Evento: XVIII Jornada de Extensão

CAPACITAÇÃO EM ANÁLISE ESTATÍSTICA DE DADOS UTILIZANDO O SOFTWARE LIVRE R[1]
TRAINING IN STATISTICAL ANALYSIS OF DATA USING FREE SOFTWARE R

Djaina Sibiani Rieger[2], Felipe Micail Da Silva Smolski[3], Tatiane Chassot[4], Denize Ivete Reis[5], Erikson Kaszubowski[6], Iara Endruweit Battisti[7]

[1] Projeto de Extensão da instituição UFFS, Edital nº 522/UFFS/2016
[2] Acadêmica do curso de Engenharia Ambiental e Sanitária, campus Cerro Largo, UFFS, aluna bolsista de extensão/Edital nº 522/UFFS/2016.
[3] Mestrando no Programa de Pós-Graduação em Desenvolvimento e Políticas Públicas da Universidade Federal da Fronteira Sul (UFFS). felipesmolski@hotmail.com
[4] Doutora em Engenharia Florestal, docente da Universidade Federal da Fronteira Sul, campus Cerro Largo. tatianechassot@uffs.edu.br
[5] Doutora em Qualidade Ambiental, docente da Universidade Federal da Fronteira Sul, campus Cerro Largo. denizeir@uffs.edu.br
[6] Doutor em Psicologia. erikson84@yahoo.com.br.
[7] Doutora em Epidemiologia, docente da Universidade Federal da Fronteira Sul, campus Cerro Largo. iara.battisti@uffs.edu.br.

Introdução

A crescente demanda por maior transparência e rigor nos processos científicos exige a constante atualização acerca dos processos estatísticos de referência e dos recursos computacionais de destaque na comunidade acadêmica, necessitando da utilização de programas que permitam agilidade, segurança e possibilitem incrementos na cooperação na produção científica entre os pesquisadores, facilitando a sua divulgação ao público.

A estatística é a ciência que faz parte da matemática, auxilia nos métodos para a coleta, organização, descrição, análise e interpretação de dados, propiciando a utilização dos mesmos na tomada de decisões (CORREA, 2003). De forma complementar, o aumento do poder de processamento e memória dos computadores nos últimos anos favoreceu a análise estatística, ao mesmo tempo em que a linguagem R foi a maior ferramenta criada pela estatística nos últimos vinte anos, explorando todo o poder computacional em um único *software* para várias necessidades (JELIHOVSCHI, 2014).

Diante da era tecnológica atual, as análises estatísticas de dados que anteriormente eram desenvolvidas em planilhas manuais, não mais necessitam de longos períodos de tempo para coleta e apreciação para a consequente produção de resultados, devido ao uso de *softwares* estatísticos (AGUIAR, J. et al, 2016). No entanto, a ciência da computação, em relação a seus avanços recentes, ainda possui limitações com relação à avaliação e publicação de seus resultados nas revistas científicas (PENG, 2011). Contudo, o desenvolvimento recente dos *softwares*, têm contribuído de forma substancial para que se realizem e distribuam pesquisas reprodutíveis na

comunidade de código aberto (KOENKER; ZEILEIS, 2009).

No campo da educação, esforços tem sido empregados para melhorar as práticas científicas no sentido de que se encontrem as evidências do que realmente funciona nesta área do conhecimento, potencializando a confiabilidade dos estudos (MAKEL; PLUCKER, 2014). Na área da saúde, se tem reconhecido as fraquezas existentes acerca do sistema de pesquisas básica e pré-clínica, destacando-se a incapacidade de replicação de grande parte dos trabalhos em pesquisas publicadas em revistas importantes (BEGLEY; IOANNIDIS, 2015). Em relação às pesquisas no campo da econometria, os desafios quanto à análise estatística e sua reprodutibilidade são enormes, com relação à incentivar uma melhor distribuição dos resultados e o enriquecimento dos detalhes disponibilizados dos trabalhos, bem como na utilização de ferramentas mais eficazes (KOENKER; ZEILEIS, 2009).

Uma vez que nenhum processo de pesquisa atual é completamente linear, pois há que se conseguir novos dados e alterar variáveis ou técnicas estatísticas, os documentos dinâmicos das pesquisas reproduzíveis deixam os processos mais fáceis de serem efetuados. Além disto, as pesquisas que se tornam reproduzíveis são mais prováveis de serem utilizadas por outros pesquisadores do que as demais (GANDRUD, 2013). São necessários incrementos na reprodutibilidade dos trabalhos para que se eleve os padrões de julgamento científicos, e uma barreira é a disponibilização dos códigos utilizados no caso da ciência da computação. Há que se cultivar, portanto uma "cultura da reprodutibilidade" (PENG, 2011). A replicação de estudos de referência se torna essencial para o desenvolvimento de políticas baseadas em evidências e práticas científicas melhores, pois eleva a confiabilidade dos ambientes educacionais (MAKEL; PLUCKER, 2014). Além disso, os diversos atores das instituições possuem responsabilidades, sendo que a atualização anual dos principais pesquisadores de projetos experimentais torna-se importante, inclusive para pesquisadores sêniores de instituições conhecidas (BEGLEY; IOANNIDIS, 2015). Nas ciências computacionais e aquelas que utilizam o ferramental quantitativo empírico, a pesquisa é replicável se outros pesquisadores podem seguir os mesmos procedimentos originalmente utilizados, utilizando os mesmos dados e códigos para a análise. É preciso aprender as ferramentas para que a pesquisa seja reprodutível, sendo resultado um processo de pesquisa efetivo, agregando hábitos de trabalho incrementados e uma pesquisa acessível á outros (GANDRUD, 2013).

Na atualidade, dispõem-se de uma variedade de *softwares* estatísticos, proprietários e livres. Os *softwares* proprietários possuem um custo para adquirir a licença para o uso, ao passo que os *softwares* livres R e seu console RStudio facilitam o acesso do público, sendo ferramentas promissoras quanto à abrangência estatística, facilidade na distribuição dos códigos utilizados e na elevação dos parâmetros de reprodutibilidade e replicabilidade dos resultados. O projeto “Software R: capacitação em análise estatística de dados utilizando um software livre” tem por objetivo conciliar o ensino de análises estatísticas e o uso das tecnologias, por meio da linguagem R utilizando o console RStudio. No presente projeto, busca-se desenvolver maiores aprendizagens no uso do *software* aplicado à Estatística Básica, além da integração entre os colaboradores do projeto nas suas diversas áreas. Paralelamente à busca do conhecimento, tem-se a transmissão do mesmo, que neste caso é realizada por meio de aulas envolvendo o público interessado, com o

propósito de abranger alunos e professores de graduação e pós-graduação, bolsistas, comunidade interna e externa em geral.

Metodologia

Realizou-se a criação de material, disponibilização e execução de aulas expositivas, pelos professores, alunos e bolsistas da Universidade Federal da Fronteira Sul (UFFS) *Campus* Cerro Largo, com certificação para os participantes com frequência mínima de 75% e contabilizando 40h de atividades. As oficinas foram realizadas entre os dias 25/05 e 12/06 de 2017, no laboratório de informática do *Campus* Cerro Largo.

Resultados e Discussão

R e RStudio

É preciso efetuar uma primeira distinção acerca de ambos os programas utilizados neste curso, pois o programa R representa linguagem de programação primária para estatísticas e gráficos; e o *software* Rstudio trata-se de um ambiente de desenvolvimento integrado que combina R e outras linguagens de marcação (LaTex, Markdown e Html por exemplo) e pacotes (GANDRUD, 2013). A linguagem R possui uma extensa gama de modelos estatísticos (modelagem linear e não-linear, testes estatísticos clássicos, análise de séries temporais, agrupamento, classificação, etc) bem como técnicas gráficas (GENTLEMAN et al., 1997). Uma característica importante do R é que possui uma comunidade ativa de desenvolvedores que está se expandindo regularmente, abrangendo uma gama de disciplinas para analisar dados e rodar diversas análises estatísticas (GANDRUD, 2013). Dentre as vantagens do R estão: rapidez e livre de custos; traz o estado da arte da estatística; possui gráficos evoluídos e uma comunidade ativa de utilizadores; é excelente para simulação; força o pesquisador a pensar sobre a análise e; possui ligação com bases de dados (GOUVEIA, 2017).

O programa RStudio é um ambiente de desenvolvimento integrado para o R, que em seu console suporta diretamente a execução dos códigos de programação, realiza a plotagem dos dados e gráficos, guarda o histórico, entre outras funções. Em sua versão *open source*, ainda possibilita o rápido acesso as definições das funções, bem como a administração de múltiplos diretórios de trabalho e o acesso a pacotes de desenvolvimento (RSTUDIO, 2017). A linguagem estatística do R juntamente com o console RStudio fornecem o ferramental necessário para a construção de pesquisas reprodutíveis de projetos inteiros, partindo da coleta dos dados, passando pela análise estatística à consequente apresentação dos resultados encontrados.

Embora pareçam semelhantes, a reprodutibilidade e a replicabilidade possui sutis diferenças. A replicação de um estudo objetiva a repetição de forma intencional, em outro contexto, da pesquisa realizada para que se corrobore ou desconfirme os resultados anteriores (MAKEL; PLUCKER, 2014). Já a reprodutibilidade tem como objetivo central repetir os experimentos efetuados inicialmente, utilizando e verificando se foram utilizados os controles legitimados e testes estatísticos apropriados. Faz, portanto com que o experimento científico revele sua validade e por consequência ciência avance (BEGLEY; IOANNIDIS, 2015).

Evento: XVIII Jornada de Extensão

O Projeto Software R

O presente projeto de extensão se caracterizou no primeiro momento pela criação e disponibilização de apostilas sobre o R (https://softwarelivrer.wordpress.com/) em formato digital. Foram realizados estudos e manipulações do *software* utilizado, buscando o desenvolvimento do saber pessoal em relação ao programa adotado. Pesquisaram-se materiais bibliográficos e a disseminação do conhecimento em grupo entre os colaboradores envolvidos, sendo que dentre estes se encontram professores da área estatística, alunos graduandos e alunos mestrandos. Em que, para o desenvolvimento das aulas, foram elaboradas apostilas dinâmicas, utilizando o RMarkdown (recurso de produção de texto disponibilizado pelo próprio software R), a fim de que os participantes interagissem ativamente e desenvolvessem um conhecimento inicial das interfaces do *software* RStudio.

As aulas subdividiram-se nos conteúdos: Módulo 1 - Introdução ao R; Módulo 2 – Estatística Descritiva; Módulo 3 – Inferência Estatística; Módulo 4 – Teste Qui-Quadrado; Módulo 5 – Modelos de Regressão. Visto que o entendimento dos conceitos estatísticos é primordial para a aplicação dos mesmos, fez-se necessário um embasamento inicial, para que se fizesse possível não só a manipulação e obtenção dos resultados, mas também, a compreensão de cada um deles. A administração das aulas se deu alternadamente entre os colaboradores frente a uma turma inicial de 32 alunos. Com a finalidade de transmitir o conhecimento à comunidade acadêmica e regional, divulgou-se a proposta de aulas através de panfletos e através de publicações online em redes sociais e em páginas administradas pela universidade, ressaltando o objetivo de promover àqueles que participassem a capacitação ao uso do software livre R.

Em suma, o objetivo de transmitir os conhecimentos sobre o software foi atendido, mais nitidamente verificado nas atividades encaminhadas aos participantes e retornadas pelos mesmos, sendo realizadas utilizando os comandos do software R. A expectativa é de que os participantes levem conhecimento suficiente para assimilar o aprendizado com os problemas que surgem ao decorrer da vida acadêmica e profissional, levando a um julgamento mais crítico em suas interpretações nas leituras de estudos científicos e relatórios (AGUIAR J. et al, 2016).

Sobretudo, o ganho de conhecimento pessoal é de suma importância, visto que poderá ser aplicado não só no decorrer do trabalho exposto, mas também em trabalhos futuros, os quais possam exigir a manipulação e análise estatística de dados.

Considerações Finais

Em virtude do que foi mencionado, os *softwares* estatísticos, sejam livres ou proprietários são de grande valor frente a realização de análises estatísticas de dados, visto que agregam mais praticidade ao desenvolvimento do trabalho. A disseminação do conhecimento através de trabalhos de extensão entre os alunos, professores e com o envolvimento da comunidade em geral é um dos pilares das Universidades Federais brasileiras, sendo que potencializa a consequente produção de conhecimento das pesquisas realizadas localmente pela UFFS Campus Cerro Largo. Assim, além da constante qualificação do capital social local, este projeto oportuniza a continuidade da utilização e aprendizagem de outras ferramentas livres em projetos de extensão.

Evento: XVIII Jornada de Extensão

Palavras-chave: recursos computacionais; softwares estatísticos; RStudio;

Keywords: computational resources; statistical software; RStudio;

Agradecimentos

À UFFS, pelo auxílio do Programa Instiicional de Bolsas de Extensão.

Referências

AGUIAR, J. et al. **Software R: Capacitação em análise estatística de dados utilizando um software livre**. VI seminário de ensino, pesquisa e extensão (sepe). v. 6, n. 1, 2016.

BEGLEY, C. G.; IOANNIDIS, J. P. A. Reproducibility in science. **Circulation research**, v. 116, n. 1, p. 116–126, 2015.

CORREA, S. **Probabilidade e estatística.** Belo Horizonte: PUC Minas Virtuais, p. 116, 2003.

GANDRUD, C. **Reproducible Research with R and R Studio**. 2. ed. London: CRC The R Series, 2013.

GENTLEMAN, R. et al. **The R project for statistical computing**.

GOUVEIA, L. B. **R: a alternativa ao SPSS e ao NVivo em software livre**.

JELIHOVSCHI, E. **Análise exploratória de dados usando o R**. Ilhéus, BA: Editus, 2014.

KOENKER, R.; ZEILEIS, A. On reproducible econometric research. **Journal of Applied Econometrics**, v. 24, n. 5, p. 833–847, 2009.

MAKEL, M. C.; PLUCKER, J. A. Facts Are More Important Than Novelty: Replication in the Education Sciences. **Educational Researcher**, v. 43, n. 6, p. 304–316, 2014.

PENG, R. D. Reproducible Research in Computing Science. **Science**, v. 334, n. 6060, p. 1226–1227, 2011.

RSTUDIO. **Take control of your R code**.