Elaborando um Compêndio de Pesquisa em RMarkdown

Emerson M. Del Ponte

2017-11-15

Contents

Prefácio

Pesquisa Reproduzível (sin. reprodutível) é um tema da ciencia que tem despertado muito a atenção de pesquisadores, agências de fomento e a mídia acadêmica nos últimos anos. São frequentes os relatos de que um estudo que foi repetido gerou resultados diferentes ou mesmo discortantes de um estudo anterior. Os próprios pesquisadores tem se manifestado com grande preocupação com uma alegada "crise de reprodutibilidade" na ciência.

As possíveis causas e algumas soluções para minimizar o problema vem sendo discutidas e algumas ações implementadas. É importante que o estudante e pesquisador em geral tenha conhecimento sobre a correta definição dos termos para melhorar a comunicação. Reprodutibilidade tem diferentes significados dependendo do contexto. Aqui, definimos como:

> Capacidade de um pesquisador em chegar ao resultado de um estudo prévio usando os mesmos materiais (dados) e métodos (estatística) da pesquisa original.

Portanto, é condição *sine qua non* que o pesquisador independente tenha:

- acesso aos dados
- saiba detalhes de como a análise foi feita

Na mídia, os resultados discordantes de estudos se refere, na verdade, à **replicabilidade**, ou a **reprodutibilidade inferencial**, segundo alguns autores. Em ciência, parte-se do princípio de que os resultados de uma pesquisa publicados em revistas com corpo editorial tenham sido obtidos segundo os princípios que regem os métodos e a ética científica. No entanto, os editores e revisores, quase sempre, não tem como verificar se todos os passos do trabalho, especialmente a análise dos dados, foram executados corretamente, uma vez que avaliam apenas um produto - o artigo científico.

Um artigo científico é escrito e submetido para publicação segundo convenções da academia que definem o conteúdo mínimo para que o trabalho seja avaliado pelos pares. Normalmente esse conteúdo consiste no texto, gráficos e tabelas e, idealmente, um material suplementar (o que ainda parece ser desconhecido para muitos pesquisadores).

Ambiente computacional

Um cientista que objetiva que sua pesquisa seja reproduzida por outros grupos deve se preocupar em disponibilizar também os dados e os procedimentos de análise para que possam ser inspecionados e reutilizados. Esse trabalho também exige procedimentos padronizados de conteúdo e formato de arquivos (dados, relatórios, imagens, etc.) os quais devem ser comentados de maneira clara para que outros pesquisadores possam reproduzir, melhorar ou expandir uma análise estatística, reutilizar os dados em outro trabalho ou combinar com outros conjuntos. Uma condição para que isso ocorra é que os pesquisadores usem um mesmo ambiente computacional que seja preferenciamente de código aberto. Dentre os ambientes, nossa preferência é pelo R e conjunto de ferramentas integradas no ambiente RStudio.

Nesse livro, o estudante ou pesquisador interessado aprenderá as boas práticas de pesquisa reproduzível que devem ser incorporadas na rotina de trabalho. Com hábitos simples, persistência, sistemática e um pouco de

dedicação na linguagem R, o pesquisador moderno estará adotando práticas que contribuirão para a maior transparência na ciência.

Slides do autor

A primeira versão deste material foi elaborada para um minicurso de 8h oferecido aos estudantes do Programa de Pós-graduação em Fitopatologia dia 13 de novembro de 2017. Os slides da apresentação podem ser visualizado abaixo.

Agradecimentos

Chapter 1

Requisitos básicos

1.1 Uma nova rotina

Para um estudante ou cientista que está iniciando um projeto é importante que as boas práticas de PR sejam incorporadas no seu dia a dia, e que sejam implementadas desde a concepção e o planejamento do mesmo.

São atividades que dependem essencialmente de uma grande capacidade organizacional e admistrativa de tempo e esforço no planejamento, condução e documentação de tudo que é feito. É preciso seguir rotinas e gerar documentos que seguem certas normas de padronização, especialmente se o trabalho é feito de forma colaborativa. Analogamente, é como escrever e formatar um artigo científico que deve ser estruturado e apresentado segundo determinadas normas. Aqui, o produto gerado não é somente o documento do manuscrito e um punhado de gráficos, mas sim tudo que foi gerado durante a pesquisa e que precisa estar bem organizado e formatado para uso posterior e publicação/divulgação. Para obter sucesso na implementação de uma PR, é preciso:

- Ser diligente e sistemático
- Aprender novas ferramentas (computacionais)
- Aprender a organizar arquivos diversos
- Documentar todas as etapas do trabalho

No dia a dia, os pesquisadores não sobrevivem se os computadores como ferramenta central de trabalho. Atualmente, não é preciso ser um "nerd" para que se possa utilizar com bastante eficiência os computadores que estão hoje cada vez mais portáteis e de fácil uso, para ser eficiente e produtivo no trabalho. Em algumas áreas da pesquisa é necessário maior envolvimento com linguagens de programação, programas específicos que exigem um esforço de aprendizado.

No entanto, na PR o mais importante e desafior é certamente aprender a sistemática de trabalho do que ser um expert em programação - mas é necessário sim aprender alguma linguagem de programação (R ou Python) para implementar as práticas de PR. Durante nossa formação acadêmica não recebemos nenhum ou muito pouco treinamento em como preparar e organizar de maneira apropriadas os arquivos diversos incluindo dados, códigos, gráficos, tabelas, manuscrito, figuras, etc.

Apender uma rotina de PR é fundamental para:

1) Facilitar o nosso próprio trabalho de análise-reanálise
2) Permitir o uso dos dados e códigos por outras pessoas (seu orientador!)
3) Compartilhar a "pipeline" da análise, ou seja, explicar o que, por que e como foi feito

Quando não somos treinados a trabalhar seguindo as boas práticas de PR, é muito comum: criar um número grande arquivos e versões desnecessárias que dificultam o processo; gerar inconsistência e redundância nas análises; não ter um controle adequado de versões e dificuldade quando é solicitado o compartilhamento do

trabalhos - ou seja, levará um tempo grande só para organizar a "bagunça" que foi gerada durante o processo e que só o próprio pesquisador entende, quando entende! Práticas que deveriam ser simples como refazer um gráfico ou estatísticas após receber os pareceres de revisores se tornam um verdadeiro pesadelo para alguns pesquisadores, o que contribui para o atraso na publicação de artigos.

1.2 Ferramentas computacionais

Segundo Yihui Xie, um dos principais desenvolvedores do R da empresa RStudio de programas (ex. knitr, rMarkdown, bookdown, etc) que visam facilitar a pesquisa reproduzível:

> The final product of research is not only the paper itself, but also the full computation environment used to produce the results in the paper such as the code and data necessary for reproduction of the results and building upon the research (Xie et al. 2014).

Dentre os ambientes de programação disponíveis, as ferramentas mais usadas para implementar uma PR de maneira efetiva (dados, análises e saídas são combinados, idealmente, em um único ambiente de programação), são baseados em duas linguagens principais: Python e R, cujos produtos principais são Jupyter Notebooks e RMarkdown, respectivamente. Esses pacotes ou rotinas facilitam sobremaneira a documentação e reprodução das análises bem como aceleram a obtenção dos resultados e visualizações assim que novos dados forem adicionados ou reanálises são necessárias.

Além de aprender a utilizar esses programas, é importante que o pesquisador aprenda como usar efetivamente planilhas eletrônicas para reunir e organizar os dados que serão usados na pesquisa. Por princípios, as planilhas eletrônicas como Excel, Libre Office Calc, Numbers e Google Sheets são usadas apenas para armazenar os dados e não para processar, transformar, visualizar ou fazer sumários prévios. O motivo é muito simples: esses procedimentos todos feitos com movimentos de mouse não são reproduzíveis! além disso, na PR os dados originais levantados ou recebidos devem ser mantidos na sua forma original. Caso seja modificado de forma que é mais fácil fazer em uma planilha como renomear variáveis, é importante manter sempre uma planilha não manipulada como referência.

Chapter 2

Compêdio de pesquisa

2.1 Definição

Um compêndio (do inglês, compendium) pode ser definido como uma compilação em que se encontra resumido o mais indispensável de um estudo. O termo "research compendium" foi cunhado em meados dos anos 2000 em uma publicação sobre pesquisa reproduzível com enfoque em análise estatística.

> We introduce the concept of a compendium as both a container for the different elements that make up the document and its computations (i.e. text, code, data, …), and as a means for distributing, managing and updating the collection.(Robert Gentleman, Department of Biostatistics, Harvard University et al., 2004)

Na visão dos autores, o compêndio tem como base um ou mais documentos dinâmicos, a partir dos quais são gerados os demais documentos estáticos como um PDF a partir de um arquivo TEX (há mais de 10 anos o TEX era o equivalente ao Markdown). Também, esses documentos dinâmicos (e seus elementos de texto, código e dados) devem ser facilmente extraídos e processados de diferentes maneiras pelo autor e pelos leitores.

Portanto, é importante que um compêndio da pesquisa seja desenvolvido de maneira padronizada para que outros pesquisadores, de forma facilitada, possam inspecionar, reproduzir e ampliar a pesquisa. Os princípios básicos para a confecção de um compêndio de pesquisa são:

1) Organização segundo a convenção criada pela academia
2) Separação clara de dados, códigos e saídas
3) Especificação do ambiente computacional
4) Documentação detalhada de cada elemento e rotina

Um dos objetivo de se criar um compêndio de pesquisa é, em primeira instância, facilitar o trabalho do pesquisador e permitir que ele distribua os produtos gerados pela pesquisa de uma maneira mais ampla possível, visando a dar visibilidade ao trabalho. Além disso, a eficiência do trabalho é aumentada com a incorporação de uma sistemática que pode ser replicada em outros projetos, dimuindo assim os custos operacionais e acelerando o trabalho. Tem se discutido que publicações que são acompanhas de um compêndio tendem a receber maior atenção, credibilidade e citações uma vez que o compendio é um tipo de publicação por si só. Imagine o quanto pode ser útil tornar o compêndio público antes da submissão do trabalho de forma que editores e revisores possam revisar os dados e os métodos (estatísticos principalmente) que foram utilizados, entender as decisões tomadas no processo analítico e os resultados obticos. Alem disso, uma vez que o compêndio é publicado, os autores poderão receber comentários dos pares e dos revisores para melhorar o trabalho como um todo.

Antes de falarmos sobre como organizar o compêndio de pesquisa, veremos aspectos e cuidados específicos de organização dos dados.

2.2 Dados

Segundo Wilkinson et al. (2016), os dados devem ser organizados segundo o princípio **FAIR** = **F**indable, **A**ccessible, **I**nteroperable and **R**eusable.

Compartilhar os dados (o que pode ser aplicado também aos códigos, significa facilitar a distribuição e o acesso pela comunidade científica, ou seja que eles sejam facilmente **encontrados** e **acessados**. Qual a vantagem disso? reproduzir os resultados originais e pertimitir que novas análises sejam feitas usando os mesmos dados, ou mesmo combinando com outros conjuntos de dados (metanálise). É importante que os dados estejam em um formato que seja de fácil entendimento para facilitar o uso. Três boas práticas são recomendadas:

1. Documentação: dados bem documentados e descritos (metadados) são mais fáceis de entender
2. Formataçao: dados formatados apropriadamente podem ser usados em diversos programas de computador
3. Distribuição: depósito em repositórios conhecidos e com licença aberta facilita que sejam encontrados e reusados

Práticas de compartilhamento de dados talvez ainda não sejam valorizadas pela maioria dos pesquisadores, haja visto que a maioria dos trabalhos ainda não disponibilizam os dados originais. As vantagens óbvias seria a reprodução e possível melhorias na análise, o reuso dos dados em metanálises para chegar a conclusões gerais e gerar novo conhecimento que só é possível com dados compartilhados em larga escala. Mas por que ainda os pesquisadores não compartilham os dados? os motivos podem ser ligados ao receio de perder uma competição por publicações ou mesmo a falta de conhecimento sobre como fazer o compartilhamento. A percepção que domina é que organizar e depositar dados é difícil tecnicamente ou leva muito tempo,

2.3 Formato

2.4 Planilhas eletronicas

2.5 Nomes

2.6 Copias

2.7 Formato

2.8 Depósito

2.9 Licença

Bibliography

Robert Gentleman, Department of Biostatistics, Harvard University, Duncan Temple Lang, Department of Statistics, University of California, Davis, and Authors (2004). Statistical analyses and reproducible research.