DESENVOLVIMENTO DE BIBLIOTECA ESTATÍSTICA EM LINGUAGEM R PARA ESTIMATIVA DE MEDIDAS CARACTERÍSTICAS DE VARIÁVEIS CONTÍNUAS PARA DADOS AGRUPADOS EM TABELA DE FREQUÊNCIA.

Matheus Moresco (PIBIC/Fundação Araucária/UEPG), Jorim Sousa das Virgens Filho (Orientador), e-mail: jvirgens@uepg.br.

Universidade Estadual de Ponta Grossa/Sexatas/Demat/LECA.

Ciência da Computação/ Metodologia e Técnicas da Computação

Palavras-chave: Dados e Estatísticos, Tabela de Frequência, Linguagem R.

Resumo

A análise de frequências de dados é um tema muito abordado na área da estatística. São de suma importância para a agrupação e organização de dados. Assim, é de grande importância ter um programa que realize cálculos estatísticos de maneira simples e rápida. Para análises estatísticas o software R se mostra uma das mais eficientes alternativas. Porém não possui funcionalidades que resolvam os mais diversos problemas de maneira eficiente. Diante disto, este trabalho objetivou desenvolver um pacote que auxiliasse em analises de frequências de dados estatísticos com uma interface gráfica que facilitasse o acesso do usuário às funcionalidades do R, uma vez que nem todos os usuários dominam a linguagem. Os resultados mostraram que o pacote desenvolvido, funcionou adequadamente e se mostrou útil no auxilio do ensino de análise exploratória de dados estatísticos.

Introdução

A relação entre o ser humano e os computadores vem aumentando consideravelmente nos últimos anos, e aos poucos vem se tornando indispensável no dia a dia das pessoas. Assim, é de suma importância a criação de novos métodos que facilitem o uso desses recursos computacionais.

De acordo com Braga (2004) o uso de interfaces gráficas facilita muito no manuseio de certas ferramentas computacionais e, uma interface amigável e bem projetada aumenta consideravelmente a aceitação de um programa pelos usuários, se tornando um importante fator no desenvolvimento de um software.

O R é um ótimo software para análises estatísticas, porém, sua interface não é muito amigável, o que dificulta a utilização do mesmo por pessoas que não estão familiarizadas com esse ambiente. Para solucionar este problema, são criados pacotes para o software que facilitam a programação, o quais também podem adaptar novas interfaces gráficas, para que as pessoas que não sabem programar em no ambiente R também possam usar a ferramenta (Valero-Mora, 2012).

Como exemplo de pacotes que geram interfaces temos o arvoRe, uma interface gráfica projetada para análise de custo e efetividade, realizando cálculos com base nos conceitos de arvore de decisão. A interface foi desenvolvida na linguagem tcl/tk e projetada Prestes e Camey (2013). Também pode-se observar o pacote deducer, que é uma interface em R que tem por objetivo facilitar o manuseio da ferramenta, abrangendo na interface, grande parte das funcionalidades do R. Este pacote foi projetado na linguagem JGR por Fellows (2012). Ainda, também é importante citar o Rcmdr um pacote bem conhecido e muito utilizado pelos usuários

do R, cuja interface tem por objetivo otimizar e facilitar o uso das funcionalidades do programa, e foi desenvolvida por Fox (2005).

Algo comum na estatística é a análise de frequências em um conjunto de dados. O processamento dos dados baseia-se em montar uma tabela de frequência e realizar os cálculos necessários por meio dos dados fornecidos pela tabela. Em um conjunto com poucos dados, é fácil a obtenção da tabela, porém, quando existem muitos dados a se analisar, acaba sendo muito difícil realizar este processo manualmente. Assim, o ambiente R se apresenta como uma boa alternativa para a realização de cálculos desse tipo. Porém, o R não possui uma interface gráfica que facilite este processo, o que dificulta novos usuários a utilizá-lo efetivamente.

Diante deste contexto, o objetivo deste trabalho foi desenvolver um pacote em R que gere uma interface gráfica para a análise descritiva de dados, e que também possa ser utilizado com fins didáticos na área de Estatística, de maneira que os usuários possam realizar os cálculos de uma maneira simples e rápida.

Materiais e Métodos

Este trabalho foi desenvolvido no Laboratório de Estatística Computacional e Aplicada-LECA da UEPG. Diante da popularidade do software R no meio acadêmico, o mesmo foi escolhido para desenvolvimento deste projeto. Inicialmente o projeto foi focado na pesquisa de funções que criam pacotes em R, para na sequência se iniciar o desenvolvimento de algoritmos para implementação das funções a serem codificadas no ambiente de programação.

Na sequência, iniciou-se a implementação de algoritmos para as funções estatísticas na linguagem R que pudessem automatizar o processo de análise exploratória de dados. Assim, foram geradas várias funções que fazem análise de frequências de um conjunto de dados que depois de finalizadas foram reunidas para compilação num pacote (biblioteca), que pudesse ser carregado no ambiente R com as funções a serem executadas pelo usuário.

Depois de implementadas as funções estatísticas, deu-se início a elaboração de uma interface gráfica, utilizando para tanto a linguagem tcl/tk, a partir de um modelo sugerido por Prestes e Camey (2013). Com as funcionalidades disponíveis no pacote tcl/tk2 disponível no R, foram criadas as interfaces gráficas, as quais foram submetidas a vários testes sucessivos para correção dos eventuais erros encontrados. A partir das funções prontas foi criado um pacote que quando carregado no R disponibiliza todas as funcionalidades propostas neste do projeto ao usuário.

Resultados e Discussão

Como resultado do projeto obteve-se um pacote em R, que contém funções estatísticas relacionadas a análise de frequência de dados, e também funções que geram interfaces gráficas que executam as funções estatísticas.

Na Figura 1 observa-se as funcionalidades principais da interface, onde é possível carregar um conjunto de dados salvo num arquivo do tipo ".csv" na opção abrir arquivo. A interface também permite que sejam salvos, todos os resultados obtidos e também os gráficos realizados pelo programa.

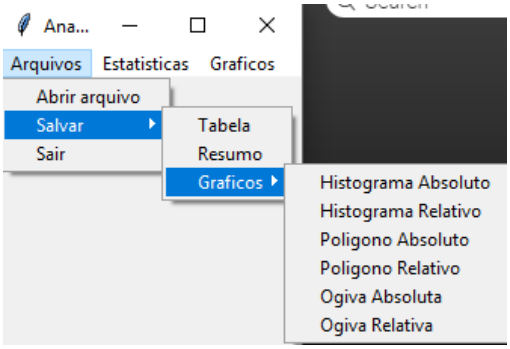


Figura 1- Menu principal da interface

Na Figura 2 são apresentadas as opções estatísticas disponibilizadas pela interface. Nas Figuras 3 e 4 são mostrados uma tabela de frequência e um resumo das estatísticas requisitadas pelo usuário.

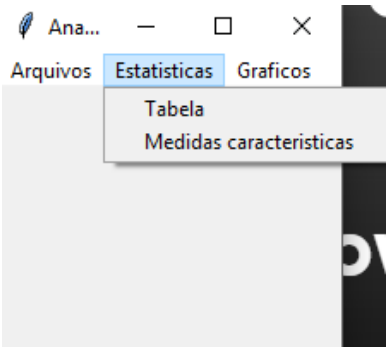
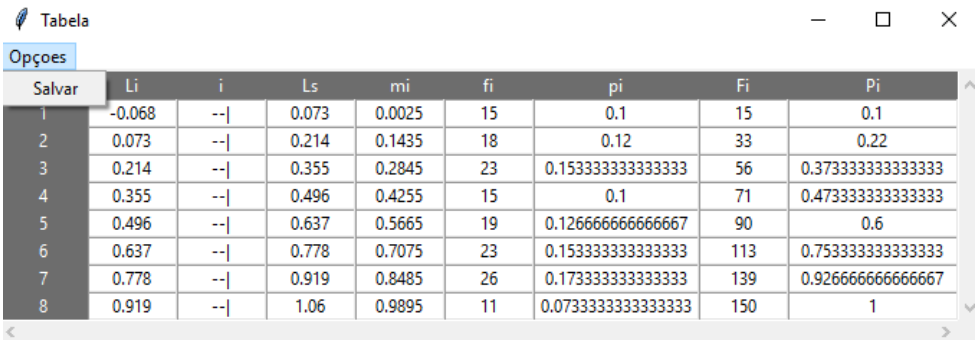


Figura 2- Opções de dados estatísticos disponíveis



	Li	i	Ls	mi	fi	pi	Fi	Pi
1	-0.068	--|	0.073	0.0025	15	0.1	15	0.1
2	0.073	--|	0.214	0.1435	18	0.12	33	0.22
3	0.214	--|	0.355	0.2845	23	0.153333333333333	56	0.373333333333333
4	0.355	--|	0.496	0.4255	15	0.1	71	0.473333333333333
5	0.496	--|	0.637	0.5665	19	0.126666666666667	90	0.6
6	0.637	--|	0.778	0.7075	23	0.153333333333333	113	0.753333333333333
7	0.778	--|	0.919	0.8485	26	0.173333333333333	139	0.926666666666667
8	0.919	--|	1.06	0.9895	11	0.0733333333333333	150	1

Figura 3- Tabela criada pela interface

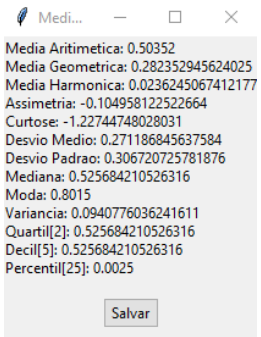


Figura 4- Medidas características calculadas pelo programa

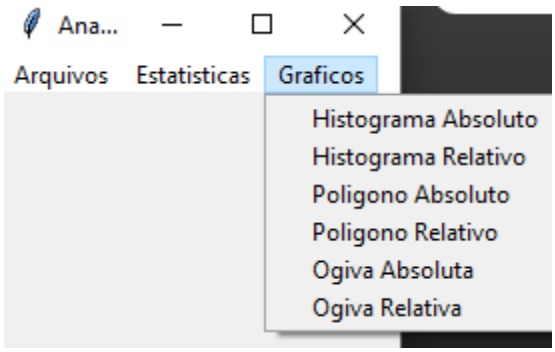


Figura 5- Opções de gráficos que são gerados pelo programa

A Figura 5 mostra os tipos de gráficos estatísticos disponibilizados pelo programa. Estes gráficos podem ser abertos no R ou também podem ser salvos em formato PDF. Pode-se perceber ao usar o pacote desenvolvido neste projeto, que o mesmo atende as expectativas, uma vez que poupa o trabalho do usuário na hora de manipular os dados, já que apresenta os resultados prontos, sendo necessário apenas que o usuário entre com os dados.

A interface se mostrou extremamente amigável ao usuário, principalmente para usuários que não têm familiaridade programação de computadores, principalmente com a programação em R. Além do mais, por não requerer maiores conhecimentos da linguagem R, didaticamente, o pacote pode facilitar o aprendizado de pessoas que ainda não tem conhecimento acerca da Ciência Estatística.

Conclusões

A biblioteca desenvolvida se mostrou eficiente na realização de cálculos estatísticos baseados na distribuição de frequência de dados. A criação da interface gráfica se mostrou de fácil manuseio, podendo auxiliar, tanto profissionais da área quanto alunos que estão iniciando o aprendizado na área estatística.

Agradecimentos

Ao Prof. Dr. Jorim Sousa das Virgens Filho, à Profª. Drª. Deyse Gebert, ao meu colega Eduardo Cordeiro e à Fundação Araucária pelo apoio na concessão de bolsas.

Referências

- BRAGA, A. S. **Desing de Interface.** São Paulo, 2004; Disponível em: < http://www.pucsp.br/~braga/dissertacao.pdf >
- FELLOWS, I. **Deducer: A Data Analysis GUI for R.** Jornal of Statistical Sftware. Vol. 49. 2012.
- FOX, J. **The R Commander: A Basic-Statistics Graphical User Interface to R.** Jornal of Statisitical Software. Vol. 14. 2005.
- PRESTES, I.; CAMEY, S. A.; **Implementação de uma interface gráfica para Análise de Custo de Efetividade no R.** Jornal of Statistical Software. Porto Alegre. 2013.
- VALERO-MORA, P. M.; LEDESMA, R. D.; **Graphical User Interfaces for R.** Jornal of Statistical Sftware. Vol. 49. 2012.