Posfácio

Ronaldo Lima Jr. (UFC)

Sinto-me extremamente honrado em redigir estas palavras finais para o livro da Livia Oushiro. Seu trabalho teve bastante impacto no meu, e, por isso, sou muito grato às suas empreitadas em prol da capacitação de linguistas em análise quantitativa de dados.

Conheci a Livia Oushiro e seu material "Introdução à Estatística para Linguistas" em 2017, por meio do grupo de Facebook que ela havia criado, "R para Linguistas". A data é memorável, pois, no segundo semestre daquele ano, ministrei pela primeira vez parte de uma disciplina sobre análise quantitativa de dados no Programa de Pós-Graduação em Linguística (PPGL) da minha universidade, a Universidade Federal do Ceará (UFC), e a versão em swirl do material, em que é possível fazer as aulas de forma interativa dentro do próprio R, caiu como uma luva. Ao consultar a Livia se poderia usar seu material com meus alunos na disciplina, percebi que estava diante de uma professora genuína, com a generosidade acadêmica que tanto admiro e da qual a academia carece. Eu fiz todas as aulas em swirl para ver como ela tinha organizado o curso, e a forma dialogada de suas aulas me chamou a atenção. Acabei planejando minha parte da disciplina seguindo a mesma sequência proposta pela Livia em seu material, utilizei diversos dos seus dados e exemplos ao longo das aulas, e as tarefas de casa dos meus alunos eram o caderno de exercícios criado por ela também em swirl. O programa da recém-criada disciplina "Análise Quantitativa de Dados em Linguística" do PPGL da UFC ainda contém muito dessa estrutura inicial. Além disso, sempre que ministro algum curso ou palestra sobre assunto[10], recomendo os materiais da Livia, e agora será um prazer recomendá-lo como livro publicado pela Editora da Abralin.

[10] Disponíveis em https://ronaldolimajr.github.io.

A importância da análise quantitativa de dados na Linguística

Se você chegou até o posfácio desta obra, provavelmente não precisa de muito convencimento sobre a importância de análises estatísticas na ciência como um todo, e, consequentemente, na Linguística. Mesmo assim, gostaria de registrar que acredito que o papel de uma capacitação em análise quantitativa de dados vai além de sua aplicação direta a dados quantitativos. Há subáreas da Linguística que exigem análises estatísticas mais do que outras (como a Fonética, a Psicolinguística, a Sociolinguística e, mais recentemente, a Linguística de Corpus), mas acredito que mesmo pesquisadores de subáreas que conduzem estudos exclusivamente qualitativos podem se beneficiar do conhecimento em análise quantitativa de dados. Isso porque o raciocínio sobre desenho experimental como um todo pode auxiliar no estabelecimento de objetivos e perguntas de pesquisa coerentes, em que a operacionabilidade do estudo é posta à prova pelo próprio pesquisador ao considerar questões como correlações espúrias, variáveis de confusão (*confounding variables*), limitações da amostra, e inferências sobre as relações de causalidade (que são de responsabilidade exclusiva do pesquisador, não informadas pelos dados observados).

Um exemplo: um alienígena observando um humano abrindo os olhos exatamente quando um despertador toca todas as manhãs poderia se questionar se os olhos abrindo estão fazendo o relógio despertar, ou se o despertador tocando está fazendo os olhos abrirem. Este é um questionamento que nunca faríamos, mas que deveríamos praticar fazer principalmente ao elaborar as questões de pesquisa, de maneira a torná-las operacionalizáveis. Humanos sabem que é o despertador que causa os olhos abrirem, que é um acidente de trânsito que causa o congestionamento (e raramente o contrário), e que estar gripado nos faz tossir (e não tossir nos causa gripe). Contudo, precisamos nos lembrar que desconhecemos as relações de causa do nosso objeto de pesquisa (por isso o pesquisamos), e que, sendo assim, precisamos de cautela nas inferências. Parece ser muito lógico, por exemplo, ao observar que aprendizes motivados

apresentam maior desenvolvimento em uma língua estrangeira, concluir que a motivação causa aprendizagem. Contudo, o raciocínio estatístico leva o pesquisador a questionar questões como: mas e os alunos motivados que não desenvolvem tanto a L2? E os desmotivados que aprendem a L2 mesmo assim? Perceber o desenvolvimento de sucesso na nova língua não aumenta também a motivação? Não haveria outros fatores que poderiam estar causando tanto a motivação como o maior desenvolvimento linguístico? Ou seja, mesmo diante de dados qualitativos (se fosse o caso aqui), ou mesmo antes de se coletar os dados, um pesquisador com conhecimento e pensamento estatístico é capaz de levantar questões sobre a metodologia de sua própria pesquisa que podem torná-la mais robusta e, consequentemente, com inferências mais confiáveis e replicáveis.

Dada essa importância do pensamento quantitativo na análise de dados linguísticos, é inquestionável o importante papel deste livro para a Linguística. Concluo, portanto, meu posfácio com breves sugestões para pesquisadores em diferentes níveis de proficiência em análise quantitativa de dados. Como iniciantes, penso em absolutamente qualquer pessoa que tomou interesse pelo assunto e que tenha apenas vagas lembranças da matemática aprendida na escola (público-alvo potencial deste livro). Como intermediário, me refiro a alguém que chegou até a Lição 11 deste livro com segurança e que talvez ainda esteja firmando seus conhecimentos das demais lições. Avançado aqui se refere aos que terminarem as Lições 12–15 com bastante familiaridade e confiança.

Para os iniciantes: meu testemunho sobre os primeiros passos em estatística

Alguns estudantes e colegas linguistas que estão buscando seus primeiros passos em análise quantitativa de dados se identificam com o início da minha trajetória. Por isso, aproveito para registrá-la como forma de motivação e encorajamento, juntamente com algumas sugestões para quem está no início da caminhada em análise estatística.

Logo no início do meu doutorado me dei conta que precisaria de alguma análise estatística dos dados que estava planejando coletar e analisar – dados acústicos de produção de fala de alunos de inglês como L2. Lendo outros trabalhos da área, descobri que precisaria de um tal de teste-t. Comprei, então, o livro "*A Guide to Doing Statistics in Second Language Research Using SPSS*" (LARSON-HALL, 2010), já que o título parecia perfeito para a minha pesquisa, e, com pressa, fui direto ao capítulo sobre teste-t. Resultado: não entendi nada! Percebi que precisaria ler desde o início para saber minimamente o que eram "variáveis", "amostra" e outros conceitos básicos. Sabia, por exemplo, que um valor de *p* menor que 0,05 indicaria uma diferença significativa entre meus grupos de aprendizes, mas não fazia ideia do que realmente ele representava (hoje percebo, que perigo!). Li o livro do início e confesso que compreendi apenas uns 50%. Passei, então, a fazer cursos on-line e a assistir videoaulas, e, quando me senti com mais base, reli o livro e experimentei um momento "aha!", quando tudo passou a fazer sentido!

Não acho que essa deva ser a rota de todos, e nem acho que o livro que utilizei seja o melhor para começar, mas registro aqui minha experiência inicial para mostrar que percebi que o conhecimento em análise quantitativa de dados é cumulativo e normalmente requer um estudo em ciclos de idas e vindas. Algumas lições que tenho aprendido na minha jornada em análise quantitativa de dados e que gostaria de deixar para quem está começando são:

- Não tem como fugir do básico! Antes de conduzir sua análise, mesmo sabendo qual será e qual valor você espera como resultado, é preciso compreender conceitos como amostra e população, (tipos de) variáveis, tamanho de efeito, intervalo de confiança, e compreender, minimamente, como se chega ao valor que você vai reportar (como o valor de *p*);
- Nem sempre compreendemos um conceito de primeira, mas isso não é motivo para desistir, e vale a pena insistir. Dê um tempo e depois volte ao conceito – procure por pessoas diferentes, em vídeos e páginas eletrônicas, explicando o mesmo conceito de maneiras diferentes, com outros exemplos;

- Às vezes um conceito que ficou muito claro em um momento passa a ficar menos claro com o tempo – volte ao conceito e busque por exemplos para reativar sua memória;
- Persista e capacite-se para realizar suas próprias análises, sem contar com programas automáticos ou com pessoas contratadas para isso, pois conduzir suas próprias análises trará uma compreensão muito maior sobre os seus dados e sobre os fenômenos linguísticos sob investigação.

Aos poucos a análise quantitativa de dados passou a fazer parte dos meus interesses acadêmicos, e, por isso, continuo estudando e colocando esses passos em prática. Em muitos momentos a curva de aprendizagem é bastante íngrime, mas compensa persistir e retornar aos trechos desafiadores.

Para os intermediários: cautelas na análise quantitativa de dados

É importante manter em mente que, ao mesmo tempo que a análise estatística dos dados é absolutamente necessária em vários tipos de estudos, ela é apenas um instrumento, uma ferramenta. Sendo assim, quanto maior o domínio da ferramenta, maior será a compreensão de suas atribuições, capacidades e limitações. Por um lado, pesquisadores iniciantes em análises quantitativas podem inadvertidamente propor inferências e conclusões de maneira muito incisiva e categórica com base em uma análise frágil. Um valor de p abaixo de 0,05, por exemplo, nem sempre é suficiente para se determinar um efeito de maneira contundente. Por outro lado, pesquisadores com bastante bagagem estatística podem esquecer-se de que suas pesquisas em Linguística não devem ter como fim a análise, e que um retorno à teoria e às contribuições para o conhecimento de Linguística devem ser sempre priorizados. Sendo assim, deixo aqui duas recomendações para quem já domina testes de hipótese ou até mesmo rode alguns modelos de regressão:

Não dependa exclusivamente do valor de p

Como a Livia Oushiro deixa claro ao longo do livro, há diversas limitações numa interpretação de análise estatística exclusivamente dependente do resultado do valor de *p*.

Lembre-se que o valor de *p* é apenas a probabilidade dos seus dados diante da hipótese nula e que, por isso, ele:

- não diz nada sobre sua hipótese de trabalho (a hipótese alternativa);
- não diz nada sobre o tamanho do efeito;
- não tem gradiência (e, por isso, valores próximos ao limiar estabelecido não estão "aproximando significância");
- tem um limiar (normalmente 5% na Linguística) que é arbitrário.

Veja, por exemplo, que o pacote lme4 do R não dá valores de *p* em seus resultados de modelos com efeitos mistos[11], deixando a interpretação para ser feita com base nos coeficientes e nos erros-padrão (que permitem calcular os intervalos de confiança).

Um dos problemas com a prática de se tirar conclusões baseadas exclusivamente em valores de *p* é que, muitas vezes, é possível ter um valor de *p* baixo em um estudo com tamanho de efeito muito pequeno e/ou com baixo poder estatístico[12]. Além disso, como Lima Jr. e Garcia (2021) demonstram, diferentes análises podem levar a resultados categoricamente diferentes se baseados apenas em valores de *p*. Por fim, o desejo por um valor de *p* baixo pode levar, mesmo que inadvertidamente, no caso de pesquisadores iniciantes, à prática antiética de *p-hacking*, em que pequenos ajustes nos dados ou decisões de técnicas diferentes de análise podem ser aplicados para abaixar o valor de *p*.

Por isso, a sugestão é de, inicialmente, utilizar o valor de *p* juntamente com as demais informações de testes e modelos estatísticos para se chegar às inferências, em especial o tamanho do efeito, os intervalos de confiança, o tamanho da amostra, e o poder

[11] Neste caso é por não haver uma única metodologia para o cálculo do valor de *p* nesses modelos, com diferentes aproximações sendo utilizadas por diferentes pacotes, como o lmerTest e o sjPlot.

[12] A chance de detectar o efeito de um teste quando de fato houver um efeito.

estatístico. Com o tempo, conforme consta na próxima seção, a ideia é abandonar por completo o valor de *p* das análises inferenciais.

Cuidado com práticas antiéticas

Existe outra prática antiética adotada por pesquisadores muitas vezes por falta de conhecimento: o *HARKing – Hypothesizing After Results are Known* (Hipotetizar depois que os resultados são conhecidos). Umas das causas das crises de replicabilidade que têm afligido diversas áreas do conhecimento, recentemente a psicologia, é o estabelecimento de hipóteses depois que se veem alguns resultados. Essa prática aumenta as chances de Erro de Tipo I, pois são hipóteses que podem ser criadas com base em resultados espúrios, chegados ao acaso quando se testava outra hipótese.

A melhor maneira de se evitar essa prática é por meio da elaboração cuidadosa e meticulosa das hipóteses antes de se conduzir o estudo, de forma que mudanças nas hipóteses não ocorrerão ao longo da análise. Essas hipóteses podem ser registradas em ferramentas on-line criadas especificamente para isso (como o https://osf.io), e etapas de qualificação de projetos de pesquisa, no caso de estudantes de pós-graduação, também podem servir esse propósito.

Para os que estão prontos para avançar: próximos passos

Para os que já dominam os modelos ensinados pela Livia Oushiro nas Lições 12–15, minha sugestão é que o próximo passo seja no investimento de mais conhecimento sobre modelos de regressão, para que, aos poucos, passem a utilizar exclusivamente modelos em suas análises. Modelos de regressão com efeitos mistos[13] deveriam ser a prática padrão por qualquer pesquisador (MCELREATH, 2020). Portanto, saber rodar e interpretar modelos lineares simples e múltiplos, modelos logísticos, multinomiais, de

[13] Também chamados de modelos de efeitos variáveis, modelos hierárquicos, modelos multinível, modelos aninhados etc.

poisson e ordinal, com a inclusão de *intercepts* e *slopes* aleatórios sempre que necessário, é crucial. Explorar as possibilidades de *contrast coding* bem como de centralização e padronização das variáveis também é desejável. Para isso sugiro três materiais, em ordem crescente de complexidade: Garcia (2021), Sonderegger (2022)[14] e Gelman, Hill e Vehtari (2020).

Como passo seguinte, sugiro investir em inferência bayesiana. A estatística bayesiana tem diversas vantagens, sendo a primeira delas a própria definição de probabilidade. Em análises bayesianas, a probabilidade não é vista como a frequência de ocorrência de um fenômeno, mas como a expectativa de sua ocorrência[15]. Essa aparente pequena diferença fez com que estatísticos frequentistas não conseguissem calcular a probabilidade de um acidente quando usinas nucleares começaram a ser construídas nos EUA, já que não tinham observado nenhum acidente ainda; por isso, a *RAND Corporation* precisou utilizar métodos bayesianos para avaliar a probabilidade de acidentes nucleares antes de acontecer um (MCGRAYNE, 2011).

Outra vantagem é que, enquanto análises frequentistas medem a probabilidade dos dados diante de uma hipótese nula, que é a própria definição de valor de *p*, análises bayesianas calculam a probabilidade das hipóteses de trabalho (hipóteses alternativas) diante dos dados, fazendo a sua interpretação ser mais direta e intuitiva. Com isso, não há valores de *p* em análises bayesianas, e os resultados não são dados como *point estimates*, mas como distribuições de probabilidade das hipóteses de trabalho. Isso retira a decisão dicotômica entre ter ou não ter efeito, e traz nuance, incerteza e gradiência para as inferências, algo desejado quando se busca inferir parâmetros de uma população com base em uma amostra.

Por fim, modelos bayesianos permitem incluir conhecimento prévio da área ou expectativas de valores plausíveis das variáveis, por meio das distribuições *a priori*. Em

[14] A ser publicado pela MIT Press, disponível publicamente no momento da escrita deste texto em https://osf.io/pnumg/

[15] Por isso análises não bayesianas são comumente chamadas de frequentistas.

análises frequentistas, todos os valores dos parâmetros são igualmente prováveis *a priori*. Sendo assim, em uma análise de tempo de reação, um tempo de 10 milissegundos e um de 10 minutos são igualmente prováveis *a priori*; ou, em um estudo fonético, um valor de 100Hz e um de 5000Hz para F1 são igualmente prováveis, e sabemos que não deveriam ser. Em uma análise bayesiana, antes mesmo de se olhar para os dados, podemos informar o modelo sobre uma distribuição de probabilidade de tempos de reação plausíveis ou de valores de F1 plausíveis com base na literatura e em estudos prévios.

As recomendações que deixo para iniciar estudos de estatística bayesiana são um artigo introdutório de Lima Jr. e Garcia (2021), e os livros de McElreath (2020)[16] e Kruschke (2015).

Conclusão

Finalizo reforçando a alegria que sinto em escrever este posfácio, dada a relevância deste livro para a Linguística no Brasil. Tenho certeza que, assim como tem sido para os alunos da Livia e para os meus, este livro será útil para muitos que estão buscando seus primeiros passos para ter o controle das análises quantitativas de seus dados.

Referências

GARCIA, G. D. *Data visualization and analysis in second language research*. New York: Routledge, 2021.

GELMAN, A.; HILL, J.; VEHTARI, A. *Regression and other stories*. Cambridge: Cambridge University Press, 2020.

KRUSCHKE, J. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. 2nd. ed. London: Academic Press, 2015.

[16] O Richard McElreath tem um canal no YouTube com videoaulas que acompanham os capítulos do seu livro.

LARSON-HALL, J. *A guide to doing statistics in second language research using SPSS and R*. New York: Routledge, 2010.

LIMA JR., R. M.; GARCIA, G. D. Diferentes análises estatísticas podem levar a conclusões categoricamente distintas. *Revista da ABRALIN*, v. 20, n. 1, p. 1–19, ago. 2021. Disponível em: https://revista.abralin.org/index.php/abralin/article/view/1790.

MCELREATH, R. *Statistical rethinking: A Bayesian course with examples in R and Stan*. 2nd. ed. Boca Raton: CRC press, 2020.

MCGRAYNE, S. B. *The theory that would not die*. New Haven: Yale University Press, 2011.

SONDEREGGER, M. *Regression modeling for linguistic data*. [S.l.]: Version 1.0. https://osf.io/pnumg/, 2022.