as well as the contribution on future updates of the package, which can make it's use as an aid tool on a larger number of areas possible.*

Keywords: R Software; Descriptive statistics; Probability; Graphical user interface.

Introdução

O R (R CORE TEAM, 2017) é um software gratuito e de código aberto que disponibiliza diversas funções estatísticas para seus usuários. O fácil acesso ao desenvolvimento proporcionado pela ferramenta estimula a criação e atualização constante de pacotes – que são grupos de uma ou mais funções executáveis dentro do ambiente – em seu repositório e permite que novas soluções e produções possam ser propostas por meio dela.

[†]Autora correspondente: *anacbse@gmail.com*.

Por ser um software com foco estatístico, grande número dos pacotes disponibilizados no R possuem foco na área da matemática. No Brasil, produções como os pacotes *ds* (ARNHOLD, 2014) e *ExpDes* (FERREIRA et al., 2014), oferecem, respectivamente, opções de estatística descritiva e análise de variância. Já o pacote *fdth* (FARIA, JELIHOVSCHI e ALLAMAN, 2016) oferece funções de distribuição de frequências sobre um conjunto de dados. Para a utilização desses pacotes é necessário um conhecimento prévio sobre a linguagem de programação R. Suas funções precisam ser invocadas manualmente por meio do terminal do software, com muitas delas exigindo um conhecimento mais avançado sobre a utilização de parâmetros dentro da ferramenta e de conceitos estatísticos. Por exemplo, para a geração personalizada de uma tabela de frequências pelo pacote *fdth*: uma variável contínua deve apresentar, no mínimo, quatro argumentos além dos dados observados para geração da tabela, que são os parâmetros para número de classes, definição dos limites inferior e superior e amplitude de cada intervalo. O mesmo ocorre para a leitura dos resultados: esses pacotes os apresentam por meio do console da ferramenta, o que deixa os valores calculados inacessíveis dentro e fora do R. Em um ambiente educativo essas características podem ser custosas aos que possuem pouca ou nenhuma experiência com a ferramenta, principalmente para estudantes que não são da área da estatística.

Pacotes que incluem uma interface gráfica exigem pouco conhecimento adicional de seus usuários e costumam ser mais acessíveis e de fácil uso. O pacote Rcmd (FOX, 2005) é um exemplo de pacote estatístico em R que oferece interface gráfica para comunicação com seus usuários (FOX, 2017; FOX E BOUCHET-VALAT, 2018). O pacote se torna atrativo por fazer uso de diferentes tipos de objetos gráficos como menus, listas e botões, facilitando o contato no ambiente. Mais exemplos de extensões gráficas executáveis no ambiente R são abordados por Valero-Mora e Ledesma (2012). Adicionalmente, softwares interativos se mostram ferramentas benéficas para o ensino da estatística, como abordado por Knypstra (1999). Estudantes mostram dificuldade na compreensão de termos estatísticos como "população" e "amostra", por exemplo, que se diferem bastante de seus significados cotidianos (KNYPSTRA, 1999, apud HAWKINS et al., 1992) e o autor propõe como solução a utilização visual dos conceitos.

O objetivo desse trabalho é o desenvolvimento de um pacote em R, o LECA, que apresente suas funções em uma interface gráfica de usuário interativa capaz de autoajuste, em português brasileiro, com foco na aprendizagem. Dessa maneira, será possível propiciar o acesso a estatística básica a todo tipo de usuário, bem como facilitar a realização de cálculos por meio das funcionalidades oferecidas pelo software R sem que o processo de aprendizagem de estudantes não familiares com essa ferramenta seja prejudicado. As funções matemáticas oferecidas por ele são de estatística descritiva e de probabilidade com foco em interação e apresentação intuitiva de resultados.

A seguir serão apresentados a metodologia de implementação do pacote no software R; o pacote LECA, com exemplos de sua funcionalidade; e as considerações finais sobre sua criação e atualizações futuras.

Metodologia de implementação do LECA

Todo o embasamento teórico utilizado na implementação do LECA foi baseado nos seguintes livros: Bussab e Morettin (2013), Larson e Farber (2010) e Martins e Domingues (2014). Essas obras também apresentam modelos que influenciam a interface gráfica do pacote, como o formato da distribuição de frequências, seus gráficos e o diagrama de árvore de probabilidades.

A linguagem de programação utilizada para a implementação da interface gráfica é a Tcl (OUTERSHOUT, 1989). O pacote tcltk (DALGARRD, 2001a; DALGARRD, 2001b) é capaz de reconhecer e executar a linguagem dentro do software R junto a sua extensão gráfica Tk, assim como

o pacote tcltk2 (GROSJEAN, 2014), e ambos são utilizados no desenvolvimento e utilização do LECA. Como a interface gráfica possui papel importante no pacote apresentado, foram realizadas pesquisas quanto a sua implementação no R com foco em usabilidade, assunto bastante discutido por Lawrence e Verzani (2014, 2016). O restante das opções oferecidas pelo LECA foi implementado com o auxílio dos pacotes básicos do R.

A entrada de dados se dá por meio de arquivos em formato CSV contendo uma ou mais variáveis com identificação no cabeçalho ou por inserção manual de forma individual e acumulativa. A comunicação com o usuário na apresentação de resultados se dá pela janela do LECA por meio de tabelas gráficas, arquivos editáveis ou pela janela de geração de gráficos.

O pacote foi desenvolvido e testado utilizando o R e sua versão gráfica, RStudio (RSTUDIO TEAM, 2016), ambos na versão 3.4.1 e em sistema operacional Windows nas versões 8.1 e 10.

O pacote LECA

A janela principal do pacote LECA contém as opções de entrada de dados no cabeçalho e abas que apresentam, separadamente, as funções para estatística descritiva (Figura 1) e probabilidade (Figura 2).

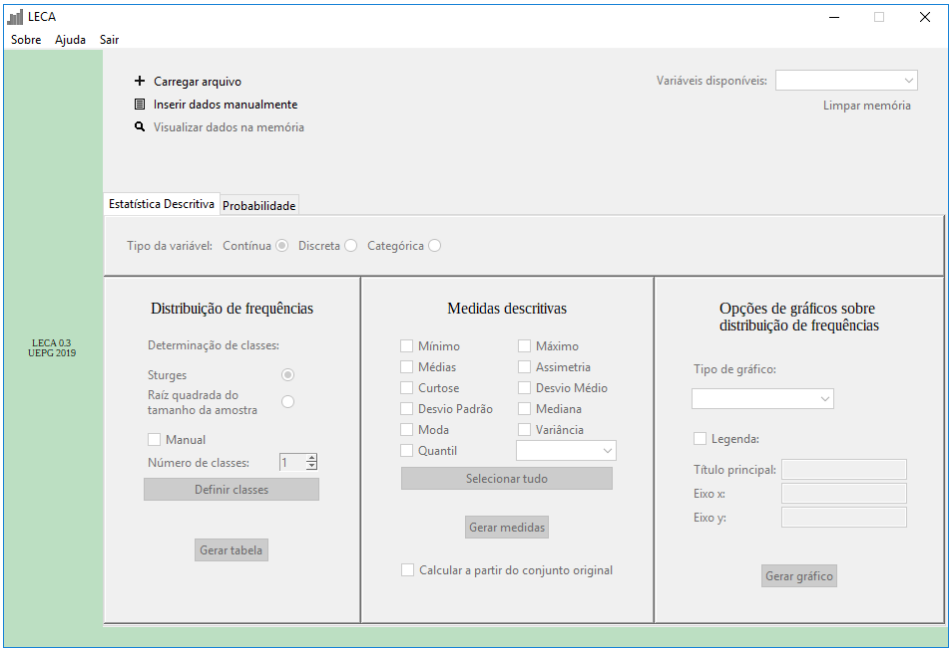


Figura 1 – Opções de estatística descritiva na janela principal do LECA

Estatística descritiva

Todas as opções de estatística descritiva são desabilitadas por padrão até que uma variável seja carregada ou inserida na memória do pacote. Essas funções são habilitadas quando aplicáveis sobre a variável que está sendo utilizada. Isso não só evita erros por parte do usuário como visa a dedução de conceitos sobre o tema. A habilitação de definição de classes na área de distribuição de frequências, por exemplo, é realizada apenas quando o tipo da variável selecionada pelo usuário for contínua, implicando que apenas nessa opção utiliza-se esse conceito.

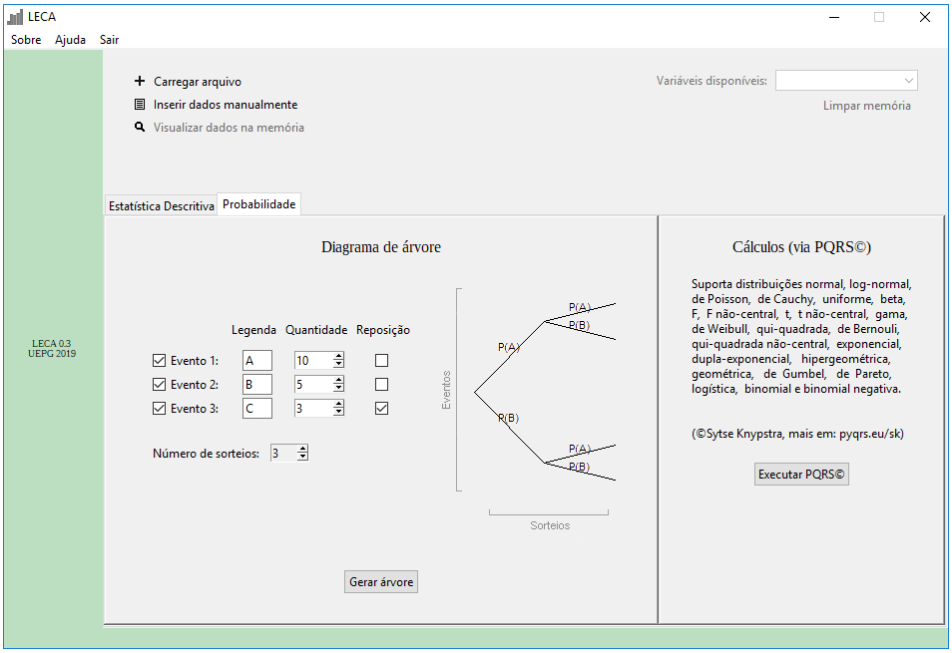
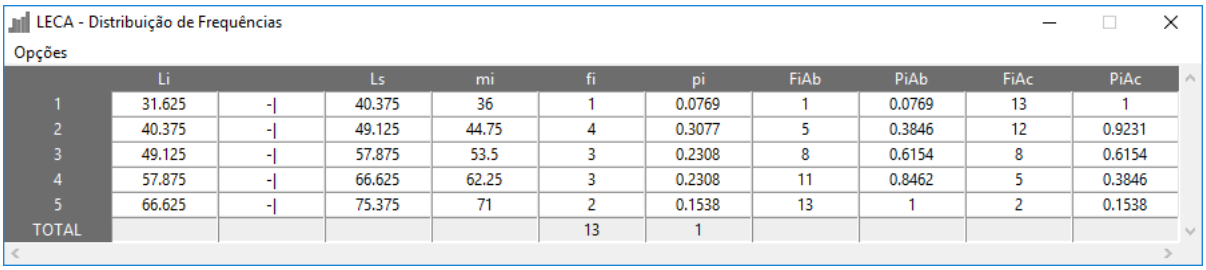


Figura 2 – Opções de probabilidade na janela principal do LECA

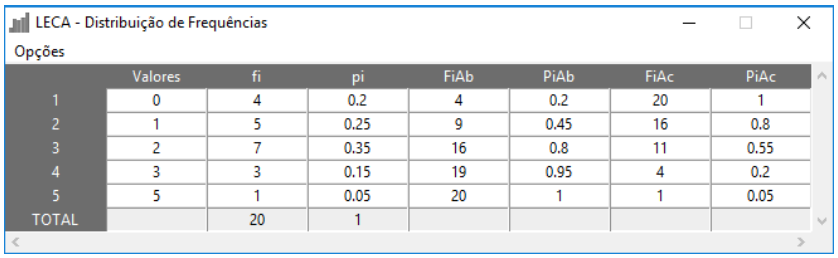
A elaboração de tabelas de distribuição de frequências se dá de duas maneiras: uma para variáveis aleatórias contínuas e outra para variáveis aleatórias discretas ou categóricas (Figura 3). Quando a tabela de frequências é gerada com a opção "Contínua" como tipo de variável, são apresentados os limites inferiores (Li), limites superiores (Ls), pontos médios (mi), frequências absolutas simples (fi), relativas simples (pi), absolutas acumuladas acima (FiAc), relativas acumuladas acima (PiAc), absolutas acumuladas abaixo (FiAb) e relativas acumuladas abaixo (PiAb) de cada classe. A determinação das classes que a tabela de frequências de uma variável contínua deve ter é dada pelo arredondamento de um número **k**, que possui diversas formas de ser calculado. No pacote, esse cálculo se dá por meio de três métodos: Sturges (MARTINS e DOMINGUES, 2014), raiz quadrada do tamanho da amostra ou de forma manual, com o valor de **k** sendo escolhido pelo usuário ou de forma automática (**k = (máx – min)/n**, onde **máx** e **min** são o maior e menor valor assumidos pela variável, respectivamente, e **n** é o tamanho da amostra). Quando a tabela de frequências é gerada com a opção "Discreta" ou com a opção "Categórica" como tipo de variável, o conceito de classes como previamente explicado não é utilizado e todas as frequências são apresentadas por observação.

Medidas descritivas implementadas no pacote podem ser visualizadas na coluna central da Figura 1 e a apresentação de seus resultados na Figura 4. Elas podem ser calculadas sobre a tabela de frequências gerada pelo pacote ou sobre o conjunto de dados original.



	Li		Ls	mi	fi	pi	FiAb	PiAb	FiAc	PiAc
1	31.625	-|	40.375	36	1	0.0769	1	0.0769	13	1
2	40.375	-|	49.125	44.75	4	0.3077	5	0.3846	12	0.9231
3	49.125	-|	57.875	53.5	3	0.2308	8	0.6154	8	0.6154
4	57.875	-|	66.625	62.25	3	0.2308	11	0.8462	5	0.3846
5	66.625	-|	75.375	71	2	0.1538	13	1	2	0.1538
TOTAL					13	1				

(a)



	Valores	fi	pi	FiAb	PiAb	FiAc	PiAc
1	0	4	0.2	4	0.2	20	1
2	1	5	0.25	9	0.45	16	0.8
3	2	7	0.35	16	0.8	11	0.55
4	3	3	0.15	19	0.95	4	0.2
5	5	1	0.05	20	1	1	0.05
TOTAL		20	1				

(b)

Figura 3 – Formato das tabelas de frequências apresentadas pelo LECA. Distribuição de frequências de uma variável (a) contínua e (b) discreta ou categórica.

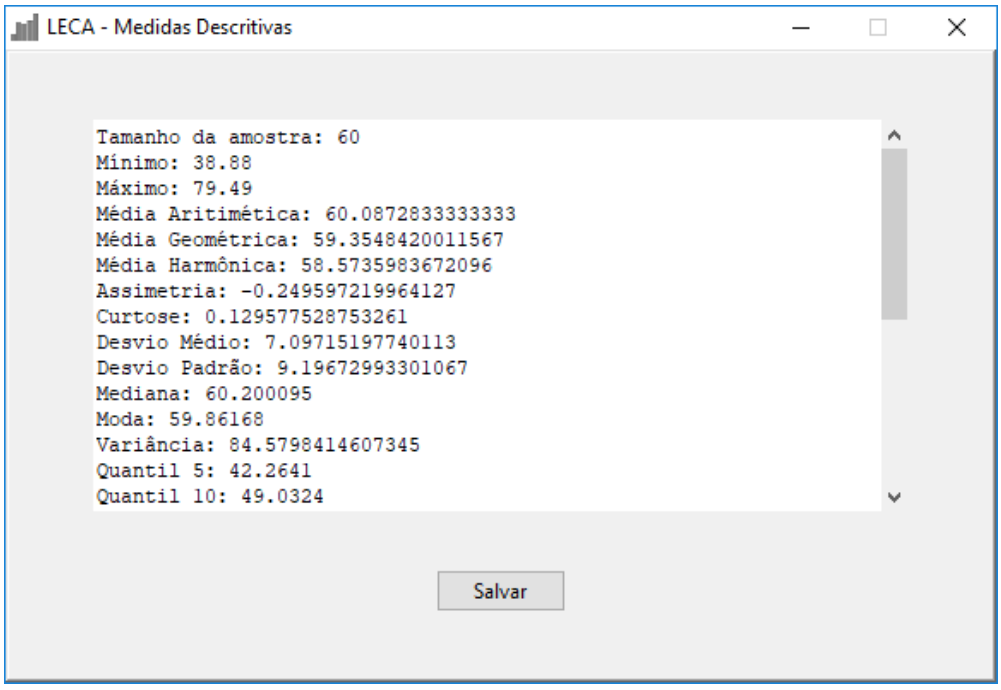
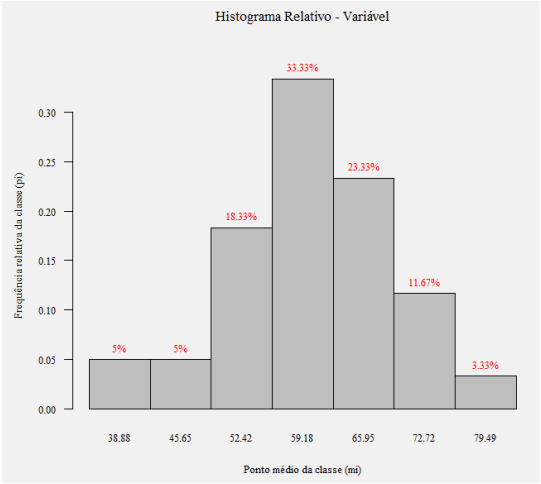
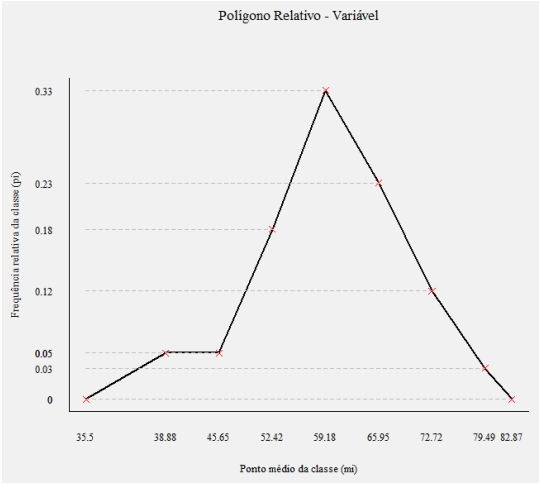


Figura 4 – Apresentação de resultados das medidas descritivas no LECA

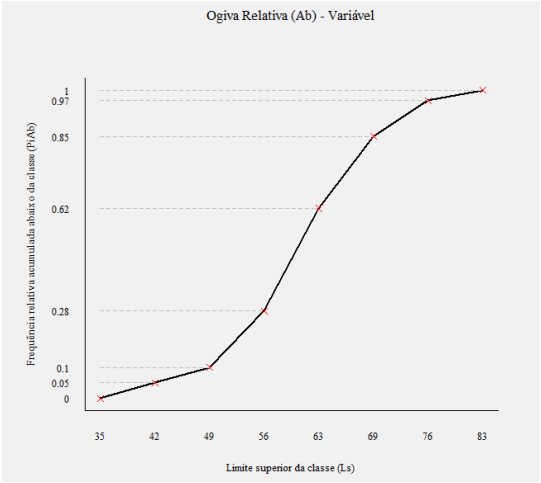
Assim como com as medidas descritivas, os gráficos são habilitados apenas de acordo com o tipo da variável na memória do LECA. Os gráficos disponíveis no pacote são histograma, polígono de frequência, ogiva, gráfico de escada, de barra e de setores e estão todos ilustrados na Figura 5. Para variáveis contínuas, os gráficos disponíveis são histograma, polígono de frequência e ogiva. Para variáveis discretas, os gráficos disponíveis são o de escada, de barras e de setores. Para variáveis categóricas, os gráficos disponíveis são o de barra e o de setores.



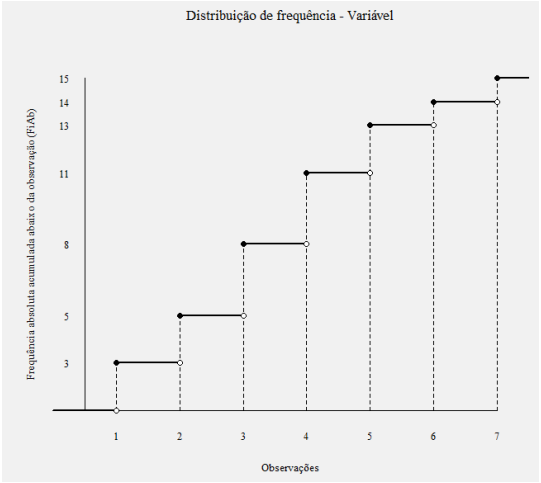
(a)



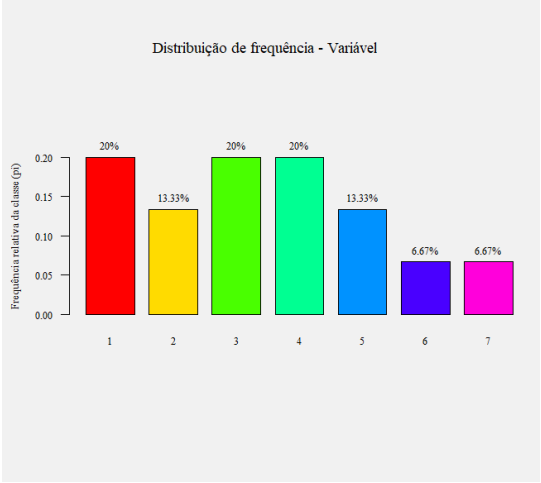
(b)



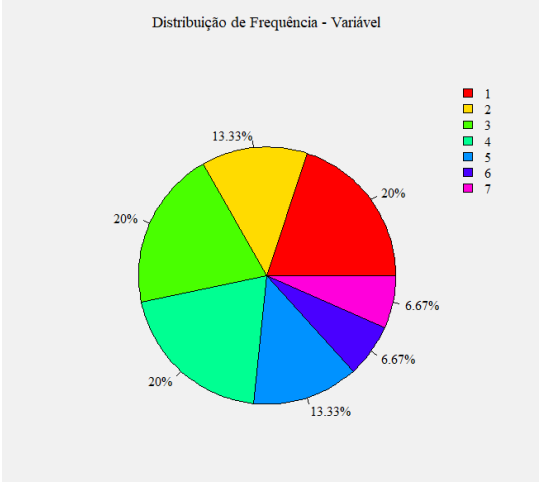
(c)



(d)



(e)



(f)

Figura 5 – Tipos de representações gráficas oferecidas pelo LECA. (a) Histograma, (b) polígono de frequência, (c) ogiva, (d) gráfico de escada, (e) gráfico de barras e (f) gráfico de setores.

Probabilidade

Na aba de probabilidade estão disponíveis o gráfico de árvore e o acesso ao software PQRS©. O diagrama de árvore (Figura 6) não faz uso de dados armazenados na memória do LECA. Os valores utilizados podem ser inseridos nas opções de probabilidade, como o número de sorteios e o uso ou não de reposição nesses sorteios, tonando possível a formulação de diversos tipos de cenário pelo usuário. O diagrama gerado apresenta a probabilidade de cada evento em cada sorteio assim como a probabilidade total de cada ramo.

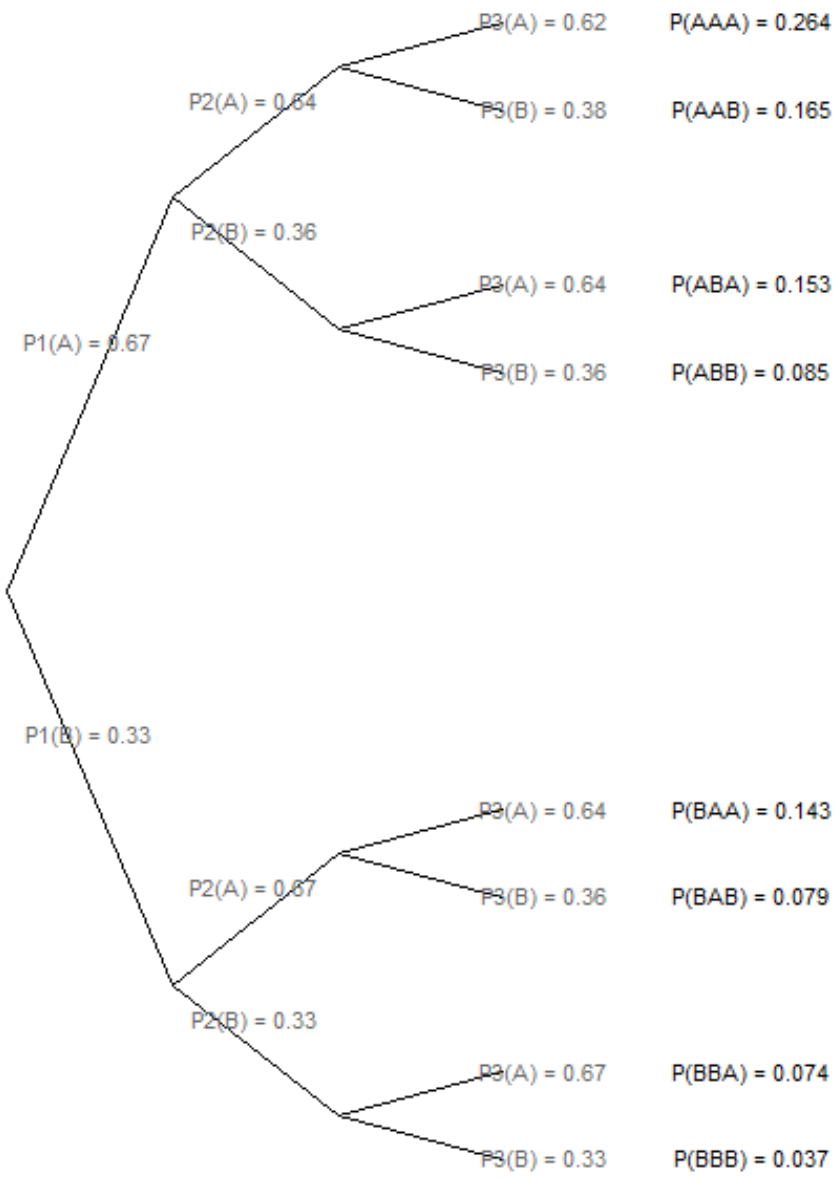
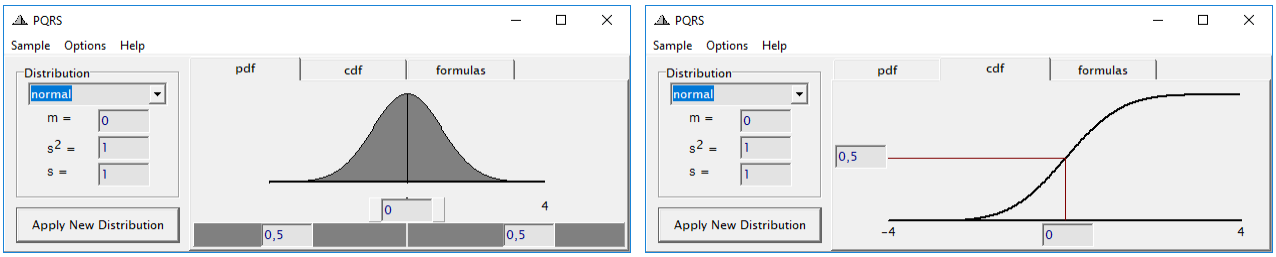


Figura 6 – Diagrama de árvore oferecido pelo LECA. O cenário proposto envolve três sorteios para dois eventos de quantidades 10 e 5 com opção de reposição para o último evento.

O software gratuito PQRS© foi desenvolvido por Sytse Knypstra (2000) e agrupa funções de probabilidade suportando uma grande quantidade de modelos de distribuições para cálculos. O programa inclui gráficos interativos (Figura 7) que facilitam a visualização de conceitos de estatística básica, como o cálculo de percentis, de probabilidade pontual e acumulada e p-valor, e apresenta resultados de forma prática. Ele também inclui um formulário em português brasileiro no menu de ajuda do LECA.



(a) (b)

Figura 7 – Gráficos interativos no PQRS©. (a) Função de distribuição acumulada e (b) função densidade de probabilidade para distribuição normal.
Fonte: Knypstra (2000).

A versão 2 do PQRS© é embutida no pacote, podendo ser utilizada dentro do ambiente R por meio do botão de execução na aba de probabilidade, que simplesmente executa o software na máquina do usuário. O programa não faz parte do desenvolvimento do LECA e todos os direitos autorais pertencem ao seu desenvolvedor. Para utilizar o PQRS© fora do ambiente R, basta fazer o download por meio do endereço pyqrs.eu/sk/.

Considerações finais

O pacote estatístico em R, LECA, oferece funções de estatística descritiva e probabilidade em uma interface gráfica interativa em português brasileiro com foco na aprendizagem. Ele pode ser facilmente utilizado como material de apoio mesmo por aqueles que não possuem conhecimento avançado sobre a linguagem R.

Um software como o LECA pode desmistificar a dificuldade da utilização da estatística por estudantes e pesquisadores que não são da área, podendo também estimular o interesse pelo software R e o desenvolvimento de novas produções em seu ambiente.

O pacote resultado desse trabalho ainda não se encontra no repositório oficial do R, *The Compreenshive R Network* (CRAN), pois novas opções de interface para probabilidade e testes de hipótese estão em desenvolvimento. Assim que essas funcionalidades estiverem implementadas, pretende-se disponibilizá-lo para que ele possa ser utilizado como material de apoio na educação de estatística básica.

Agradecimentos

Agradecimentos a Universidade Estadual de Ponta Grossa e a Fundação Araucária pelo incentivo à pesquisa e pelas bolsas concedidas dentro do Programa Institucional de Bolsas de Iniciação Científica no período de desenvolvimento do pacote.

Referências Bibliográficas

ARNHOLD, E. *Pacote em ambiente R para automatizar estatísticas descritivas.* Sigmae, 2014. v.3, n.1, p.36-42.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica***.** 8ª ed., São Paulo: Saraiva, 2013. 548 p.

DALGARRD, P. *The R-Tcl/Tk interface.* Proceedings of the 2nd International Workshop on Distributed Statistical Computing, 2001a. ISSN 1609-395X.

DALGARRD, P. *A Primer on the R-Tcl/Tk Package.* R News, 2001b. v.1/3, p.27-31. ISSN 1609-3631.

FARIA, J.C.; JELIHOVSCHI, E. G.; ALLAMAN, I. B. *fdth:* Frequency Distribution Tables, Histograms and Polygons. UESC, Bahia, Brasil, 2016. Disponível em: https://CRAN.R-project.org/package=fdth. Acesso em: 04 de março de 2019.

FERREIRA, E. B.; CAVALCANTI, P. P.; NOGUEIRA, D. A. *ExpDes:* An R Package for ANOVA and Experimental Designs. Applied Mathematics, 2014. v.5, p.2952-2958.

FOX, J. *The R Commander:* A Basic Statistics Graphical User Interface to R. Journal of Statistical Software, 2005. v.14, n.9, p.1-42.

FOX, J. *Using the R Commander:* A Point-and-Click Interface of R. Boca Raton FL: Chapman and Hall/CRC Press, 2017. 233 p.

FOX, J; BOUCHET-VALAT, M. *Rcmdr:* R Commander. 2018. Pacote em R versão 2.5-1. Disponível em: https://CRAN.R-project.org/package=Rcmdr. Acesso em: 04 de março de 2019.

GROSJEAN, P. *SciViews:* A GUI API for R. UMONS, Mons, Belgium, 2014. Disponível em: http://sciviews.org/SciViews-R. Acesso em: 04 de março de 2019.

KNYPSTRA, S. *PyQRS*. 2000. Disponível em: https://pyqrs.eu/sk/. Acesso em: 10 de março de 2019.

KNYPSTRA, S. *Inference as a dynamic concept map*. 1999.

LARSON, R.; FARBER, B. *Estatística Aplicada.* 4ª ed., São Paulo: Pearson Prentice Hall, 2010. 637 p.

LAWRENCE, M.; VERZANI, J. *ProgGUIinR:* support package for "Programming Graphical User Interfaces in R". 2014. Pacote em R versão 0.0-4. Disponível em: https://CRAN.R-project.org/package=ProgGUIinR. Acesso em: 04 de março de 2019.

LAWRENCE, M.; VERZANI, J. *Programming Graphical User Interfaces in R.* CRC Press, 2016. 479 p.

MARTINS, G. A.; DOMINGUES, O. *Estatística Geral e Aplicada.* 5ª ed., São Paulo: Atlas, 2014. 416 p.

OUTERSHOUT, J. K. *TCL:* An Embeddable Command Language. Proceedings of the 1990 Winter USENIX Conference, 1989. p. 133-146.

R CORE TEAM. *R:* A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. Disponível em: https://r-project.org. Acesso em: 04 de março de 2019.

RSTUDIO TEAM. *RStudio:* Integrated Development for R. RStudio, Inc., Boston, 2016. Disponível em: https://rstudio.com. Acesso em: 04 de março de 2019.

VALERO-MORA, P. M.; LEDESMA, R. D. *Graphical User Interfaces for R.* Journal of Statistical Software, 2012. v.49, n.1, p.1-8.