Um mapa para a transparência e replicabilidade na ciência social empírica: o Protocolo TIER

Amanda Domingos (Universidade Federal de Pernambuco)

Ian Rebouças Batista (Universidade Federal de Pernambuco)

Resumo: Transparência e replicabilidade são o padrão ouro da pesquisa científica. Mas, alcançar esse padrão pode ser um caminho longo e trabalhoso. Nesse artigo, apresentamos uma maneira de aumentar a transparência de sua pesquisa através da documentação. Argumentamos que o Protocolo TIER é uma importante ferramenta de transparência e de auxílio na organização de materiais para replicação. Trata-se de uma metodologia de documentação que compila os principais materiais produzidos numa pesquisa empírica social. Além disso, destacamos a adequação do protocolo nas três dimensões do padrão de replicação (substantiva, pedagógica e transparente), uma vez que contribui para o avanço da ciência, introduz jovens graduandas à análise de dados de uma forma prática e permite a transparência quanto aos procedimentos empregados na pesquisa. Por fim, discutimos atividades relacionadas ao Projeto TIER que contribuem com os objetivos do protocolo. Especificamente, apresentamos um exercício de introdução à análise de dados a ser realizado no R, onde a documentação do protocolo é requerida e sua aplicação, portanto, ilustrada.

Palavras-chave: Transparência; Replicabilidade; Documentação; Protocolo TIER

A map to transparency and reproducibility in empirical social science: TIER protocol

Abstract: Transparency and replicability are the golden standard in empirical research. But reaching that standard can be a long and laborious journey. In this article, we present a manner to increase the transparency of your research through documentation. We argue that TIER Protocol is an important tool for transparency and assistance in organizing replication materials. It is a documentation methodology that compiles the main materials produced in an empirical social research. In addition, we highlight the adequacy of the protocol in the three dimensions of the replication standard (substantive, pedagogical and transparent), since it contributes to the advancement of science, introduces young undergraduate students to data analysis in a practical way and allows transparency of the research procedures. Finally, we discuss activities related to the TIER Project that contribute to the objectives of the protocol. Specifically, we present a data analysis introduction exercise for R, where the protocol documentation is required and its application, therefore, exemplified.

Key-words: Transparency; Reproducibility; Documentation; TIER Protocol

Introdução

Baixar o banco de dados, tratar o banco, elaborar *scripts*, programar funções, executar comandos, analisar os resultados. Se essas são etapas corriqueiras de sua pesquisa acadêmica, existe um protocolo que irá te ajudar a organizar a sua pesquisa e a divulgar, de maneira transparente e replicável, seus resultados.

Princípios e práticas de transparência e replicabilidade na pesquisa científica têm mobilizado crescente atenção da comunidade acadêmica nas ciências sociais, ganhando destaque em publicações nacionais (PARANHOS, 2012; PARANHOS et al, 2014; FIGUEIREDO et al, 2019; AVELINO e DESPOSATO, 2018) e internacionais (CHRISTENSEN, 2016; MUNAFÓ et al, 2017; CHRISTENSEN e MIGUEL, 2018; STOCKEMER, KOEHLER E LENZ, 2018; DE LA GUARDIA e STUDY, 2019), sendo ainda incentivados através de iniciativas inter-universitárias como *Berkeley Initiative for Transparency in the Social Sciences* (BITSS) e *Teaching Integrity in Empirical Research* (TIER), que promovem cursos, *workshops* e eventos voltados à troca de experiências e ideias em prol da transparência.

Identificada como padrão-ouro (KING, 1995), a transparência quanto aos procedimentos adotados em uma pesquisa é aqui entendida como a disponibilização dos materiais utilizados e a indicação das etapas percorridas até se chegar aos resultados. Para além dos benefícios que isso traz ao adequado funcionamento da ciência, permitindo acesso público a suas escolhas metodológicas e possibilitando que seu trabalho seja replicado e avaliado mais rigorosamente, a transparência permite: (i) teste de hipóteses mais preciso, (ii) a colaboração com outros pesquisadores com os mesmos interesses e (iii) aumento na credibilidade da autora. Argumentamos que a pesquisadora transparente, que documenta seu trabalho à medida que é feito, tem maior facilidade na elaboração de sua pesquisa. Uma documentação maior traz benefícios em termos de reputação científica, uma vez que contribui para boas práticas da ciência e gera benefícios práticos, como o aumento de citações (FIGUEIREDO et al, 2019).

A partir disso, alguns instrumentos têm sido empreendidos para difundir e implementar os princípios e as práticas da transparência e replicabilidade nas Ciências Sociais, sendo as três práticas principais: planos de divulgação, pré-registro e abertura de dados e materiais (MIGUEL et al, 2014). O objeto desse estudo se enquadro no terceiro tipo. Nesse artigo, apresentamos e sugerimos a adoção do protocolo TIER, uma metodologia para documentação adequada dos arquivos utilizados no desenvolvimento da pesquisa (bancos de dados, *scripts*, *outputs*) e posterior disponibilização desses materiais para replicação. Acreditamos que a adoção de uma regra específica quanto à organização dos arquivos de pesquisa aumenta a transparência e favorece a replicabilidade do estudo.

O artigo se divide como segue. Na próxima seção, apresentamos o padrão-ouro da ciência em termos de transparência dos procedimentos e replicabilidade dos resultados, indicando porque a leitora deve almejar esses valores e práticas. Na sequência, apresentamos o protocolo TIER, destacando os seus aspectos fundamentais. Em seguida, enfatizamos os benefícios e incentivos a se adotar uma estratégia de documentação e disponibilização para tornar a pesquisa transparente. Por fim, elencamos algumas atividades desenvolvidas pelo Projeto TIER que corroboram e buscam difundir as boas práticas de transparência e replicabilidade nas ciências sociais e apresentamos um exercício introdutório ao protocolo. As considerações finais encerram este trabalho.

Por que um Protocolo para transparência?

Dentre as características de uma boa pesquisa científica estão a inovação conceitual, o rigor metodológico e a riqueza do conteúdo substantivo (PRZEWORSKI e SALOMON, 1998). Nas

Ciências Sociais, no entanto, outros dois elementos foram adicionados a essas características nas últimas décadas: a transparência e a replicabilidade, tidos como o padrão-ouro da pesquisa empírica (KING, 1995; PARANHOS et al, 2014; JANZ, 2016). Os benefícios de uma pesquisa transparente são diversos, tanto para a disciplina como um todo quanto para a pesquisadora, em particular. A adoção de práticas transparentes permite que a roda da ciência gire de forma mais rápida, ao permitir o teste de hipótese mais preciso, facilitar a colaboração com outros pesquisadores com os mesmos interesses, aumentar a credibilidade da pesquisa, além de retirar a costumeira atenção da significância estatística dos resultados e aumentar a relevância da pesquisa *per se* (DE LA GUARDIA e STUDY, 2019).

No que se refere a benefícios para autores e periódicos, Figueiredo et al (2019) apresentam sete razões para adotar procedimentos transparentes em pesquisas empíricas nas ciências sociais: (I) prevenir que as pesquisas contenham erros honestos ou, até mesmo, fraudes em suas análises empíricas; (II) tornar a escrita de artigos científicos mais fácil, uma vez que permite o teste de outras hipóteses com a mesma base de dados; (III) procedimentos transparentes permitem que pareceristas tenham acesso a dados e *scripts* e deem recomendações mais precisas sobre como melhorar o trabalho; (IV) ao garantir a sistematização de dados brutos e analisados, permite que o trabalho cientifico do autor continue; (V) ajudam a construir reputação científica, ao demonstrar a boa-fé por apresentar material de replicação; (VI) ajudam a aprender análise de dados, ao permitir que os alunos tenham contato com dados e procedimentos estatísticos utilizados na vida real e (VII) aumenta o impacto do trabalho acadêmico, uma vez que evidências empíricas demonstram que trabalhos que adotam procedimentos e práticas transparentes são duas vezes mais citados.

Por outro lado, os custos da não adoção de procedimentos de transparência e replicabilidade também são variados. Pesquisas que não mantém procedimentos públicos podem contribuir para a diminuição da confiança em seus resultados, e, além disso, podem ser uma barreira à publicação em periódicos de alto fator de impacto que solicitam matérias de replicação – o que diminui o alcance da pesquisa (MUNAFO et al, 2017; DE LA GUARDIA e STUDY, 2019). Diante de tudo isso, uma pessoa poderia argumentar que não adota princípios e práticas de transparência por ser muito trabalhoso. Aqui, vale a pena reforçar que o padrão-ouro de replicação não solicita que um indivíduo reproduza os resultados apresentados em livros ou artigos científicos, apenas recomenda que sejam disponibilizadas tanta informação quanto for suficiente para que, caso algum pesquisador deseje, consiga reproduzi-los (KING, 1995).

Apesar disso, a prática de transparência não é a regra. Cerca de 67% dos artigos publicados na *American Political Science Review* entre 2013 e 2014 não são replicáveis (KEY, 2016). No Brasil, apenas 5% dos artigos publicados entre 2012 e 2016 em periódicos da área de Ciências Sociais eram passíveis de replicação completa (AVELINO e DESPOSATO, 2018)[1,2]. O padrão ainda é a obscuridade, não a transparência. Nesse sentido, não é incomum encontrar pesquisadores que não tornam públicos os procedimentos e práticas que os levaram às conclusões de seus produtos científicos.

Questionamos a cara leitora: quantas vezes, na produção de sua própria pesquisa empírica, se esforçou para lembrar os caminhos que a levaram ao resultado que encontrou? Mais do que isso, o que você tem feito para ser transparente? Se pensou um tanto antes de responder essa última pergunta, esse artigo tem uma contribuição a te fazer. Nas próximas seções, vamos sugerir um

---

1 O relatório é fruto de do projeto "*Research Transparency in Brazilian Political and Social Science: A First Look*" dos autores junto à *Berkeley Initiative for Transparency in the Social Sciences* (BITSS). Para maiores informações, visite: < https://www.bitss.org/projects/research-transparency-in-brazilian-political-and-social-science-a-first-look/>. Acesso em: 20 mar. 2010.

2 O relatório final está publicamente disponível em: <https://osf.io/f279z/>. Acesso em 20 mar. 2020.

protocolo de organização para facilitar a implementação de práticas transparentes e discutir um pouco sobre o processo de documentação de uma pesquisa empírica.

O protocolo TIER

O protocolo TIER é uma metodologia de documentação dos materiais de replicação de uma pesquisa empírica. De modo simplificado, o protocolo traz uma indicação geral sobre a organização dos arquivos utilizados em um estudo, indica a construção de pastas e quais informações devem ser disponibilizadas em cada uma delas. O protocolo foi desenvolvido originalmente para lidar com um problema de sala de aula. Os trabalhos que eram entregues pelos alunos de Introdução à Estatística, disciplina ministrada por Richard Ball na graduação de economia na Haverford College (EUA), apresentavam problema similar àqueles existentes nos artigos científicos publicados nos periódicos acadêmicos de grandes revistas: eles não eram transparentes. Ao entregarem os trabalhos finais da disciplina com resultados errados, era praticamente impossível para o professor entender em que momento da atividade de análise dos dados foram realizados procedimentos equivocados, o que impedia a adequada correção do professor e, consequentemente, o processo de aprendizado do aluno.

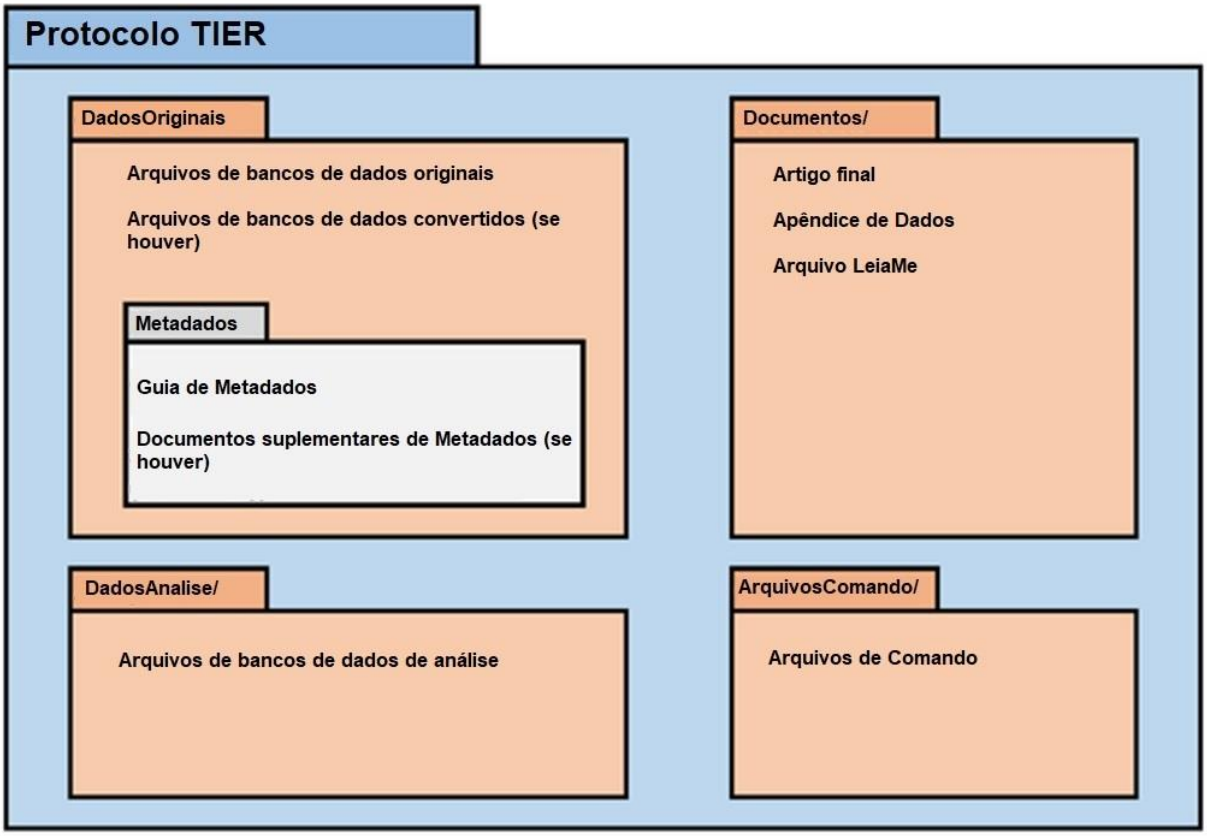
Para garantir que existisse um padrão para a entrega das atividades, tornando possível que o professor corrigisse de forma eficiente, tendo acesso aos *scripts* e dados de forma inteligível e organizada, Ball elaborou, junto com o bibliotecário do College, Norm Medeiros, o protocolo TIER (BALL e MEDEIROS, 2012). O protocolo consiste em uma organização e hierarquia de pastas específica que separa as informações sobre determinada pesquisa (dados brutos, dados analisados, *scripts*, etc.) e facilita a replicação dos achados. Apesar dessa origem pedagógica, o Protocolo TIER vai além das portas da sala de aula. Diversos pesquisadores na área de Ciências Sociais podem utilizar o protocolo como uma maneira de aumentar o nível de transparência de suas pesquisas. Nesse sentido, argumentamos que a adoção do Protocolo TIER acaba por colaborar nas três dimensões do padrão de replicação: substantiva, pedagógica e transparência (PARANHOS et al, 2013).

O protocolo TIER possui uma orientação geral, tal como se encontra ilustrado na Figura 1, e possui versões que aprimoram a organização e a operacionalização da replicação de seu trabalho, como a versão 4.0 apresentada na Figura 2. Discutiremos as duas versões na sequência, mas antes alguns esclarecimentos procedimentais dessa metodologia organizacional que o TIER propõe ajudam a entender sua proposta.

As diferentes pastas que irão compor o material de replicação de seu trabalho devem ser criadas no início da pesquisa. Elas acompanharão a pesquisadora ao longo de todo o processo: tratamento do banco de dados original, processamento das análises, elaboração do artigo e dos apêndices. A pesquisadora deve ir salvando e trabalhando com os arquivos nas pastas adequadas e referenciando, nos seus *scripts* ou arquivos de comando, os diretórios de trabalho de acordo com a hierarquia de pastas, o que torna fundamental que essa organização inicial seja mantida ao longo do trabalho. Por exemplo, no *script* onde você irá tratar o banco de dados original e irá gerar o banco da análise, você deve carregar o banco de dados original a partir da pasta *DadosOriginais* e deve salvar o banco tratado na pasta *DadosAnalise* (Figura 1). A ideia é que usando um *software* como o R, seu *script* contenha as linhas necessárias para abrir o banco original, que estará na pasta adequada, e salvar o banco de análise na pasta correta[3].

3 Como uma medida de replicabilidade, o ideal é que você adote um diretório de trabalho relativo. Para maiores informações, ver: https://ssc.wisc.edu/sscc/pubs/R_intro/book/1-10-paths-and-working-directories.html. Acesso em 07 de outubro de 2020.

Figura 1: Visão geral da documentação sugerida pelo Protocolo TIER



Fonte: Project TIER. Tradução dos autores[4]

Mas que organização de pastas você deve adotar, uma vez que as pastas propostas na Figura 1 e na Figura 2 são distintas? O objetivo do protocolo é fornecer uma orientação para organização e disponibilização de todo o material digital utilizado na elaboração de uma pesquisa empírica. É importante que essa orientação seja coerente e que seja, se não intuitiva, justificada e apresentada. Sendo assim, a exata configuração das pastas e dos arquivos fica a gosto do freguês. A Figura 1 é o exemplo mais genérico, a partir do qual se elaboraram quatro outras versões, e a Figura 2 é a formatação sugerida mais recente pelo Projeto (4.0). Contudo, se o argumento que o protocolo faz, sobre organização da pesquisa e sobre transparência, convencer a pesquisadora/leitora, você é encorajada a organizar o seu material para disponibilização a partir de um protocolo que funcione para você, ou seja, na sua própria hierarquia de pastas que achar conveniente.

Para que os protocolos individuais funcionem a qualquer pesquisador interessado em replicar o seu trabalho e conferir os resultados, são necessárias algumas orientações básicas que compõem a mensagem do TIER. O primeiro deles é a elaboração de arquivos LeiaMe (geralmente em .txt, para facilitar a leitura independente do sistema operacional utilizado) que orientará a atividade de replicação desde o início. Uma das premissas de uma replicação transparente é que arquivos desse tipo são os primeiros a serem lidos, pois além de apresentar e explicar a

[4] O *template* utilizado é o *template* original utilizado pelo Projeto TIER. Disponível em: https://www.projecttier.org/tier-protocol/specifications/#overview-of-the-documentation. Acesso em 23 de março de 2020

configuração de pastas adotada, que a replicadora estará se deparando ao ler, são esses documentos de texto que dizem que arquivos abrir primeiro e qual será a sequência do trabalho de replicação. É interessante, portanto, que ele apareça logo na primeira pasta, como consta na Figura 2. A elaboração de um arquivo LeiaMe claro e objetivo acaba demonstrando a preocupação da pesquisadora com a replicação futura de seus achados. Assim, deve fornecer um panorama completo da maneira com que as pastas estão organizas e das etapas de replicação.

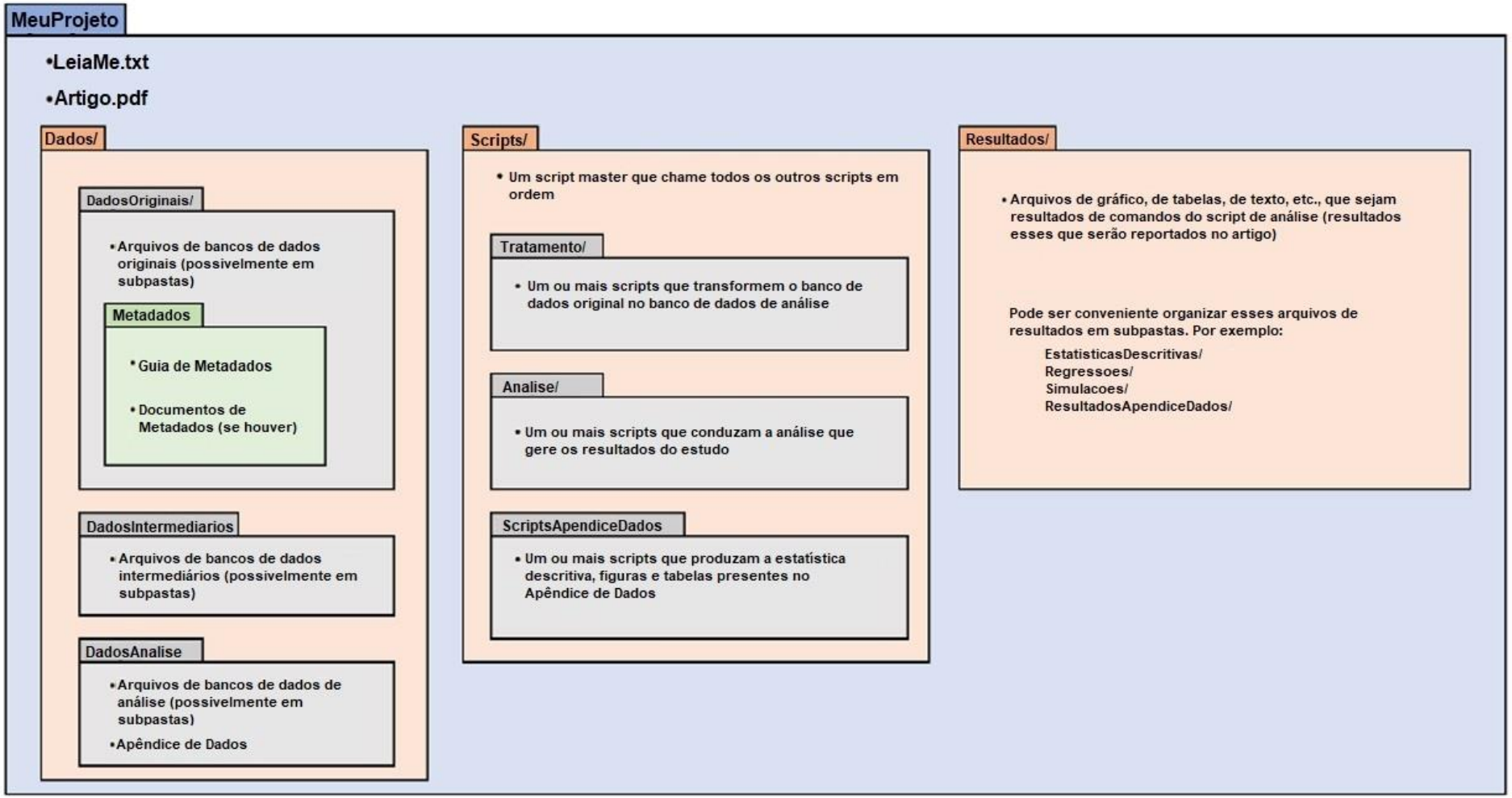
Por exemplo, para um trabalho que utiliza o protocolo inspirado no apresentado na Figura 2, o arquivo LeiaMe pode conter um texto semelhante a esse: "Inicialmente, vá até a subpasta *Tratamento*, localizada na pasta *Scripts*, e abra o *script* Tratamento.R. Nele você irá tratar o Banco.dta, que está na subpasta *DadosOriginais*, dentro da pasta *Dados*, e irá gerar o BancoNovo.dta, que será salvo na subpasta *DadosAnalise*. Em seguida, também dentro da pasta *Scripts*, vá até a subpasta *Analise*, abra o *script* Analise.R. Nele você irá gerar os Gráficos 1, 2 e 3, que serão salvos na pasta *Resultados"*. Os arquivos LeiaMe podem ainda conter informações sobre o *software* de análise de dados utilizado, os endereços da *web* onde os bancos originais foram baixados ou ainda informações para contato com a pesquisadora.

Semelhante ao LeiaMe, mas com outra função, o protocolo TIER considera ainda que é fundamental que a pesquisadora elabore um Apêndice de Dados (ou *codebook*). É nesse documento que serão descritas com detalhes as variáveis que constam no banco de dados de análise, incluindo o tipo dessa variável, como ela é mensurada e como ela foi construída (se já estava pronta no banco de dados original ou se foi elaborada pela pesquisadora no processo de tratamento dos dados). Pode-se ainda incluir estatísticas descritivas para cada uma das variáveis. É nesse documento também que maiores especificidades sobre as versões do *software* utilizado e dos pacotes estatísticos necessários são fornecidas.

Outra orientação básica que permeia o protocolo TIER em qualquer de suas versões, e que deve estar presente também num possível protocolo pessoal inspirado no TIER, é a separação dos arquivos por tipos. Tanto na Figura 1 quanto na Figura 2 existem pastas próprias para bancos de dados, *scripts* e documentos. Analisemos de modo mais aprofundado essas figuras agora.

Na Figura 1, a orientação geral do protocolo, tem-se quatro pastas ao todo: *DadosOriginais*; *DadosAnalise*; *AquivosComando*; *Documentos*. Dentro da pasta *DadosOriginais*, encontramos os arquivos de bancos de dados originais, baixados diretamente da fonte que for, os bancos de dados convertidos, caso a pesquisadora tenha que converter o formato do banco de dados original baixado (note que mesmo nesse caso é importante manter o banco também no formato original), e uma subpasta chamada de *Metadados*, onde se encontram um Guia dos Metadados (um Apêndice de Dados para o banco original, incluindo a citação bibliográfica indicada, o identificador DOI, a data em que o banco foi baixado e o endereço de web) e documentos suplementares desse banco original (se houver). Na pasta *DadosAnálise*, tem-se o(s) banco(s) de dado(s) tratado(s), com o qual a pesquisadora fará as análises do seu artigo. Na pasta de *Documentos*, encontram-se o artigo final, o Apêndice de Dados e o arquivo LeiaMe. Por fim, na pasta *ArquivosComando*, estarão todos os *scripts* utilizados na pesquisa, desde o tratamento do banco de dados original até as análises realizadas.

Figura 2: Protocolo TIER na sua versão 4.0 (2020)



Fonte: Project TIER. Tradução dos autores

A Figura 2, por sua vez, baseia-se em uma elaboração preliminar da versão 4.0 do Protocolo, apresentada em novembro de 2019. Nela, na pasta principal, tem-se logo de cara dois arquivos, o artigo final e o arquivo LeiaMe, que irá orientar o procedimento daquele que deseja replicar o presente trabalho, e três pastas: *Dados*, *Scripts* e *Resultados*. Na pasta *Dados*, encontram-se três subpastas: *DadosOriginais* (com a pasta de *Metadados* dentro dela); *DadosIntermediarios*, onde estarão bancos de dados por ventura convertidos em outro formato; e *DadosAnalise*, onde também estará o Apêndice de Dados. Na pasta *Scripts*, encontra-se um *script* "*master*", que combinará todos os *scripts* do trabalho, mas também três subpastas: *Tratamento*, *Analise* e *ScriptApendiceDados*, cada pasta contendo o *script* que realiza cada uma das etapas referentes. Por fim, na pasta *Resultados*, estarão todos os arquivos que foram gerados ao longo da análise a partir dos *scripts*, como tabelas, gráficos e figuras, que podem, por sua vez, estar subdivididas também em subpastas.

Como se percebe, as duas versões do protocolo compartilham do mesmo ideal de documentação e transparência da pesquisa realizada, ainda que adotando organizações e hierarquias de pastas diferentes. Compreendendo quais os objetivos do protocolo e a lógica da separação dos arquivos presentes em ambas as versões, é possível então que a pesquisadora interessada numa pesquisa transparente e replicável possa, se não adotar exatamente alguma das versões expostas, adotar uma versão própria do protocolo TIER. Uma versão que faça sentido para a pesquisadora e beneficie o seu fluxo de trabalho, ao mesmo tempo que seja clara e acessível a qualquer interessado em replicar os resultados de sua pesquisa.

Após elaborado todo o material seguindo o Protocolo TIER, a disponibilização deve ocorrer da maneira mais acessível possível. Como a transparência no uso dos dados é um assunto em ascensão, temos a nosso dispor uma gama de repositórios digitais que fornecem a estrutura adequada para essa disponibilização. O *Open Science Framework*[5] (OSF), o *Dataverse*[6] e o *Github*[7] são três exemplos de repositórios adequados para o compartilhamento de todo o material de seu trabalho, inclusive com a possibilidade de divisão nas próprias pastas do protocolo. Nesse artigo, consideramos que o trabalho da pesquisadora tenha uma abordagem quantitativa, dada a natureza do protocolo que beneficia trabalhos deste tipo. Mas, também existem repositórios digitais para trabalhos com uma abordagem qualitativa[8], a saber: *Qualitative Data Repository*[9], UK data archive[10], National Anthropological Archives[11].

Documentação, protocolo TIER e dimensões do padrão de replicação

Em 1995, ano de lançamento do dossiê "*verification/replication*"[12], a realidade da produção científica nas ciências sociais era diferente. O dossiê propôs aos pesquisadores da área tentar tornar disponíveis maiores detalhes sobre as escolhas empíricas realizadas em suas pesquisas em notas de rodapé e apêndices metodológicos, mas esses esforços esbarravam nos limites impostos pelos periódicos e editoras de livro para divulgar tais detalhes. Por outro lado, a disponibilização de informações e armazenamento eram um problema: as facilidades de *drives* e

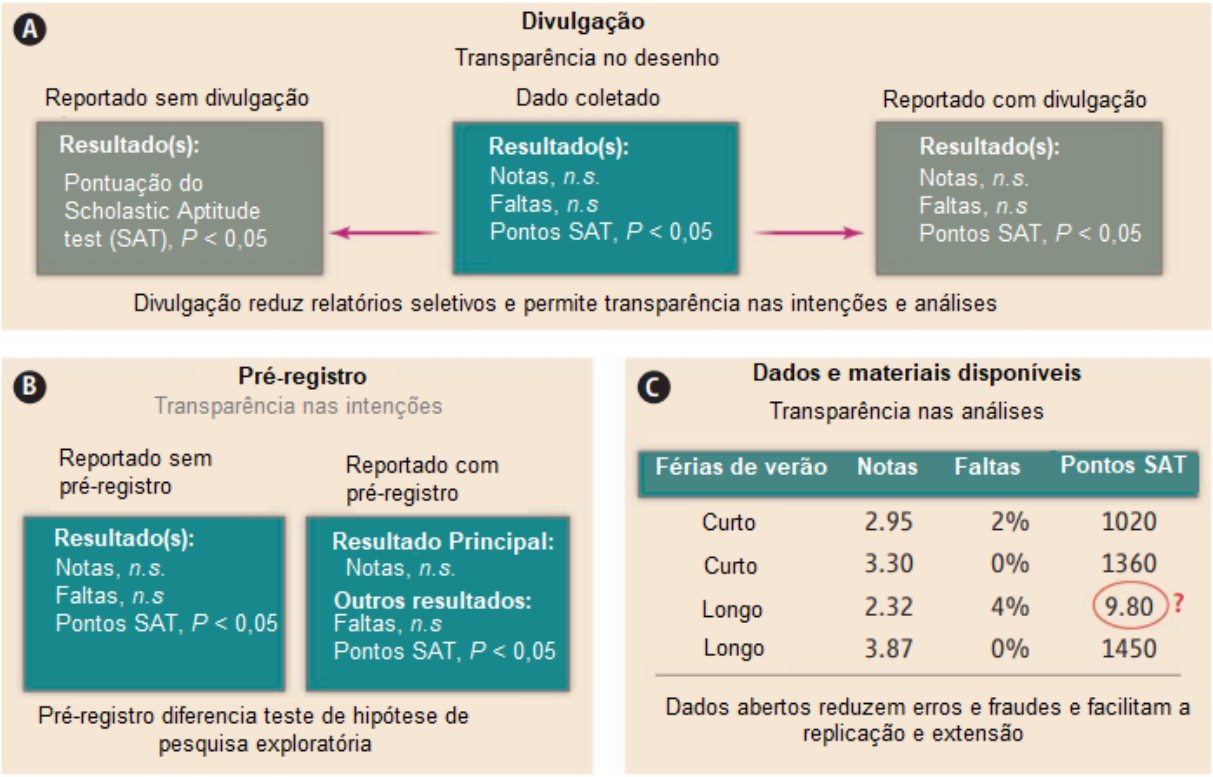
---

5 Ver: https://osf.io . Acesso em 20 mar. 2020.
6 Ver: https://dataverse.org/ . Acesso em 20 mar. 2020.
7 Ver: https://github.com/ . Acesso em 20 mar. 2020.
8 Para maiores informações sobre repositórios que aceitam dados qualitativos, ver: https://www.sesync.org/list-of-qualitative-data-repositories. Acesso em 24. Mar. 2020.
9 Ver: https://qdr.syr.edu/. Acesso em 24 mar. 2020.
10 Ver: https://www.ukdataservice.ac.uk/manage-data. Acesso em 24 mar. 2020.
11 Ver: http://anthropology.si.edu/naa/search.html. Acesso em 24. mar. 2020.
12O dossiê está disponível em: < https://www.cambridge.org/core/journals/ps-political-science-and-politics/issue/DAEAAA8412CD56791EAFCC75082A4C15>. Acesso em 24 mar. 2020.

nuvens não existiam; o arquivo digital relacionado a publicações do *Inter-University Consortium for Political and Social Research* (ICPSR)[13], por exemplo, apenas seria lançado algum tempo depois do dossiê (ALVAREZ et al, 2018).

Nos dias atuais, com o surgimento de plataformas e repositórios *online* adequados, a disponibilização de informações tem se tornado cada dia mais fácil. Nesse sentido, diversos esforços têm sido empreendidos para difundir e implementar os princípios e as práticas da transparência e replicabilidade nas Ciências Sociais. Tais empenhos têm se concentrado em três práticas principais: planos de divulgação, pré-registro e abertura de dados e materiais (MIGUEL et al, 2014), conforme ilustrado na Figura 3.

Figura 3: Três mecanismos para aumentar a transparência nos relatórios científicos (Demonstração com a pergunta de pesquisa "as férias de verão mais curtas melhoram os resultados educacionais?", onde "n.s" significa p-valor > 0,05.)



Fonte: Miguel et al (2014), p.30. Tradução dos autores.

Os planos de divulgação surgem para advogar pela produção de relatórios sistemáticos que buscam garantir que os pesquisadores documentem e divulguem a forma com que procederam quanto a coleta e análise dos dados – essa é uma estratégia muito utilizada na área da medicina e implementada em alguns periódicos acadêmicos na Ciência Política e Psicologia[14] (MIGUEL et al, 2014). Por sua vez, o pré-registro é um documento público onde o pesquisador especifica variáveis, procedimentos de processamentos de dados e técnicas de análise (OLKEN, 2015). A adoção do pré-registro visa diminuir o viés de publicação[15], prevenir a apresentação de resultados seletivos da pesquisa e reforçar a credibilidade desta (CASEY et al, 2012). Por fim, a disponibilização de dados e materiais permitem que outros pesquisadores reproduzam os

13 Disponível em: <www.icpsr.umich.edu/icpsrweb/deposit/pra/index.jsp>. Acesso em 24 mar. 2020.

14 De acordo com o levantamento dos autores, as revistas que adotaram políticas em direção a esse tipo de procedimento são: *Journal of Experimental Political Science*, *Management Science* e *Psychological Science*. Ver: Miguel et al (2014), p. 30.

15

resultados da pesquisa, testem hipóteses remanescentes com os dados e identifiquem resultados não declarados ou, até mesmo, fraudes (MIGUEL et al, 2014; MUFANO et al, 2017; CHRISTENSEN e MIGUEL, 2018).

É na terceira prática que o Protocolo TIER se enquadra. Mais especificamente, na documentação que precede a publicitação de dados e materiais de replicação. A adoção de políticas de replicabilidade por periódicos de alto fator de impacto forçam que as pesquisadoras documentem, de forma sistematizada, os arquivos de suas pesquisas. Apesar da demanda, a qualidade dos materiais disponibilizados não é alta. Stockemer, Koehler e Lenz (2018) analisaram todos os artigos publicados em 2015 nos três principais periódicos da área de comportamento político[16] e identificaram que em 25% deles os dados e os códigos estavam tão mal organizados que tornaram a replicação impossível.

Materiais de replicação bem organizados aprimoram a extensão de pesquisas relacionadas e possibilitam que as análises sejam mais acessíveis (ALVAREZ et al, 2018). No nível mais básico de preparação para se tornar um pesquisador transparente, é recomendado que se adote algum tipo de protocolo de organização desses dados, com o objetivo último de ter dados que sejam passíveis de reutilização e humanamente compreensíveis (FIGUEIREDO et al, 2019; OSINSKI, 2019).

Alvarez, Key e Nuñez (2018) apresentam algumas recomendações para aumentar a organização, clareza e utilização dos dados. Nele, os autores dão dicas sobre como arquivar os materiais relacionados à pesquisa de forma útil e segura, como produzir arquivos LeiaMe que sirvam como ajuda para replicações futuras, indicações sobre como devem ser nomeados os arquivos nos materiais de replicação, além de indicar como devem ser reportados *softwares* (utilizados ou criados pelo autor), *scripts* e *outputs* gerados na pesquisa. O Quadro 1 sumariza as sugestões dos autores.

Quadro 1 - Recomendações para criação de materiais de replicação (Alvarez et al, 2018)

Processos	**Recomendações**
Arquivamento	• Os materiais de replicação devem ser armazenados em arquivos permanentes que estejam disponíveis por longos períodos. Atualmente, os mais confiáveis são os que têm garantias institucionais, sejam de consórcios ou universidades; • Autores devem salvar seus arquivos de modo que possam ser utilizados no futuro (ex: arquivo separado por vírgula), ao invés de formatos *software-specifc*.
Arquivos LeiaMe	• O arquivo deve ser claro e suficientemente detalhado, mas não em demasia. Alguns itens deveriam ser incluídos, a saber: (a) referência do artigo ou publicação associada; (b) uma breve explicação dos arquivos e tipos de arquivos incluídos, ex: dados originais, dados processados, scripts, etc; (c) uma indicação da ordem em que os scripts devem ser executados; (d) uma lista de softwares, pacotes de softwares e sistema operacional utilizado para produzir os resultados do artigo; (e) se extensões incomuns forem utilizadas, o autor deve indicar como proceder com esses arquivos.
Nomes dos arquivos	• Os nomes dos arquivos devem ser fáceis de compreender e prover informação sobre seu conteúdo, isso é especialmente importante para matérias de replicação que incluem vários scripts e arquivos de dados. Caso existam arquivos que devem ser utilizados seguindo determinada ordem, eles devem ser nomeados de modo que incluía a ordem (ex: "1_DadosProcessados"; "2_Estimação", etc.).
Softwares e pacotes criados pelo usuário	• Em casos onde o autor tenha criado um pacote estatístico, é necessário que estes estejam (idealmente) arquivados em um local estável, como o CRAN para o R. Se possível, o autor deve incluir uma cópia do software ou pacote nos materiais de replicação.

16 Os autores analisaram todos os artigos presentes nos volumes de 2015 dos seguintes periódicos acadêmicos*: Electoral Studies, Party Politics* e *Journal of Elections, Public Opinion & Parties*.

Compatibilidade com sistemas operacionais	• Idealmente, é necessário que os autores garantam que seus materiais de replicação funcionem nos sistemas operacionais mais comuns. Caso não seja possível, devem indicar em que sistema eles funcionam.
Scripts	• (a) devem conter uma breve discrição do que fazem, quais as dependências, quais pacotes são necessários e quais são os *outputs*; (b) deve-se evitar linhas sem especificação do que se faz, pois tal indicação ajuda a identificar partes especificas do gerenciamento e processamento de dados e a identificar erros ou possíveis problemas; (c) deve-se planejar bem e evitar recodificação de dados no meio da análise, ao menos quando seja extremamente necessário.
Resultados	• Os materiais de replicação devem conter tabelas, gráficos e figuras da mesma forma que aparecem no artigo.
Resultados excluídos	• Geralmente, materiais de replicação apresentam *scripts* ou dados que não foram apresentados no artigo ou que estavam vinculados a um apêndice online. Nesse caso, o autor deve diferenciá-los dos que estão no artigo.
Resultados intermediários	• Algumas vezes os artigos apresentam resultados intermediários que levam a algo essencial na pesquisa. Nesse caso, devem estar presentes nos materiais de replicação.
Sequências aleatórias	• Para estudos de simulação ou estratégias de estimativa que usem randomização, os autores devem sempre incluir a sequência aleatória de números usada para produzir os resultados no artigo.
Diretórios e pastas	• Os caminhos de diretórios devem ser facilmente identificáveis no *script*, facilitando a alteração pelos usuários. É recomendado que se coloque o diretório apenas no início, modificá-los ao longo do script pode gerar confusão.
Computing time	• Os autores devem indicar (no código e no arquivo LeiaMe) se algum *script* específico demora muito tempo para calcular. Isso é importante para que o usuário não se depare com uma estimativa longa e acredite que esteja passando por algum problema.
Parallel Computing	• Cada vez mais os autores estão utilizando *parallel computing* nos seus *scripts* a fim de acelerar o processo computacional. Infelizmente, isso não funciona em todos os computadores. Logo, devem indicar em que momento do *script* a estratégia foi utilizada.
Avisos e erros	• Não é incomum encontrar *scripts* que produzam erros ou avisos – que geralmente estão relacionados ao uso inadequado de pacotes ou compatibilidade com o sistema operacional. Nesse caso, os autores devem indicar as razões pelas quais eles não são um motivo de preocupação.

Fonte: Alvarez et al (2018), p. 425-426. Tradução dos autores.

As sugestões de Alvarez et al. (2018) são, definitivamente, um caminho para a produção de uma pesquisa mais transparente e replicável. Apesar disso, a quantidade de passos necessários e a falta de um direcionamento mais específico sobre como implementá-los acaba tornando-as um caminho trabalhoso. É nesse sentido que defendemos a adoção do protocolo TIER como uma ferramenta para a organização de materiais de replicação.

Acreditamos que a adoção e ampla divulgação do Protocolo TIER oferece um benefício claro à comunidade acadêmica: ao apresentar uma maneira de padronizar os materiais de replicação, acaba fazendo com que pesquisadores e diferentes áreas saibam, de forma intuitiva, o que devem encontrar em cada uma das pastas. Em outras palavras, a organização dos materiais de replicação se tornará mais orgânica para os pesquisadores e mais intuitivas para os replicadores. Nesse sentido, o protocolo TIER surge como um caminho menos tortuoso à adoção do padrão de replicabilidade.

Tais benefícios estão amplamente fincados nas três dimensões da replicabilidade, a saber: substantiva, pedagógica e transparente. Na dimensão substantiva, o padrão de replicação possibilita que os trabalhos possam ser falseados e, com isso, que a ciência avance mais rápido (PARANHOS et al, 2012; PARANHOS et al, 2013). De modo similar, a adoção do Protocolo TIER permitiu que os trabalhos dos alunos de Ball fossem avaliados de forma mais cautelosa e que os alunos recebessem comentários mais direcionados sobre os problemas enfrentados na

produção do artigo. Por sua vez, para os pesquisadores que resolvem adotar o protocolo, esta estratégia de organização permite que seus trabalhos recebam melhores avaliações pelos seus pares e pelos avaliadores de periódicos especializados.

No que diz respeito à dimensão pedagógica, o padrão de replicação permite que os alunos sejam iniciados no mundo da pesquisa e da análise de dados ao terem contato com dados e problemas do mundo real (KING, 1995). Mais recentemente, os alunos de Ball resolvem exercícios *soup-to-nuts* (mais sobre eles abaixo), que levam os alunos a implementar o protocolo e exercer todo o processo de pesquisa empírica com dados estatísticos, desde a coleta dos dados à escrita do relatório com as análises estatísticas. Nesse mesmo sentido, o *hub* do Projeto TIER no Brasil, baseado em Recife, tem iniciado a produção de *workshops* para alunos de graduação em Ciência Política e Relações Internacionais com o objetivo de difundir a utilização do protocolo e iniciá-los na pesquisa empírica.

Por fim, na dimensão da transparência, o padrão de replicação permite que os procedimentos e dados sejam publicamente disponíveis para todos os pesquisadores que se interessem sobre o tema (KING, 1995; PARANHOS et al, 2013; FIGUEIREDO, 2019). O protocolo permite que a disponibilização das informações necessárias para a replicação seja humanamente inteligível e disponibilizada de modo com que qualquer pesquisador possa realizar a replicação dos resultados do estudo sem necessitar de ajuda do autor. Para os alunos do Ball, possibilitou que os resultados fossem replicados sem a presença dos alunos. Para a ciência, é a maneira das pesquisas, discussões e hipóteses conflitantes avançarem rumo à produção do conhecimento.

Nesse sentido, acreditamos que a adoção do protocolo pode ser uma ferramenta útil na caminhada para os autores das Ciências Sociais que desejam alcançar maiores níveis de transparência em suas pesquisas, auxiliando não só quem está conhecendo o valor da transparência agora, mas também fornecendo um mapa claro sobre como proceder para aqueles que já tinham o interesse, mas não sabiam como. Na próxima seção, apresentamos um exercício *soup-to-nuts* de replicação que serve como uma introdução para aplicação do protocolo.

Exercício de Aplicação do Protocolo

Nesta seção, apresentaremos o primeiro exercício *soup-to-nuts* desenvolvido para exemplificação da aplicabilidade do protocolo (BALL, 2018). Os exercícios *soup-to-nuts* são originalmente desenvolvidos com o intuito de serem aplicados em turmas de Introdução à Estatística ou Introdução à Análise de Dados[17]. Nessas atividades, o intuito é levar os alunos a realizarem um processo completo de pesquisa empírica, desde o download de um banco de dados, até o tratamento, análise e entrega dos resultados. A descrição das etapas é detalhada, e originalmente é prevista para ser introduzida em sala de aula e concluída pelos alunos em casa. Nessa atividade, os alunos devem construir a sua hierarquia de pastas logo no início do projeto, trabalhando, portanto, com o Protocolo TIER ao longo de toda a jornada, uma vez que serão cobrados de realizar a entrega dos resultados com a documentação completa e organizada. Além de facilitar o trabalho do professor na correção dos exercícios, como já comentamos da experiência do Richard Ball, o protocolo ajuda aos alunos em seus fluxos de trabalho, configurando assim um exemplo poderoso das potencialidades dessa metodologia organizacional. Os exercícios de *soup-to-nuts* são, portanto, verdadeiros tutoriais para os alunos trabalharem com análise de dados de maneira transparente e replicável.

17 Ver: https://www.projecttier.org/tier-classroom/soup-nuts-exercises/#shorter-exercises-for-teaching-transparency-and-reproducibility . Acesso em 20 de março de 2020.

Neste artigo, utilizaremos uma versão encurtada do exercício "Consumo de Álcool em Universidades Americanas" como exemplificação do protocolo[18]. Recomendamos que dedique algum tempo para sua elaboração, de maneira que compreenda a maneira pela qual o protocolo contribui ao fluxo de trabalho e facilita a replicação de uma pesquisa. A versão completa e em português deste exercício você encontra na página do OSF[19], e a versão original você encontra no site do TIER[20]. O exercício resolvido, com modelo de relatório e de *scripts*, também estão disponíveis nos endereços indicados. A versão apresentada abaixo é para ser resolvida no *software* R.

Atividade de Replicação (Protocolo TIER): Consumo de Álcool em Universidades Americanas

Nesta atividade, você irá explorar se o consumo de bebidas alcoólicas é menor entre estudantes que vivem em dormitórios onde a prática é proibida em relação àqueles que moram em dormitórios onde não existe essa regra. Os dados utilizados são fruto de uma pesquisa conduzida em 2001 em universidades americanas, da *Harvard School of Public Health College Alcohol Study, 2001*. Essa atividade pedirá que você use os dados dessa pesquisa para criar alguns gráficos de barra e que responda a uma série de questões sobre a interpretação desses gráficos no que se refere aos padrões de consumo de álcool entre estudantes, comparando os que podem beber em seus dormitórios e os que não deveriam. A entrega deverá ser não somente o relatório com os gráficos e as respostas, mas todo o material utilizado para processamento e análise realizados nesta atividade, seguindo a documentação indicada.

I. PREPARATIVOS

IA. Crie uma hierarquia de pastas para guardar e organizar seu trabalho.

Crie uma pasta chamada de **ExercicioAlcool/.**

Dentro de **ExercicioAlcool/**, crie as seguintes pastas e subpastas:

- **Dados/** (uma subpasta de **ExercicioAlcool/)**
    - **DadosOriginais/** (uma subpasta de **Dados/)**
    - **DadosAnalise** (uma subpasta **Dados/**)
- ***Scripts/*** (uma subpasta de **ExercicioAlcool/)**
- ***Graficos/*** (uma subpasta de **ExercicioAlcool/)**

Salve essa hierarquia de pastas no local que escolheu guardar o trabalho para esse exercício.

I.B Instale os pacotes

Dois pacotes serão úteis nessa atividade. Instale antes de iniciar o seu trabalho.

- *haven* (para ler o banco de dados no formato *.dta*)

---

[18] Ball (2018) é o autor da versão original deste exercício, e Jenna Krall é creditada enquanto colaboradora. A tradução para o português e para a linguagem R foi realizada pelos membros do TIER hub Recife. Esta versão encurtada é de responsabilidade dos autores deste artigo, com autorização do autor da versão original.

[19] Versão completa do exercício em Português: https://osf.io/mc2e9/?view_only=32e9fa640c83422ba7d6e9ab91e65743. Acesso em 06 de outubro de 2020.

[20] Versão original completa: https://www.projecttier.org/tier-classroom/soup-nuts-exercises/#shorter-exercises-for-teaching-transparency-and-reproducibility . Acesso em 10 de outubro de 2020.

- *tidyverse (*para facilitar os procedimentos de processamento e criação dos gráficos)

II. BAIXE E EXAMINE O BANCO DE DADOS E O *CODEBOOK* QUE VOCÊ UTILIZARÁ NESSE EXERCÍCIO.

II.A. Encontre e explore o site onde os dados para esse exercício estão disponíveis.

Os dados da pesquisa está arquivado no *Inter-University Consortium for Political and Social Research* (ICPSR).

- Vá até o site do ICPSR (www.icpsr.umich.edu) e procure por *Harvard School of Public Health College Alcohol Study, 2001*.
- Quando você tiver encontrado a página desse estudo, leia as informações fornecidas.

II.B. Baixe o banco de dados e o *codebook*.

- Antes de baixar o banco de dados, você precisará criar uma conta no site do ICPSR.
- Uma vez feito isso, faça o login e vá para a página do estudo: https://www.icpsr.umich.edu/icpsr-web/ICPSR/studies/4291
- Clique no botão de *download*. Você verá um menu de formatos para escolher, incluindo SAS, SPSS, Stata ou ASCII. Nesse menu, selecione Stata (com o pacote *haven* o R irá ler o formato *.dta*).

O arquivo do banco será baixado numa pasta zipada, *ICPSR_04291-V2.zip*.

Quando você extrair os arquivos, verá que são vários documentos, incluindo uma subpasta chamada **DS0001**. Para esse exercício, você usará somente dois desses arquivos que vieram na pasta zipada e ambos estão na subpasta **DS0001**:

- *04291-0001-Data.dta*. Esse arquivo contém os dados da pesquisa em formato *.dta*. Esse arquivo será referido como "o banco de dados original" nesse exercício.
- *04291-0001-Codebook.pdf*. Esse é o *codebook* para a pesquisa utilizada. Será referido como "metadados do banco de dados original".

Você deve salvar cópias desses dois arquivos na pasta **DadosOriginais/**.

II. C. Examine o *codebook* e o banco de dados.

Examine o *codebook* e o banco de dados original para se familiarizar com as informações que eles contêm. Para esse exercício, você precisará utilizar cinco variáveis desse estudo. No banco de dados original, essas variáveis têm os nomes *A6, B8, B9, C13* e *G11*. Para cada uma dessas variáveis, analisando o *codebook*, responda as seguintes questões (as respostas a essas perguntas não precisarão estar no relatório final, mas são fundamentais para que você compreenda a natureza dos dados com que lidará):

- Como a variável é definida? O que ela representa?
- Qual a classe da variável (contínua, ordinal, categórica, discreta?)
- É qualitativa ou quantitativa?
    - Se for quantitativa, encontre a média, o desvio padrão, os três quartis, o valor mínimo e o valor máximo.
    - Se for qualitativa categórica, quantas categorias existem, o que elas representam e qual proporção das observações por categorias? Para cada variável categórica, decida se as categorias estão ordenadas ou não.

III. ESCREVENDO OS *SCRIPTS*

Convenções para o diretório de trabalho

Em todos os *scripts*, adote as seguintes convenções para o diretório de trabalho:

- Quando seus *scripts* forem executados, o diretório de trabalho do R deve ser definido para a pasta ***Scripts/***.
- Não escreva comandos em seus *scripts* que mudem o diretório de trabalho. Por exemplo, evite utilizar a função *setwd()*.
- Quando você precisar abrir ou salvar um arquivo em outra pasta que não ***Scripts/***, especifique a localização da pasta em questão utilizando um *relative directory path*.

Para esse exercício, você escreverá três *scripts*:

III.A. *tratamento.R*

- O propósito desse *script* é modificar o banco original (*04291-001-Data.dta*) para prepará-lo para a análise que você vai realizar.
    - As modificações incluem a criação de uma série de novas variáveis e a exclusão de algumas observações do banco.
- Salve o novo banco em um arquivo chamado *analise.RData*.
    - Esse arquivo será chamado de banco de análise.

Criando o *script*

- Abra o RStudio.
- Abra um novo *script* de R.
- No início do novo *script*, escreva um comentário indicando que o diretório de trabalho do R deve ser a pasta **Scripts/**. Por exemplo:

    # Quando esse *script* for executado, o diretório de trabalho deve ser a pasta **Scripts/**.
- Salve esse novo *script* na sua pasta **Scripts/**, com o nome *tratamento.R*.
    - Para evitar perder trabalho, salve o *scripts* constantemente enquanto o escreve.
- Carrege os pacotes *haven* e *tidyverse*
- Abra (importe) o banco *04291-0001-Data.dta*.
    - Use a função *read_dta()* do pacote *haven*.
    - Como seu diretório de trabalho é a pasta **Scripts/**, você precisa usar um *relative directory path* para especificar que o arquivo *04291-0001-Data.dta* deve ser importado de **DadosOriginais/**.
- Para essa atividade, nós queremos considerar apenas estudantes que vivem no campus ou em fraternidades/sororidades. Assim, exclua todos os casos os quais a residência do aluno seja "*Off campus house/apt*" ou "*other*" ("fora de casa/apto no campus" ou "outro"). Também exclua todos os casos onde a variável representando a residência do aluno seja uma NA.
- Crie uma variável chamada *drunk*, que seja:
    - Igual a 0 se o estudante bebeu o suficiente para ficar bêbado menos do que três vezes nos últimos trinta dias.
    - Igual a 1 se o estudante bebeu o suficiente para ficar bêbado três ou mais vezes nos últimos trinta dias.
    - Igual a NA se a variável indicando quantas vezes o estudante bebeu o suficiente para ficar bêbado nos últimos trinta dias for NA.
- Crie uma variável chamada *hsdrunk*, que seja:
    - Igual a 0 se o estudante bebeu cinco ou mais bebidas seguidas em duas ou menos ocasiões no último ano do High School

    - Igual a 1 se o estudante bebeu cinco ou mais bebidas seguidas em três ou mais ocasiões no último ano do High School
    - Igual a NA se a variável indicando quantas vezes o estudante bebeu cinco ou mais bebidas seguidas no último ano do High School for uma NA
- Crie uma variável chamada *free*, que seja:
    - Igual a 0 se o estudante não vive num dormitório onde é proibido consumir bebida alcoólica (*alcohol-free housing*)
    - Igual a 1 se o estudante vive num dormitório onde é proibido consumir bebida alcoólica (*alcohol-free housing*)
    - Igual a NA se a variável indicando se o estudante vive ou não num dormitório onde é proibido consumir bebida alcoólica for NA.
- Crie uma variável chamada *volfree*, que seja:
    - Igual a 1 se *free*=1 (o estudante vive num dormitório onde é proibido ingerir bebida alcoólica) e o aluno requisitou morar num dormitório desse tipo
    - Igual a 0 se *free=1* (o estudante vive num dormitório onde é proibido ingerir bebida alcoólica) e o aluno foi designado a viver num dormitório desse tipo
    - Igual a NA em todos os outros casos
- Crie uma variável chamada *housing*, que seja:
    - Igual a 1 se *free*=0 (o estudante não vive num dormitório onde é proibido ingerir bebida alcoólica)
    - Igual a 2 se *free*=1 (o estudante vive num dormitório onde é proibido ingerir bebida alcoólica) e o aluno foi designado a viver num dormitório desse tipo
    - Igual a 3 se *free* = 1 (o estudante vive num dormitório onde é proibido ingerir bebida alcoólica) e todas as habitações do campus possuem a regra de proibição de consumo de bebida alcoólica.
    - Igual a 4 se *free* = 1 (o estudante vive num dormitório onde é proibido ingerir bebida alcoólica) e o aluno requisitou morar num dormitório desse tipo
    - Igual a NA em todos os outros casos

- Exclua todas as variáveis que não sejam as seis que você acabou de criar.
- Exclua todos os casos onde o valor de *hsdrunk, drunk* ou *housing* seja NA.
- Salve o novo banco de dados na pasta **DadosAnalise/,** nomeando o arquivo de *Analise.RData*.
    - Você salva o banco com a função *save()*.
    - Aqui também você precisará indicar um diretório de trabalho relativo para especificar a localização da pasta **DadosAnalise/**.

Lembre-se de salvar tratamento.R na sua pasta ***Scripts/***.

III.B. *ApendiceDados.R*

O *script* ApendiceDados.R irá conter comandos que geram as informações que vão aparecer no Apêndice de Dados (mais informações abaixo).

Criando o *script*

- Abra o RStudio.
- Abra um novo *script* de R.
- Abra o arquivo do banco de análise, analise.RData (que, após ter escrito e rodado tratamento.R, deve estar salvo na pasta **DadosAnalise/**).
- Crie os seguintes resultados para cada uma das cinco variáveis no banco de análise:

- O número de NAs e o número total de observações (incluindo NAs)
- A distribuição da frequência e a distribuição da frequência percentual
- Um gráfico de barra que apresente a distribuição da frequência percentual.

Lembre-se de escrever comentários explicando quais linhas produzem qual informação e gráfico. Cada linha que produzir informação ou gráfico que componha o Apêndice de Dados deve ser precedido por um comentário que descreva o que se espera que esse comando faça. Por exemplo, acima de uma linha que gere uma tabela mostrando a distribuição da frequência da variável *housing*, você escreve um comentário do tipo "A seguinte linha gera uma tabela mostrando a frequência da distribuição de *housing*".

Lembre-se de salvar ApendiceDados.R em sua pasta ***Scripts/***.

III.C. *analise.R*

O analise.R gera os gráficos de barra que constituem a principal análise que você fará para essa atividade, utilizando os dados do seu banco de dados de análise (analise.RData).

Criando o *script*

- Abra o RStudio.
- Abra um novo *script* de R.
- Abra o analise.RData (que, depois de você ter escrito e rodado tratamento.R, deve estar salvo na sua pasta **DadosAnalise/**).
- Crie um gráfico de barras, com duas barras, onde:
  - Uma barra represente estudantes vivendo em dormitórios onde é proibido consumir bebida alcoólica
  - Uma barra represente estudantes que não vivem em dormitórios onde é proibido consumir bebida alcoólica
  - A altura de cada barra é igual à proporção de estudantes (do grupo de estudantes que a barra representa) que beberam o suficiente para ficar bêbado três ou mais vezes nos últimos trinta dias.
  - Salve esse gráfico na pasta **Graficos/** (subpasta de **ExercicioAlcool/**), com o nome Figura1.jpg. (Use um diretório de trabalho relativo para especificar a localização da subpasta **Graficos/)**

- Crie um gráfico de barra, com duas barras, onde:
  - Uma barra represente estudantes que vivem em dormitórios onde é proibido consumir bebida alcoólica porque eles requisitaram isso
  - Uma barra represente estudantes que vivem em dormitórios onde é proibido consumir bebida alcoólica porque eles foram designados
  - A altura de cada barra deve ser igual à proporção de estudantes (no grupo representado pelas barras) que beberam o suficiente para ficar bêbados três ou mais vezes nos últimos trinta dias
  - Salve esse gráfico na pasta **Graficos/** (subpasta da pasta **ExercicioAlcool/**), com o nome Figura2.jpg.

- Crie um gráfico de barrar, com quatro barras, onde:
  - Cada barra represente estudantes em uma das quatro categorias definidas na variável *housing*
  - A altura de cada barra é igual à proporção de estudantes (no grupo representado pelas barras) que beberam o suficiente para ficar bêbados três ou mais vezes nos últimos trinta dias

- Salve esse gráfico na pasta **Graficos/** (subpasta da pasta **ExercicioAlcool/**), com o nome Figura3.jpg.

- Crie dois gráficos de barra (lado-a-lado), cada um possuindo duas barras, onde:
  - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em três ou mais ocasiões no último ano do *High School*
  - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em menos de três ocasiões em seu último ano de *High School*
  - E em cada um desses gráficos,
    - Uma barra representa estudantes que vivem em dormitórios onde o consumo de álcool é proibido
    - Uma barra representa estudantes que não vivem em dormitórios onde há essa proibição
    - A altura de cada barra é igual à proporção de estudantes (no grupo representado pelas barras) que beberam o suficiente para ficar bêbados três ou mais vezes nos últimos trinta dias
    - Salve esse gráfico na pasta **Graficos/** (subpasta da pasta **ExercicioAlcool/**), com o nome Figura4.jpg.

- Crie dois gráficos de barra (lado-a-lado), cada um possuindo duas barras, onde:
  - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em três ou mais ocasiões no último ano do *High School*
  - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em menos de três ocasiões em seu último ano de *High School*
  - E em cada um desses gráficos,
    - Uma barra represente estudantes que vivem em dormitórios onde é proibido o consumo de bebidas alcoólicas porque requisitaram
    - Uma barra represente estudantes que vivem em dormitórios onde é proibido o consumo de bebida alcoólica por que eles foram designados
    - A altura de cada barra é igual à proporção de estudantes (no grupo representado pelas barras) que beberam o suficiente para ficar bêbados três ou mais vezes nos últimos trinta dias
    - Salve esse gráfico na pasta **Graficos/** (subpasta da pasta **ExercicioAlcool/**), com o nome Figura5.jpg.
- Crie dois gráficos de barra (lado-a-lado), cada um possuindo quatro barras, onde:
  - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em três ou mais ocasiões no último ano do *High School*
  - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em menos de três ocasiões em seu último ano do *High School*
  - E em cada um desses gráficos,
    - Cada barra represente estudantes em uma das quatro categorias definidas na variável *housing*
    - A altura de cada barra é igual à proporção de estudantes (no grupo representado pelas barras) que beberam o suficiente para ficar bêbados três ou mais vezes nos últimos trinta dias
    - Salve esse gráfico na pasta **Graficos/** (subpasta da pasta **ExercicioAlcool/**), com o nome Figura6.jpg.

Escreva comentários indicando quais linhas produziram que gráficos. Por exemplo, antes da linha que gera a Figura 4, escreva um comentário como “A seguinte linha de comando gera a Figura 4”.

Lembre-se de salvar analise.R em sua pasta de ***Scripts&Relatorio/***.

IV. O RELATÓRIO

O relatório que você deve entregar para essa atividade consistirá na apresentação das Figuras que foram geradas e de respostas a uma série de perguntas que pedem que você interprete os gráficos de barra que você criou. Salve em *.pdf.*

Apresente as Figuras e escreva uma legenda abaixo de cada gráfico que indique a numeração e um título informativo. Por exemplo, uma legenda apropriada para a Figura 1 seria: “Figura 1: Taxa de Consumo Excessivo de Álcool em Dormitórios onde é proibido consumir álcool vs. Dormitórios onde não há essa proibição”.

Questões

1) Descreva a amostra no arquivo BancoAnalise.dta que você criou. Qual a unidade de análise? Quantas observações existem? Descreva o método utilizado para criar a amostra e quais critérios foram considerados para incluir ou excluir grupos de indivíduos.

Note que essa pergunta é sobre seu banco limpo e processado, ao invés do banco original que você baixou do ICPSR. De qualquer forma, para responder, você terá que recorrer ao *codebook* do banco original.

2) Explique em suas palavras o que a Figura 1 apresenta. Era o que você esperava? Se não, como a Figura 1 é diferente do que você esperava?

3) Explique em suas palavras o que a Figura 2 apresenta. O que você vê na Figura 2 repercute de alguma forma o que você viu na Figura 1? Explique.

4) Quantos indivíduos são representados na Figura 1? (Ou seja, quantas observações foram utilizadas para gerar a figura?). E quantos indivíduos estão representados na Figura 2? Se o número de indivíduos representados na Figura 1 e Figura 2 não é o mesmo, explique: alguns indivíduos apresentados na Figura 1 não estão representados na Figura 2? Se sim, que indivíduos foram esse? E alguns indivíduos apresentados na Figura 2 não foram representados na Figura 1? Se sim, quais?

5) Considerando a Figura 3, quais partes contêm informação que já foram apresentadas nas Figuras 1 e 2? Que partes apresentam nova informação? Resuma em suas palavras o que a Figura 3 apresenta.

6) Compare a Figura 4 com a Figura 1. Explique que novas informações são fornecidas pela Figura 4 e apresente uma interpretação dessas informações.

7) Compare a Figura 5 com a Figura 2. Explique que nova informação é fornecida pela Figura 4 e apresente uma interpretação.

8) Compare a Figura 6 com a Figura 3. Explique que nova informação é fornecida pela Figura 4 e apresente uma interpretação.

9) Das figuras que você criou, quais conclusões gerais podem ser retiradas sobre os fatores associados ao consumo de álcool entre estudantes universitários?

V. O APÊNDICE DE DADOS

O Apêndice de Dados deve consistir em múltiplas seções, uma para cada variável do banco de dados de análise. Salve também em *.pdf*.

Para cada variável, a seção deve fornecer:

- O nome da variável.
- O número de NAs no formato NA/Total (quantidade de NAs/número total de observações).
- Uma definição da variável.
- Possíveis valores da variável.
- Uma explicação do que cada valor da variável representa (codifição dos valores).
- Uma tabela apresentando a distribuição da frequência e a distribuição da frequência percentual. Se a variável categórica é ordenada, essa tabela deve também apresentar distribuição da frequência percentual acumulativa.
- Um gráfico de barra que ilustre a distribuição da frequência percentual.

VI. O ARQUIVO LEIA-ME

O arquivo Leia-Me é um documento que apresenta informações sobre os arquivos eletrônicos que você organizou para documentar o trabalho desta análise. Deve consistir em duas seções principais:

> A primeira seção deve apresentar um mapa de todos os arquivos incluídos na documentação da replicação, assim como as pastas e subpastas onde estão localizados.
>
> A segunda seção deve fornecer instruções passo-a-passo explicando como utilizar a documentação para (i) replicar todo o processamento de dados necessário para transformar o banco original no banco de análise, (ii) gerar todos os resultados que você apresenta no Apêndice de Dados, e (iii) reproduzir os seis gráficos de barra que você criou.

Essas instruções devem ser escritas em português claro e devem ser detalhadas o suficiente para que alguém não familiarizado com esse exercício consiga seguir e reproduzir tudo que você fez. Salve em *.pdf*.

VII. DOCUMENTAÇÃO ELETRÔNICA

Uma vez finalizado o exercício, os arquivos utilizados e criados devem estar alocados na hierarquia de pastas da seguinte maneira:

ExercicioAlcool/

Leia-Me.pdf

Relatorio.pdf

Dados/ (subpasta de **ExercicioAlcool/**)

DadosOriginais (subpasta de **Dados/**)

04291-0001-Data.dta

04291-0001-Codebook.pdf

DadosAnalise (subpasta e **Dados/**)

BancoAnalise.dta

Scripts/ (subpasta de **ExercicioAlcool/**)

tratamento.R

ApendiceDados.R

analise.R

Graficos/ (subpasta de **ExercicioAlcool/**)

Figura1.jpg

Figura2. jpg

Figura3. jpg

Figura4. jpg

Figura5. jpg

Figura6. jpg

Não deixe nada além disso nessas pastas

- Sem nenhuma pasta além dessas.
- Sem nenhum arquivo além desses

Teste seus arquivos para saber se rodam

- Encontre um computador que não seja o que você realizou a atividade e copie para ele a pasta inteira, com todo seu conteúdo.
- Siga os seguintes passos, sem modificar a ordem das pastas ou os arquivos:
  - Inicie o R;
  - Defina ***Scripts/*** como seu working directory
  - Rode o tratamento.R e após isso, confira se o arquivo BancoAnalise.dta que você tem na pasta **DadosAnalise/** foi alterado por uma nova cópia criada recentemente (confira a data e hora de modificação do arquivo na pasta)
  - Rode o ApendiceDados.R e verifique se gera os resultados esperados. Confira também se os gráficos na pasta **Graficos/** foram atualizados por novas versões recentes.

Rode o analise.R e confira se os seis arquivos de gráficos localizados na pasta **Graficos/** foram atualizados.

Outras atividades do TIER

Com o intuito de promover a importância da transparência e de difundir os valores da ética na ciência, o Projeto TIER desenvolve algumas atividades e eventos de promoção de seu protocolo e de boas práticas científicas. Todas podem ser encontradas em maiores detalhes em seu endereço da web[21].O Projeto TIER promove workshops voltados a professores e pesquisadores universitários, tanto para ouvi-los sobre suas práticas em prol da transparência e da replicabilidade na ciência, quanto para apresentá-los às ideias mais recentes sobre o protocolo (e colher

21 Ver: https://www.projecttier.org/ . Acesso em 20 de março de 2020.

comentários e impressões)[22]. Esses workshops são ótimas oportunidades de cooperação interuniversidades e de cooperação internacional, em termos de boas práticas científicas, reunindo professores de distintos campos das ciências sociais e de diferentes países.

Há ainda o programa de *Fellowship*, voltado a jovens professores pesquisadores engajados com ensino e treinamento de métodos quantitativos sob uma lente transparente e replicável. As bolsas valem por um ano acadêmico e buscam incentivar que o recipiente desenvolva e difunda métodos de ensino de pesquisa transparente – como exercícios *soup-to-nuts* e *workshops*. Além disso, o TIER promoveu também nos meses de fevereiro e março de 2020 uma série de conferências no formato de webcast. A cada semana, um convidado fazia uma fala sobre os seus mais recentes desenvolvimentos e ideias sobre transparência e replicabilidade. Dentre esses "líderes da transparência", estavam Gary King, Dorothy Bishop e Scott Long. As apresentações estão disponíveis no site do TIER[23].

Considerações Finais

Os princípios e as práticas de transparência têm se tornado o padrão-ouro da ciência empírica. Iniciativas para difundir essas ideias e apresentar ferramentas para aumentar a transparência nas pesquisas e na produção científica sobre o assunto tem se tornado cada vez mais relevantes. Materiais de replicação ganham força enquanto elementos que compõe uma pesquisa concluída, e a divulgação e o compartilhamento desses têm sido um requisito para publicação em diversos periódicos especializados, seja em território americano e, embora ainda de forma incipiente, aqui no Brasil.

Nesse trabalho, tivemos o objetivo de discutir formas de aumentar a transparência de pesquisas empíricas nas Ciências Sociais, principalmente através da documentação dos materiais de replicação. Apresentamos, ao longo do artigo, o surgimento do Protocolo TIER e seu enquadramento nas três dimensões do padrão de replicação (substantiva, pedagógica e transparência), uma vez que o protocolo permite a continuidade da produção acadêmica e a movimentação da roda da ciência, possibilita que jovens pesquisadoras sejam iniciadas no mundo da pesquisa empírica através de dados estatísticos e análise de dados, além de aumentar o nível de transparência dos trabalhos que o adotam. Além disso, apresentamos a estrutura organizacional do protocolo, os arquivos comportados e a hierarquia de pastas que pressupõem a documentação sugerida pelos criadores. Apresentamos, ainda, uma versão encurtada de um exercício *soup-to-nuts* para implementação do protocolo, ideal para aplicação em turmas de introdução a estatística e análise de dados. Esperamos que os alunos que sejam introduzidos a essas análises já dentro de um paradigma transparente e replicável tenham não só mais facilidade em aprender e em desenvolvê-las, mas que também carreguem consigo esses valores de boa ciência.

Com isso, visamos indicar um caminho menos tortuoso para que a pesquisadora alcance níveis mais altos de transparência em sua pesquisa. Além disso, esperamos ter difundido e apresentado de forma simples e didática o protocolo. Dessa maneira, esperamos contribuir com a agenda de pesquisa sobre transparência e replicabilidade, que vem crescendo na Ciência Social como um todo, e na Ciência Política brasileira em especial.

---

22 Ver: https://www.projecttier.org/fellowships-and-workshops/fall-2019-faculty-development-workshop/ . Acesso em 20 de março de 2020.
23 Ver: https://www.projecttier.org/fellowships-and-workshops/weekly-webcast-leaders-research-transparency/ . Acesso em 20 de março de 2020.

Referências bibliográficas

ALVAREZ, R. M; KEY, E. M; NÚÑEZ, L. (2018). "Research replication: Practical considerations". *PS: Political Science & Politics*, 51(2), 422-426.

BALL, R.; MEDEIROS, N. (2012). "Teaching Integrity in Empirical Research: A Protocol for Documenting Data Management and Analysis". *The Journal of Economic Education*, 43(2), 182-189.

BALL, R. (2018). Animal House in Alcohol-Free Dorms? TIER Project. Disponível em: https://www.projecttier.org/tier-classroom/soup-nuts-exercises/#shorter-exercises-for-teaching-transparency-and-reproducibility. Acesso em 06 de outubro de 2020.

CASEY, K; GLENNERSTER, R; MIGUEL, E. (2012). "Reshaping institutions: Evidence on aid impacts using a preanalysis plan". *The Quarterly Journal of Economics*, 127(4), 1755-1812.

CHRISTENSEN, G.; SODERBERG, C. (2016). "Manual of best practices in transparent social science research". *Retrieved April*, 2, 2017.

CHRISTENSEN, G.; MIGUEL, E. (2018). "Transparency, reproducibility, and the credibility of economics research". *Journal of Economic Literature*, 56(3), 920-80.

DE LA GUARDIA, F. H.; STURDY, J. (2019). "Best Practices for Transparent, Reproducible, and Ethical Research". *IDB Technical Note* ; 1635.

FIGUEIREDO FILHO, D.; LINS, R.; DOMINGOS, A.; JANZ, N.; SILVA, L. (2019). "Seven Reasons Why: A User's Guide to Transparency and Reproducibility". *Brazilian Political Science Review*, 13(2).

JANZ, N. (2016). "Bringing the gold standard into the classroom: replication in university teaching". *International Studies Perspectives*, 17(4), 392-407.

KEY, Ellen M. (2016). "How are we doing? Data access and replication in political science. PS: Political Science & Politics, 2016, 49.2: 268-272

KING, G. (1995). "Replication, replication". *PS: Political Science & Politics*, 28(3), 444-452.

MIGUEL, E., CAMERER, C., CASEY, K., COHEN, J., ESTERLING, K. M., GERBER, A., LAITIN, D. (2014). "Promoting transparency in social science research". *Science*, 343(6166), 30-31.

MUNAFÒ, M. R., NOSEK, B. A., BISHOP, D. V., BUTTON, K. S., CHAMBERS, C. D., DU SERT, N. P., IOANNIDIS, J. P. (2017). "A manifesto for reproducible science". *Nature human behaviour*, 1(1), 1-9.

OLKEN, B. A. (2015). "Promises and perils of pre-analysis plans". *Journal of Economic Perspectives*, 29(3), 61-80.

OSINSKI, L. (2019). "PROOF course Open Science: The new default in science, 01-10-2019". *Presentation at Eindhoven University of Technology*. Disponível em: <https://www.projecttier.org/tier-classroom/course-materials/#modal230>. Acesso em 23 mar. 2020.

PARANHOS, R., FIGUEIREDO FILHO, D. B., da Rocha, E. C., da Silva Jr, J. A., & SANTOS, M. L. W. D. (2012). “Levando Gary King a sério: desenhos de pesquisa em Ciência Política”. *Revista Eletrônica de Ciência Política*, 3(1-2).

PARANHOS, R., FIGUEIREDO FILHO, D. B., da ROCHA, E. C., CARMO, E. F. (2013). “A importância da replicabilidade na ciência política: o caso do SIGOBR”. *Revista Política Hoje*, 22(2), 213-229.

PRZEWORSKI, A., SALOMON, F. (1988). “The Art of Writing Proposals: Some candid suggestions for Applicants to Social Science Research Council”. *SSRC: New York*.