DESENVOLVIMENTO DO PACOTE ESTATÍSTICO, EM LINGUAGEM R, PARA ANÁLISE EXPLORATÓRIA DE DADOS: TABELA DE FREQUÊNCIAS E GRÁFICOS PARA VARIÁVEIS CONTINÚAS.

Eduardo Guilherme Cordeiro (PROVIC) Deyse Márcia Pacheco Gebert (Orientadora), e-mail: dmpgebert@uepg.br.
Universidade Estadual de Ponta Grossa/Departamento de Informática.

Probabilidade e Estatística; Probabilidade e Estatística Aplicadas.

Palavras-chave: software R, pacotes, estatística, LECA.

Resumo
O Software estatístico R é muito utilizado nas mais diversas áreas do conhecimento. Algumas das vantagens desse software são: ser gratuito e de código aberto. Com isso é possível que o desenvolvimento de bibliotecas/pacotes seja livre e os desenvolvedores possam contribuir com suas implementações e disponibilizá-las para o público. Os pacotes já existentes na parte da estatística direcionada à análise exploratória apresentam poucas opções em funções para tabelas de distribuição de frequências, o que dificulta seu uso em sala de aula. Devido a esta carência objetivou-se na construção do pacote estatístico LECA para fornecer mais opções no momento da análise exploratória, tanto para uso em sala de aula, como para pesquisas. Este pacote disponibiliza funções para a construção de uma tabela de distribuição de frequências completa e também a elaboração de gráficos provenientes da mesma. Com isso criou-se uma biblioteca mais didática e de fácil utilização, pois o público usuário em maior parte será dado por alunos de graduação que ainda não estão familiarizados com o ambiente R.

Introdução

Em meados de 1990, os professores de estatística Ross Ihaka e Robert Gentleman deram início ao desenvolvimento da linguagem R devido a necessidade de um melhor ambiente de software para laboratório de ensino (BECKER, 1994).

A linguagem R criada por esses professores é um ambiente livre, ou seja, obedece aos termos GNU (Gnu's Not Unix) (WHAT IS GNU?, 2017), que são: A liberdade de executar o programa como desejar, para qualquer finalidade; A liberdade de estudar como o programa funciona, e adaptá-lo às suas necessidades. O acesso ao código fonte é uma condição prévia para isso. A liberdade de redistribuir cópias para que você possa ajudar o seu vizinho. A liberdade de melhorar o programa, e liberar suas melhorias para o público, de modo que toda a comunidade se beneficie.

Devido a isso é possível o desenvolvimento de bibliotecas e os desenvolvedores podem contribuir com suas implementações e disponibilizá-las para o público. Tais bibliotecas/pacotes podem ser encontradas, por qualquer usuário, nos servidores de "download" CRAN (The Comprehensive R Archive Network, 2017).

Os pacotes disponibilizados pelo CRAN atendem diversas áreas com diversos temas, mas tem sua grande utilização na área estatística, e dentro da estatística existe um nicho específico denominado análise exploratória de dados. Esse nicho é o foco de alguns pacotes como "Rcmdr" (FOX, 2005) e "ds" (ARNHOLD, 2014),

porém estes apresentam poucas opções em funções para tabelas de distribuição de frequências, o que dificulta seu uso em sala de aula.

A importância da tabela de distribuição de frequências se dá quando o pesquisador quer estudar o comportamento de determinada variável analisando a ocorrência de suas possíveis realizações (BUSSAB; MORETTIN, 2013). Essa análise é possível através do cálculo dos diversos tipos de frequências que compõem essa tabela, porém os pacotes conhecidos ainda não apresentam essas opções de forma didática.

Com isso verificou-se a necessidade da construção de um pacote estatístico, mais didático, que forneça mais opções na construção de tabelas de distribuição de frequências para variáveis contínuas, cálculos de medidas e gráficos feitos a partir desta tabela. Desta forma a análise exploratória de dados se torna mais completa e de fácil entendimento, o que é extremamente importante para um estudo.

Materiais e Métodos

Para a construção das funções estatísticas desejadas, foi necessário utilizar um ambiente que abrangesse a linguagem R, tal ambiente foi o software R (R Core Team, 2017). Este software apresenta um desenvolvimento através de linhas de códigos, com interface simples e ferramentas pouco intuitivas, tornando assim o desenvolvimento de funções mais árduo e lento. Por esse motivo optou-se por uma interface mais amigável do programa chamada RStudio, disponível no site oficial (TAKE CONTROL OF YOUR R CODE, 2016).

A solução algorítmica para a construção da tabela de distribuição de frequências inicia-se com a contagem do número de observações (*n*), seguindo com o cálculo do número de classes (*k*) através da formula de "Sturges" (MARTINS, 2014): $k = 1 + 3{,}3\log_{10}(n)$,sendo *n* o tamanho da amostra.

Em seguida foi determinada a amplitude total dos dados (*A*) dada pela diferença entre o valor da maior (max) e menor (min) observação. Após isso a amplitude de classe (*h*) foi determinada por meio da seguinte equação: $h = A/(k-1)$.

Para a construção das classes foi necessário determinar inicialmente o limite inferior da primeira classe (L_i) pela equação: $L_i = \min - (h/2)$. Os limites superiores (L_s) de todas as classes foram definidos pela equação $L_s = L_i + h$. O limite superior de uma determinada classe se torna o limite inferior da classe seguinte, sendo o intervalo aberto em L_i e fechado em L_s. Após a determinação dos limites das classes o seu ponto médio foi calculado por meio da equação $m_i = (L_i + L_s)/2$.

A frequência absoluta (f_i) foi determinada fazendo-se a contagem do número de observações encontradas dentro dos limites das classes definidos anteriormente. Em seguida utilizando a frequência absoluta determinou-se a frequência relativa ($p_i = f_i / n$). A frequência acumulada absoluta abaixo do limite superior de cada classe ($F_i = \Sigma_{j=1}^{i} f_j$), que é o somatório das frequências absolutas até a classe que está sendo calculada. A frequência relativa absoluta abaixo do limite superior foi calculada por meio da equação $P_i = \Sigma_{j=1}^{i} p_j$, que é o somatório das frequências relativas até a classe que está sendo calculada, completando assim a construção total da tabela através de uma repetição de cálculos até a última classe.

Utilizando as informações computadas na tabela de distribuição de frequências foram implementados códigos para criação de um gráfico de colunas, o

histograma, que foi construído utilizando o ponto médio no eixo das abscissas e a frequência absoluta ou relativa no eixo das ordenadas. O polígono de frequência foi construído utilizando os mesmos parâmetros, diferenciando-se do histograma por ser um gráfico de linha. As ogivas, que também são gráficos de linhas, foram criadas utilizando-se os limites superiores das classes no eixo X e a frequência absoluta ou relativa acumulada no eixo Y.

Resultados e Discussão

Foi desenvolvido um pacote de funções, no qual por meio do comando "tabela", o usuário necessitará inserir um conjunto de dados e definir o arredondamento a ser utilizado. Essa função retornará um data frame com a tabela de distribuição de frequências completa (Figura 1).

```
> tabela(dados,2)
    Li   i   Ls   mi fi   pi  Fi   Pi
1 2.61 --| 3.15 2.88  4 0.04   4 0.04
2 3.15 --| 3.69 3.42 12 0.12  16 0.16
3 3.69 --| 4.23 3.96 14 0.14  30 0.30
4 4.23 --| 4.77 4.50 32 0.32  62 0.62
5 4.77 --| 5.31 5.04 20 0.20  82 0.82
6 5.31 --| 5.85 5.58 12 0.12  94 0.94
7 5.85 --| 6.39 6.12  4 0.04  98 0.98
8 6.39 --| 6.93 6.66  2 0.02 100 1.00
```

Figura 1 - Tabela de distribuição de frequência gerada pelo comando "tabela" do pacote LECA.

Utilizando o comando "gráfico", o usuário precisará inserir um número que é referente ao tipo de gráfico (1 - Gera histograma com frequência absoluta; 2 – Gera histograma com frequência relativa; 3 – Polígono com frequência absoluta; 4 – Polígono com frequência relativa; 5 – Ogiva com frequência absoluta; 6 – Ogiva com frequência relativa) e a tabela gerada anteriormente (Figura 2, 3 e 4).

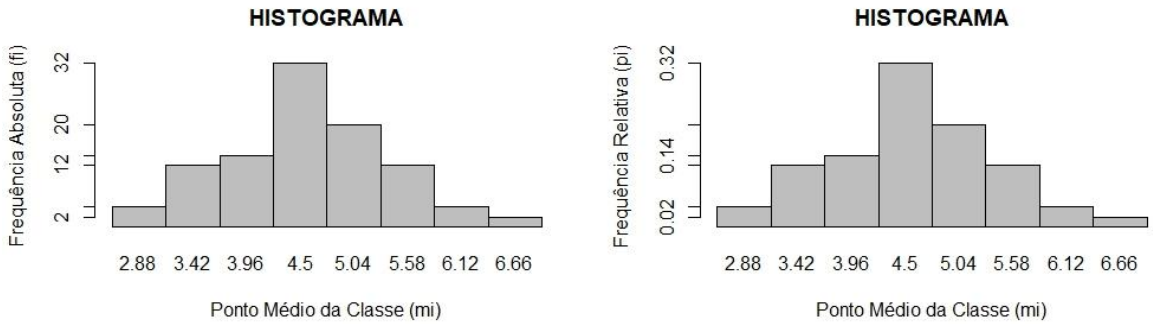


Figura 2. Histogramas gerados pelo comando gráfico do pacote LECA.

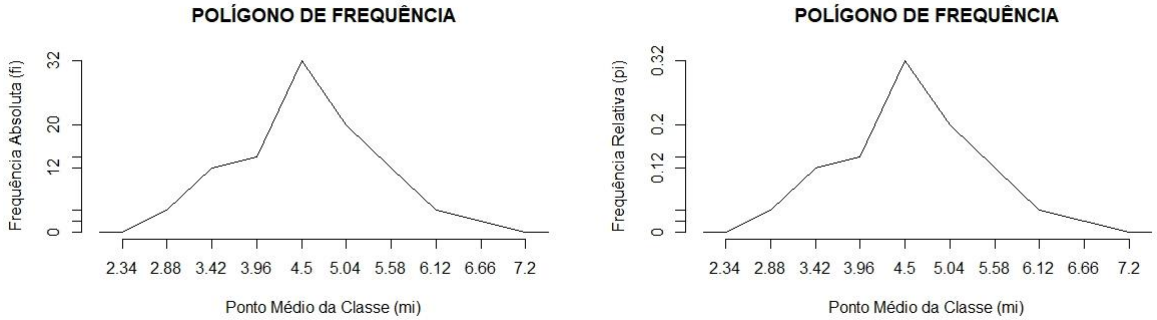


Figura 3. Polígonos de frequência gerados pelo comando gráfico do pacote LECA.

Na figura 2 estão os histogramas, que se diferem apenas pela frequência que utilizam na sua construção, sendo o primeiro frequência absoluta e o segundo frequência relativa.

Na figura 3 estão os polígonos de frequências que diferem entre si pelo mesmo motivo, o tipo de frequência utilizada. Nas ogivas (Figura 4) utilizam-se os limites superiores das classes, no eixo das abscissas, e no eixo das ordenadas as frequências acumuladas absoluta ou relativa, respectivamente.

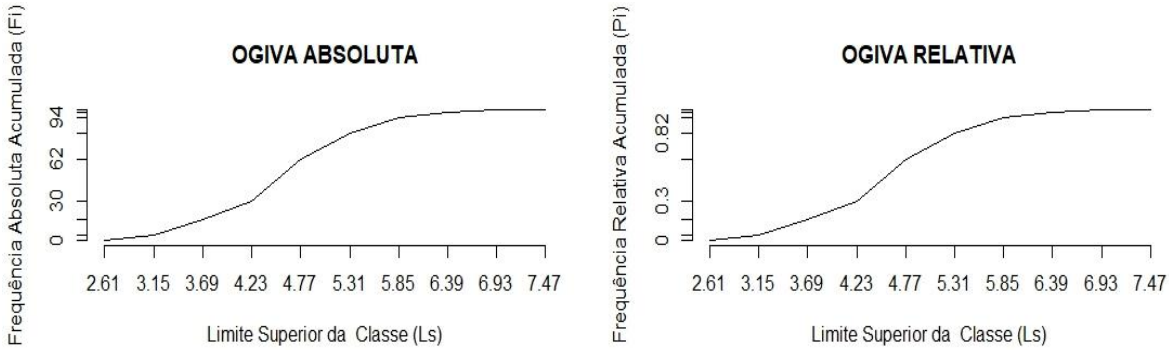


Figura 4. Ogivas geradas pelo comando gráfico do pacote LECA.

Conclusão

O pacote LECA disponibiliza funções que facilitam a análise estatística descritiva de conjuntos de dados, bem como a sua utilização em sala de aula, pois foi elaborado de forma mais didática. Todavia que o público usuário em maior parte será dado por alunos de graduação que ainda não estão familiarizados com o ambiente R.

Referências

ARNHOLD, E. Pacote em ambiente R para automatizar estatísticas descritivas. **Sigmae**, v. 3, p. 36-42, 2014. Disponivel em: <http://CRAN.R-project.org/package=ds>. Acesso em: 16 agosto. 2017.

BECKER, R. **A brief history of S. Technical Report 11**, AT&T Bell Laboratories, 1994.

BUSSAB, W.de O.; MORETTIM, P.A. **Estatística Básica**, São Paulo, Saraiva, 8 ed. 2013, 548p.

FOX, J. The R Commander: A Basic Statistics Graphical User Interface to R. **Journal of Statistical Software**, 14(9): 1–42. 2005. Disponivel em:<http://CRAN.R-project.org/package=Rcmdr>. Acesso em: 16 agosto. 2017.

MARTINS, G. **Estatística geral e aplicada**. 5ª. ed. Atlas: 2014.

R Core Team (2017). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing,Vienna,Austria.URL-https://www.R-roject.org/.

Take Control of Your R Code.Disponível em: <https://www.rstudio.com/products/RStudio/>. Acesso em: 23 abril. 2017.

The Comprehensive R Archive Network. Disponível em: < https://cran.r-project.org/>. Acesso em: 16 agosto. 2017.

What is GNU?. Disponível em: < http://www.gnu.org/>. Acesso em: 10 de abril de 2017.