REVISTA DA **ABRALIN**

TUTORIAL

Introdução à estatística bayesiana aplicada à linguística

Guilherme Duarte GARCIA
Ball State University (BSU)

Ronaldo Mangueira LIMA JR
Universidade Federal do Ceará (UFC)

OPEN ACCESS

EDITADO POR
- Raquel Freitag (UFS)

AVALIADO POR
- Livia Oushiro (Unicamp)
- Pablo Arantes (UFSCar)

SOBRE OS AUTORES
- Guilherme Duarte Garcia
Conceptualização, Curadoria de Dados, Análise Formal, Metodologia, Software, Validação, Visualização, Administração do Projeto, Escrita – rascunho original.
- Ronaldo Mangueira Lima Jr
Conceptualização, Curadoria de Dados, Análise Formal, Metodologia, Administração do Projeto, Escrita – análise e edição.

DATAS
- Recebido: 05/09/2021
- Aceito: 08/12/2021
- Publicado: 21/12/2021

COMO CITAR
Garcia, G. D.; Lima Jr, R. M. (2021). Introdução à estatística bayesiana aplicada à linguística. *Revista da Abralin*, v. 20, n. 2, p. 1-24, 2021.

RESUMO

Neste artigo, apresentamos os conceitos fundamentais de uma análise estatística bayesiana e demonstramos como rodar um modelo de regressão utilizando a linguagem R a partir de códigos comentados em detalhe e de pacotes amigáveis que otimizam a implementação de modelos completos. Ao longo do artigo, comparamos estatística bayesiana e estatística frequentista, destacamos as diferentes vantagens apresentadas por uma abordagem bayesiana, que dispensa valores de p e estima distribuições *a posteriori* de efeitos estatisticamente plausíveis com base nos dados modelados. Por fim, demonstramos como rodar um modelo simples e visualizar efeitos de interesse em gráficos intuitivos. Ao longo do artigo, sugerimos leituras adicionais aos interessados neste tipo de análise.

ABSTRACT

In this paper, we introduce the basics of Bayesian data analysis and demonstrate how to run a regression model in R using linguistic data. We provide commented code and employ user-friendly packages that optimize the implementation of full-fledged statistical models. Throughout the paper, we compare Bayesian and Frequentist statistics, highlighting the different advantages of a Bayesian approach, which dispenses with the notion of p-values and instead focuses on parameter estimation using posterior distributions of credible effect sizes given the data. We also show how to run a simple model and how to visualize effects of interest. Finally, we suggest additional readings to those interested in Bayesian analysis more generally.

PALAVRAS-CHAVE
Análise quantitativa de dados. Estatística bayesiana. Modelos de regressão.

KEYWORDS
Quantitative data analysis. Bayesian data analysis. Regression models.

Introdução

Em uma trágica noite de maio de 2009, o voo Air France AF 447 desapareceu entre Rio de Janeiro e Paris com 228 pessoas de 33 nacionalidades a bordo. O acidente, que chocou o mundo, iniciou uma maratona de Brasil e França na busca pelos destroços da aeronave em uma região de difícil acesso, a aproximadamente 1.200 quilômetros de Fortaleza, CE. Depois de dois anos de buscas sem nenhum sucesso, a esperança de encontrar a caixa-preta do voo era mínima—aparentemente, o mistério do voo AF 447 não seria solucionado. Além de ser uma tragédia sem resolução para centenas de famílias, a indústria aérea nunca entenderia o que exatamente levou o Airbus A330 ao fundo do mar naquela noite. Foi então que, em abril de 2011, uma nova busca foi iniciada. Desta vez, o método utilizado envolveria o teorema de Bayes, que utilizaria todos os dados de antes e depois do acidente, e geraria mapas de probabilidades sobre a possível localização dos destroços do Airbus. Em uma semana, o local do acidente foi encontrado, e a investigação sobre as causas do acidente pôde finalmente ser iniciada.

Há diversos exemplos de solução de problemas complexos com análise bayesiana na história, como a máquina de Turing, precursora dos computadores, e utilizada para decifrar códigos secretos alemães durante a Segunda Guerra Mundial, possivelmente salvando as tropas aliadas. Foram análises bayesianas também que permitiram à marinha americana encontrar uma bomba de hidrogênio perdida, assim como submarinos soviéticos. São ainda mais numerosos os casos de solução de problemas em curso, uma vez que a análise bayesiana é central na aprendizagem de máquinas e, consequentemente, na inteligência artificial. Alguns exemplos são os sistemas *antispam* de serviços de e-mail, os sistemas que possibilitam carros autônomos, a previsão de resultados de eleição em tempo real, e a definição de preços por empresas de seguros. McGrayne (2011) narra alguns desses exemplos em detalhe, mas todos têm um aspecto central: resolver problemas que envolvem incerteza à medida em que novos dados são adquiridos e atualizam esse grau de (in)certeza.

Existem basicamente duas grandes escolas de pensamento quando o assunto é análise de dados. De um lado, temos a estatística frequentista,[1] que vê a noção de probabilidade com base na *frequência* de ocorrência de um dado evento a longo prazo. Ou seja, após observarmos diversos dias

[1] Antes de prosseguir, recomendamos ver Lima Jr. e Garcia (2021) para uma breve revisão de conceitos frequentistas utilizando os mesmos dados do presente artigo.

ensolarados, nublados, e chuvosos em uma dada cidade (nossa *amostra*), podemos calcular a probabilidade de cada um dos três cenários para essa cidade. Um frequentista ortodoxo, portanto, entende probabilidades somente a partir de eventos que podem ser repetidos. Além disso, em uma análise frequentista, nossa conclusão se resume à probabilidade dos dados. Afinal, valores de p nos dão a probabilidade de observarmos os dados coletados se partimos do princípio de que a hipótese nula[2] é verdadeira. Em outras palavras, toda vez que você lê um artigo que utiliza valores de p, está diante de uma análise frequentista, entre as mais comuns o teste de qui-quadrado, o teste t, correlações, ANOVAs, e diversos modelos de regressão.

A segunda grande escola de análise de dados é a estatística bayesiana, apoiada no teorema de Bayes, que vê probabilidades como uma combinação de expectativa *a priori* e dados coletados. Diferentemente de uma análise frequentista, uma análise em Bayes utiliza probabilidades tanto para os dados quanto para as hipóteses—algo que foi essencial nas buscas do voo AF 447. Neste caso, embasamos nossas expectativas em nosso conhecimento de área, elaboramos uma hipótese inicial, e incorporamos nosso grau de certeza sobre essa hipótese em nossos modelos analíticos. Diferente de um frequentista, um analista bayesiano não requer eventos repetidos para gerar uma dada probabilidade. Sendo assim, podemos definir probabilidades de um dia ensolarado, nublado, ou chuvoso mesmo que não tenhamos observado dias com essas condições na cidade hipotética mencionada acima—algo impossível em uma análise frequentista, que exige a observação de eventos reais. Poderíamos, por exemplo, analisar essas probabilidades com base na localização da cidade, na estação do ano, no índice de umidade do ar, e no conhecimento prévio que temos sobre a influência dessas variáveis sobre previsões climáticas. Frequentistas, por exemplo, não conseguiram calcular a probabilidade de um acidente quando usinas nucleares começaram a ser construídas, 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).

Você talvez nunca tenha lido um artigo que utilize uma análise bayesiana em linguística. Não se surpreenda: há relativamente poucos estudos que utilizam esse método em linguística quando o comparamos ao método frequentista (IDSARDI, 2006; HAYES ET AL, 2009; GARCIA, 2019)—embora Bayes seja um método relativamente comum em áreas como psicolinguística e cognição (e.g., inúmeros trabalhos de Edward Gibson, Roger Levy, Steven Piantadosi, Joshua Tenenbaum, dentre vários outros). Contudo, a cada dia a estatística bayesiana ocupa o centro da análise de dados em diversos campos. Como veremos abaixo, análises bayesianas oferecem diversas vantagens sobre análises frequentistas, e o poder computacional disponível atualmente permite uma transição frequentista-bayesiana sem grandes problemas.

Nosso objetivo neste artigo é apresentar noções básicas de estatística bayesiana para análise de dados. É preciso salientar que há, também, diversas implementações de modelos bayesianos aplicados à cognição (e.g., CHATER et al, 2006; TENENBAUM et al, 2006; LEE e WAGENMAKERS, 2014).

---

[2] A hipótese nula expressa o contrário da hipótese real de trabalho (hipótese alternativa), e normalmente afirma que não há efeito da variável preditora. Por exemplo, se investigamos uma possível diferença entre dois grupos de falantes, a hipótese nula é a de que não há diferença entre os grupos. A estatística frequentista se baseia fortemente na avaliação da hipótese nula.

Além disso, devemos esclarecer que há diferentes formas de implementar modelos em Bayes (e.g., em uma linguagem de programação como Python ou em um pacote estatístico como Stata). Aqui, contudo, utilizaremos a linguagem R (R CORE TEAM, 2021) no aplicativo RStudio (RSTUDIO TEAM, 2021). Ao fim deste artigo, você conseguirá rodar, interpretar, e reportar um modelo bayesiano simples. Disponibilizamos o script utilizado no link https://osf.io/bvj4w/. Os leitores ainda não familiarizados com o R poderão ignorar esses trechos e, mesmo assim, se beneficiar do conteúdo e das discussões propostas.

Naturalmente, discutiremos apenas o básico neste artigo, e recomendaremos diversos materiais para que você de fato entre no mundo bayesiano. Por fim, este artigo é uma sequência natural de Lima Jr. e Garcia (2021), e tem como público-alvo pessoas que tenham um conhecimento mínimo de análise quantitativa de dados frequentista, em especial de modelos de regressão.

1. O teorema de Bayes

O teorema de Bayes (equação 1) foi proposto pelo reverendo britânico Thomas Bayes em algum momento da década de 1740 (BAYES, 1763). Contudo, foi o matemático francês Pierre-Simon Laplace que, de forma independente e aproximadamente na mesma época, desenvolveu o potencial do teorema—não seria estranho, portanto, se falássemos em teorema de Laplace. Em sua essência, o princípio de Bayes é simples: aprendemos com a experiência, ajustando nossas conclusões proporcionalmente às evidências que encontramos—um conceito tão avançado para o século XVIII que hoje, quase três séculos depois, ainda temos dificuldade em internalizar.

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

EQUAÇÃO 1 - Teorema de Bayes. H = hipótese; E = experimento

De acordo com a equação 1, coletamos dados a partir de um experimento e calculamos a probabilidade de uma dada hipótese (H) com base nesses dados (E), ou seja, P(H|E)—conhecida como *a posteriori*. Perceba que estamos calculando a probabilidade de uma *hipótese* de trabalho diante dos dados observados, e não a probabilidade dos dados diante de uma hipótese nula, como é feito no cálculo do valor de *p*. Além disso, incorporamos ao cálculo nossa expectativa da probabilidade da hipótese *a priori*, P(H), com base em conhecimento e experimentos prévios. No caso de querermos investigar o efeito da estação do ano sobre a quantidade de dias chuvosos em uma cidade, imagine que foram registrados 40 dias chuvosos no outono e 47 no inverno nessa cidade. Um frequentista calcularia a probabilidade de se observar essa quantidade de dias chuvosos em cada estação caso *não* houvesse diferença entre as estações (probabilidade dos dados frente à hipótese nula). Um

bayesiano, por outro lado, calcularia a probabilidade de as estações terem efeito sobre a quantidade de chuva uma vez que foram observados 40 dias de chuva no outono e 47 no inverno (probabilidade da hipótese diante dos dados observados), e incorporaria ao cálculo a expectativa do efeito dessas estações com base em conhecimento prévio sobre o clima nessa cidade ao longo do ano.

Para vermos o teorema em ação, vamos imaginar uma situação bastante simples, em que temos uma hipótese binária. Imaginemos dois grupos de participantes em um estudo qualquer. Ambos os grupos, A e B, possuem falantes monolíngues de português ou de inglês. No grupo A, 80% dos participantes são lusófonos; no grupo B, 40% (tabela 1). Se selecionarmos aleatoriamente um falante e verificarmos que ele é falante de português (E = Por), qual é a probabilidade de este falante pertencer ao grupo A, ou seja, qual o valor de P(A|Por)? Responderemos essa pergunta utilizando o teorema de Bayes.

	A	**B**
Português	80%	40%
Inglês	20%	60%
P(A) = P(B) = 0.5 P(Por|A) = 0.8 P(Ing|A) = 0.2 P(Por|B) = 0.4 P(Ing|B) = 0.6		

TABELA 1 - Dois grupos hipotéticos com falantes monolíngues de português ou inglês

Substituindo o (H) e (E) da equação 1 por (A) e (Por) da tabela 1, temos:

$$P(A|Por) = \frac{P(Por|A)P(A)}{P(Por)}$$

EQUAÇÃO 2 – Teorema da Bayes aplicado ao exemplo de falantes de português ou inglês.

O P(A|Por) é a probabilidade *a posteriori* que queremos descobrir, a probabilidade de o falante aleatório ser do grupo A uma vez que observamos que ele é falante de português. O P(Por|A), primeiro elemento do numerador, é 80%, a probabilidade na primeira célula da tabela 1. O P(A), segundo elemento do numerador, é a probabilidade *a priori* de que um falante aleatório pertença ao grupo A, que é igual à probabilidade *a priori* de ser do grupo B, já que ambos têm o mesmo número de participantes, ou seja, P(A) = P(B) = 50%. O P(Por), no denominador, requer uma breve explicação: neste caso, as duas hipóteses (ser do grupo A ou ser do grupo B) são mutuamente exclusivas, ou seja, um participante pertence ao grupo A *ou* ao grupo B, ninguém pertence a um terceiro grupo, e nenhum participante pertence a ambos os grupos simultaneamente. Consequentemente, podemos reescrever P(Por) como P(A)P(Por|A) + P(B)P(Por|B), a partir da lei de probabilidade total. Como todos esses valores são conhecidos (estão na tabela 1), basta colocá-los na equação e proceder com o cálculo.

No cálculo abaixo, vemos que a probabilidade de um sujeito aleatório que é falante de português vir do grupo A é de aproximadamente 67%. Intuitivamente, faz sentido que o valor esteja acima de 50%, uma vez que o grupo A possui uma maior proporção de falantes de português.

$$P(A|Por) = \frac{P(Por|A)P(A)}{P(A)P(Por|A) + P(B)P(Por|B)}$$
$$P(A|Por) = \frac{0.8 \cdot 0.5}{0.5 \cdot 0.8 + 0.5 \cdot 0.4}$$
$$P(A|Por) = \frac{0.4}{0.6} \approx 0.67$$

EQUAÇÃO 3 - Probabilidade condicional usando o teorema de Bayes

O exemplo acima é bastante simples, mas já nos mostra que o teorema de Bayes combina probabilidades condicionais e probabilidades totais, tendo papel fundamental em lógica indutiva (e.g., HACKING, 2001). Imagine agora que o grupo B tenha mais participantes do que o grupo A. Isso afetaria os valores P(A) e P(B), que não seriam mais idênticos. Por exemplo, se o grupo A tiver 50 participantes e o grupo B tiver 75 participantes, a probabilidade de um participante aleatório vir do grupo A vai de 0.5 para 0.4, o que afetaria nosso cálculo. No mundo real, nosso conhecimento de área e pesquisas anteriores podem nos informar sobre o que esperar de um dado experimento. Usando Bayes, podemos incorporar esse conhecimento ao definirmos a distribuição *a priori*, da mesma forma que podemos ajustar P(A) e P(B) no exemplo hipotético acima. Você pode interagir com o teorema de Bayes visitando a página https://guilhermegarcia.github.io/resources e, em seguida, clicando em "Bayesian statistics".

Apesar de instrutivos, exemplos simples raramente nos ajudam diretamente em aplicações reais. Quando desejamos descobrir o tamanho do efeito de um fator, nossas hipóteses não são binárias. Lidamos, nesses casos, com um *continuum* de valores plausíveis. Além disso, modelos realistas geralmente têm diversos parâmetros (variáveis preditoras), o que torna o cálculo impraticável (e quase sempre impossível). Por exemplo, imagine que queiramos examinar a probabilidade de 1.000 valores plausíveis para um dado efeito em nosso experimento. Se tivermos 5 variáveis em nosso modelo, teremos uma distribuição conjunta de 1.000^5, um valor complicado demais para nossos computadores. Essa é a principal razão técnica por que métodos frequentistas dominaram a análise de dados no século XX: simplesmente não havia poder computacional suficiente para que conseguíssemos utilizar modelos bayesianos de forma realista (KRUSHKE, 2013; MCELREATH, 2020).

1.1 Amostragem do *a posteriori*

Felizmente, em vez de calcularmos P(H|E) analiticamente, utilizamos algoritmos que compilam amostras *a posteriori*. O princípio de amostragem é essencial para tornar a análise de dados em Bayes

possível: em vez de lidarmos com a matemática, lidamos com amostras. Um método bastante utilizado de amostragem é o Markov Chain Monte Carlo, ou MCMC. Markov Chain é um método estocástico que consiste em uma cadeia onde a probabilidade de um evento n depende apenas da probabilidade do evento $n - 1$. O modelo que utilizaremos neste artigo é baseado em um método mais recente chamado de Hamiltonian Monte Carlo, ou HMC.[3]

Os detalhes desses algoritmos de amostragem são bastante complexos e fogem do escopo deste artigo[4]. Contudo, para entendermos intuitivamente o funcionamento do algoritmo MCMC, por exemplo, podemos pensar em uma "caminhada" aleatória no espaço de possíveis valores de parâmetros plausíveis. Vamos entender o que isso significa com um exemplo simples.

Imagine novamente os grupos A e B do exemplo anterior. Coletamos alguns dados, como notas em um teste, e queremos saber se os grupos são diferentes. Ou seja, o parâmetro que mais nos interessa é a diferença das médias das notas das populações A e B, ou seja $\mu_A - \mu_B$. Naturalmente, essa diferença real (desconhecida) pode ser qualquer número racional $\mathbb{Q}$. O algoritmo escolhe um valor candidato e calcula a probabilidade desse valor considerando-se os dados do experimento—seguindo o teorema de Bayes discutido acima. Em seguida, o algoritmo escolhe outro valor e faz o mesmo cálculo. A decisão de ir do primeiro para o segundo valor é a chave do algoritmo. Ao final de uma longa caminhada, em que um alto número de valores é considerado, teremos uma cadeia de valores plausíveis e podemos criar um histograma para verificar a distribuição dos valores mais confiáveis a partir dos dados que temos, ou seja, P(H|E), o nosso *a posteriori*. Quanto mais provável for um dado valor, mais vezes o algoritmo visitará esse valor (o mesmo pode ser dito sobre valores bastante próximos ao valor em questão). Consequentemente, os valores mais frequentes nessa caminhada aleatória são os valores mais prováveis para $\mu_A - \mu_B$, nosso efeito de interesse. A figura 1 ilustra essa distribuição. Como a distribuição *a posteriori* neste caso segue uma distribuição normal, podemos utilizar a média dessa distribuição como "valor simbólico representativo" do efeito em questão, mas sem nenhum impedimento para que se utilize a mediana ou a moda, por exemplo. Neste caso, o valor mais provável para a diferença de médias entre A e B é aproximadamente 7—linha pontilhada na figura. Note que esse valor está bastante longe de zero. De fato, a distribuição inteira está acima de zero, o que nos mostra que há uma diferença positiva entre os grupos neste caso: o grupo A tem uma média superior à do grupo B, portanto.

---

[3] A principal diferença prática entre HMC e MCMC é o método usado para explorar o espaço de parâmetros: enquanto MCMC utiliza distribuições de probabilidades, HMC utiliza dinâmica hamiltoniana, um método que reduz a correlação de valores propostos e que, portanto, torna o algoritmo mais eficiente na busca de valores plausíveis para parâmetros de interesse.

[4] Veja o capítulo 7 de Kruschke (2015) ou o capítulo 9 de McElreath (2020) para explicações didáticas sobre o funcionamento dos principais algoritmos.

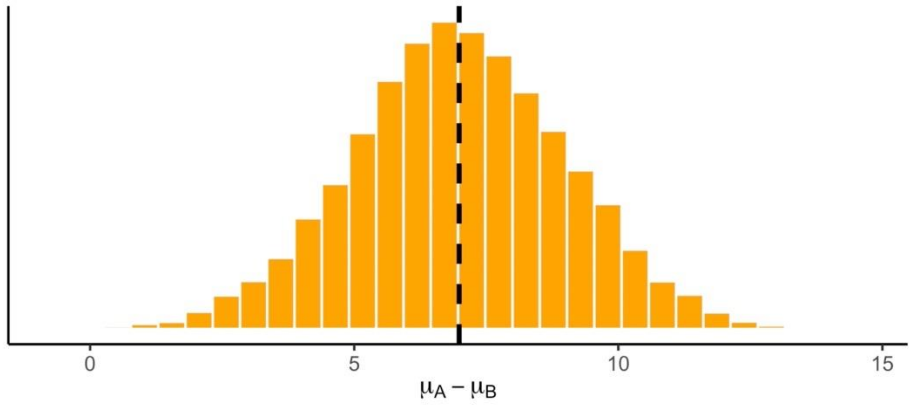

FIGURA 1 – Distribuição *a posteriori* para $\mu_A - \mu_B$

Em uma análise frequentista, poderíamos rodar um teste *t* e encontraríamos um valor de *p* abaixo de 0,05. Lembre-se, no entanto, de que valores de *p* nos dão a probabilidade dos dados a partir de uma hipótese nula. Nosso *a posteriori*, por outro lado, nos dá a probabilidade de um efeito com base nos dados. Consequentemente, não veremos valores de *p* em uma análise bayesiana. Além disso, no resultado de nosso teste *t*, teríamos *um único valor* para a diferença das duas amostras, uma estimativa pontual (*point estimate*). Uma análise em Bayes, por outro lado, nos proporciona uma distribuição de efeitos plausíveis (de diferenças entre A e B neste caso). A diferença real entre A e B pode ser qualquer valor da distribuição *a posteriori*, sendo que os valores mais frequentes (neste caso, aqueles mais ao centro da distribuição da figura 1, por exemplo) são os mais prováveis. O resultado da análise bayesiana adiciona uma camada de incerteza ao resultado, algo desejável ao se inferir parâmetros desconhecidos da população com base em uma amostra.

Por fim, um teste *t* geraria um intervalo de confiança. Na figura 1, poderíamos facilmente gerar um intervalo de credibilidade cuja interpretação seria bastante intuitiva: valores mais próximos do centro de tal intervalo são mais prováveis do que valores às margens desse intervalo—algo que não podemos concluir a partir de intervalos de confiança tradicionais, uma vez que não são distribuições.[5] Ou seja, nossa distribuição *a posteriori* nos proporciona um intervalo que consiste em uma distribuição de probabilidades.

Na comparação acima, utilizamos testes *t* como referência. Naturalmente, podemos efetuar um procedimento equivalente utilizando Bayes (e.g., KRUSHKE, 2013). Ao longo deste artigo, não utilizaremos testes *t*, já que há métodos muito superiores para analisarmos respostas contínuas. Nosso exemplo será uma regressão linear, que nos ajudará a entender diferenças importantes sobre como rodamos modelos e interpretamos resultados em estatística bayesiana.

[5] Intervalos de confiança apenas demarcam dois pontos limítrofes, e não são, portanto, uma distribuição em princípio. Parece intuitivo concluir que valores que estão mais próximos do centro de um intervalo de confiança são "mais robustos", ou mais "confiáveis". Essa conclusão, contudo, é incorreta. Para mais informações, sugerimos a leitura de Kruschke (2015, pp. 323–324).

2. Por que migrar para uma análise bayesiana?

Existem diversas razões por que deveríamos migrar de uma análise frequentista para uma análise bayesiana. Primeiramente, como vimos acima, o teorema de Bayes nos dá probabilidades sobre hipóteses, e não sobre dados. Ou seja, temos P(H|E) e não P(E|H). A probabilidade de um efeito é quase sempre mais relevante do que a probabilidade de um dado. Afinal, coletamos dados com o intuito de descobrirmos um efeito. Conceitualmente, portanto, Bayes nos proporciona um resultado mais relevante.

Uma consequência importantíssima de acessarmos P(H|E) e não P(E|H) é o abandono de valores de p. Há uma literatura bastante ampla sobre os problemas inerentes ao foco em significância estatística (e.g., NUZZO, 2014; ver também o capítulo 11 de Kruschke, 2015). Lima Jr. e Garcia (2021), por exemplo, demonstram como as intenções de um pesquisador podem afetar a significância estatística quando temos comparações múltiplas. De fato, um grande problema conceitual sobre valores de p é sua natureza simplista e binária: a ideia de que efeitos "existem ou não existem" é bastante ingênua quando temos acesso limitado a dados e quando nosso desenho experimental está longe da perfeição. Uma análise em Bayes nos fornece mais nuance, e é, portanto, muito mais realista e apropriada às complexidades envolvidas em análise de dados.

Em segundo lugar, uma análise bayesiana nos fornece uma fotografia muito mais completa sobre os efeitos de interesse, uma vez que temos acesso a uma distribuição *a posteriori* de efeitos plausíveis, como mencionado acima. Ou seja, em vez de termos apenas um valor (estimativa pontual) para um dado efeito, temos uma distribuição inteira. Uma consequência adicional dessa distribuição é a facilidade com que definimos e interpretamos intervalos de credibilidade—algo que veremos em mais detalhe abaixo. Essa facilidade de interpretação é uma vantagem adicional de modelos bayesianos, embora possuam uma implementação computacionalmente muito mais complexa.

Uma terceira grande vantagem de modelos bayesianos é o alto poder de personalização oferecido. Se temos dados que seguem uma distribuição não normal, podemos configurar nosso modelo com diferentes distribuições (*t*, por exemplo). Naturalmente, customizar um modelo exige um conhecimento relativamente avançado de estatística. Podemos traçar um paralelo entre modelos bayesianos e fotografia: uma câmera profissional quase sempre oferece os melhores resultados em seu modo manual, que exige maior conhecimento por parte de quem utiliza a câmera. Contudo, uma câmera profissional também possui um modo automático e fácil, que já proporcionará bons resultados na maioria das situações. Da mesma forma, podemos utilizar modelos em Bayes no seu "modo automático". Embora deixemos de utilizar todo o potencial desses modelos, nossos resultados ainda se beneficiarão das vantagens discutidas acima.

Em quarto lugar, modelos bayesianos são muito mais robustos quando o assunto é convergência, algo que pode ser um problema em modelos tradicionais frequentistas, especialmente quando possuem uma estrutura mais complexa—comum em modelos de efeitos mistos. Por mais complexo que um modelo bayesiano seja, se sua especificação for adequada, ele convergirá—basta esperarmos a compilação e amostragem terminarem.

Em quinto lugar, como mencionamos acima, modelos bayesianos permitem que incorporemos nosso conhecimento de área em nossas análises estatísticas a partir de distribuições *a priori*—lembre-se do exemplo do voo AF 447. A possibilidade de unirmos teoria e análise estatística abre um leque riquíssimo de estudos. Por exemplo, em aquisição de segunda língua, sabemos que aprendizes não começam do zero, e que dependem em parte de suas gramáticas nativas. Um modelo tradicional é incapaz de incorporar esse fato em sua análise. Em Garcia (2020), por exemplo, distribuições *a priori* informativas são utilizadas na simulação de diferentes premissas teóricas na transferência de padrões fonológicos entre primeira e segunda línguas. Naturalmente, você pode escolher não utilizar uma distribuição *a priori* informada. O modelo simplesmente utilizará um *a priori* bastante vago, e seus resultados serão relativamente similares àqueles que você teria em uma análise frequentista equivalente—é isso que mostraremos abaixo. Exemplos com diferentes distribuições *a priori* aplicados a dados linguísticos podem ser vistos em Garcia (2020; 2021) e Arantes e Lima Jr. (2021).

Por fim, modelos bayesianos têm a vantagem de lidar bem com dados ausentes ou com dados desbalanceados, quando há quantidade diferente de dados para participantes ou grupos. Infelizmente, um tratamento comum para casos de dados ausentes ou desbalanceados acaba sendo a exclusão parcial ou total dos dados de certos participantes (BARKAOUI, 2014), algo que não precisa ser feito, já que esses dados podem trazer informações importantes para o modelo. McElreath (2020) dedica um capítulo inteiro (capítulo 15) sobre como lidar com dados ausentes de maneira bayesiana.

Reconhecemos que as vantagens acima têm um custo. É preciso entender as desvantagens envolvidas na utilização de modelos em Bayes. Por exemplo, há uma curva acentuada de aprendizado, uma vez que há aspectos conceituais e técnicos que são distintos de análises frequentistas. Além disso, estimar efeitos usando amostragens do *a posteriori* é um processo computacionalmente exigente, o que demandará mais tempo de processamento, especialmente para modelos de efeitos mistos mais complexos—não é raro que um modelo em Bayes leve uma hora ou mais para rodar.

Pesquisadores que utilizarem métodos bayesianos em linguística também precisarão lidar com pareceristas que, muitas vezes, não estarão familiarizados com o método. Em muitos casos, haverá desconfiança sobre distribuições *a priori* informativas e sobre a ausência de valores de *p*. O argumento costuma ser o seguinte: se podemos escolher a distribuição *a priori*, e se sabemos que nossos resultados podem consequentemente ser afetados, uma escolha informada de distribuição *a priori* pode enviesar nossos resultados a favor do argumento feito pelo estudo em questão.

A crítica acima está embasada em algo real: de fato, se escolhermos uma distribuição alinhada com os resultados que desejamos, e se utilizarmos um desvio-padrão minúsculo para essa distribuição, certamente nossas conclusões serão basicamente a imagem de nossas expectativas. Essa relação faz sentido, e é bastante conhecida: se acreditamos cegamente em algo, nenhuma evidência nos fará mudar de ideia. Ou seja, se o nosso *a priori* for absolutamente intransigente, nossos dados serão virtualmente irrelevantes: nosso *a posteriori* simplesmente imitará nosso *a priori*. O problema, contudo, é que nenhum *a priori* é escolhido com base em nossa própria vontade (GELMAN, 2008), e essa aparente subjetividade sem critérios é um exemplo da famosa falácia do espantalho.

Além disso, modelos frequentistas também trazem expectativas *a priori*. Neles, entretanto, todos os valores dos parâmetros são igualmente prováveis a priori. Em uma análise da diferença de altura entre homens e mulheres, por exemplo, um modelo frequentista parte do *a priori* de que uma diferença de 3 km é tão provável quanto uma diferença de 10 centímetros ou de 1 milímetro (BÜRKNER, 2018a). Em dados linguísticos, modelos frequentistas iniciam suas análises com a expectativa de que tempos de reação de identificação lexical de 500 milissegundos e de 3 minutos são igualmente prováveis; ou que diferenças de 1, 30 ou 90 pontos entre dois grupos em um exame de proficiência são igualmente prováveis.

Em ciência, todo e qualquer estudo está embasado em estudos anteriores—essa cumulatividade de conhecimento está no cerne do fazer científico. A possibilidade de incorporarmos conhecimento de área em nossos modelos a partir de distribuições *a priori* informadas é, portanto, uma característica não apenas desejável, mas essencial a qualquer estudo. Evidentemente, a escolha de distribuições *a priori* precisa ser criteriosa e estar embasada no corpo de conhecimento de área a partir de estudos anteriores, que alimentarão, assim, estudos atuais e futuros.

3. Demonstração em R

3.1. Pacotes

Para rodarmos modelos bayesianos em R utilizaremos indiretamente uma linguagem chamada Stan, que foi criada para a implementação de modelos bayesianos (CARPENTER et al., 2017). Faremos isso a partir de um pacote que "traduz" para Stan as especificações de modelos já familiares em R. Antes de prosseguirmos, portanto, você precisará instalar o pacote `brms` (BÜRKNER, 2018b), que, por sua vez, instalará alguns pacotes adicionais necessários[6]. Também será necessário instalar o pacote `languageR` (BAAYEN, 2007), que contém os dados `danish`, que utilizaremos abaixo—os mesmos dados utilizados em Lima Jr. e Garcia (2021). Por fim, também usaremos o pacote `tidyverse` (WICKHAM et al., 2019), que você já deve ter instalado se utiliza R em suas análises de dados.

[6] O `brms` roda nos bastidores o `Rstan`, que é uma interface do R para a linguagem Stan. O Stan, por sua vez, é construído na linguagem de programação C ++. Sendo assim, é preciso primeiramente configurar o computador para que possa usar C ++. Isso deve ser feito apenas uma vez, e o procedimento depende do sistema operacional. As instruções podem ser encontradas em https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started.

3.2. Exemplo de modelo bayesiano em R

3.2.1. Importando, simplificando, e visualizando os dados

Primeiramente, carregaremos os pacotes e os dados mencionados acima (linhas 1–3 do quadro 1). Os dados em questão vêm de uma tarefa de decisão lexical do dinamarquês. Nos dados, temos diferentes sufixos. Em seguida, a fim de tornar nossa demonstração comparável a Lima Jr. e Garcia (2021), simplificaremos o número de variáveis (linhas 4–6) e filtraremos nossos dados para que tenhamos apenas cinco sufixos: "bar", "ende", "ede", "ere", e "lig" (linhas 7–8).

```
1| library(tidyverse)
2| library(languageR)
3| library(brms) # instale via install.packages("brms") primeiro

3| data(danish)
4| dan = danish %>%
5|   select(Subject, Word, Affix, LogRT) %>%
6|   as_tibble()

7| dan = dan %>%
8|   filter(Affix %in% c("bar", "ende", "ede", "ere", "lig")) %>% droplevels()
```

QUADRO 1 - Linhas de comando para carregar pacotes, carregar e filtrar os dados a serem analisados.

Em nossa análise, queremos descobrir se diferentes afixos afetam o tempo de reação dos participantes. Mais importante do que definirmos se há um efeito ou não é quantificarmos o *tamanho do efeito* de cada sufixo relativo a um nível de referência. Como sabemos, o nível de referência de um dado fator em uma regressão linear é escolhido alfabeticamente ao rodarmos um modelo—esse nível pode ser facilmente alterado, mas não entraremos nessa discussão aqui. Aqui, portanto, todos os sufixos serão comparados a "bar". Naturalmente, poderíamos alterar esse nível, mas nenhuma escolha será mais justificável ou menos arbitrária do que "bar" para o exemplo de análise a seguir.

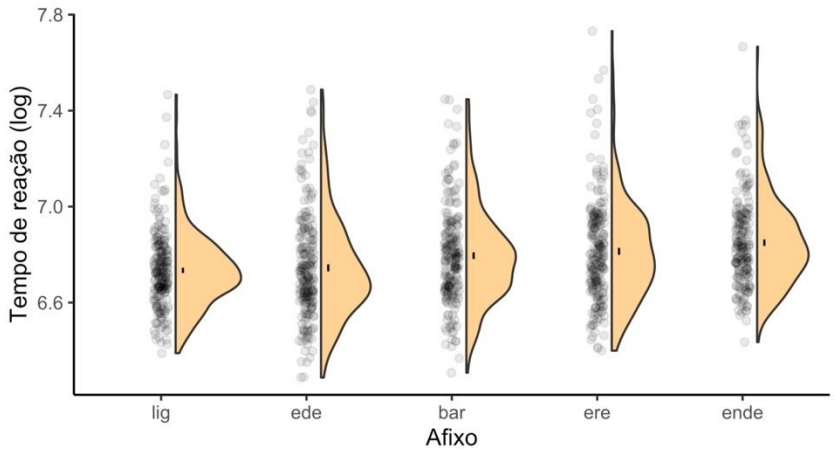

FIGURA 2 – Tempo de reação (log) para palavras com diferentes sufixos. A barra de erro em preto no interior de cada distribuição representa o erro-padrão.

A figura 2[7] apresenta três informações sobre os sufixos com base nos tempos de reação dos participantes (em escala logarítmica para reduzir a assimetria das caudas em uma típica distribuição de tempos de reação). Temos a dispersão dos dados (círculos cinzas semitransparentes), a distribuição dos dados (em laranja), e o erro-padrão de cada sufixo (pequena barra preta no interior de cada distribuição). Esse tipo de visualização permite uma checagem visual rápida sobre o grau de normalidade dos dados e, é claro, sobre qualquer possível efeito (ou diferença) dos sufixos em questão. Com base na inspeção visual dos erros-padrão, podemos prever um efeito de sufixo, dada a distância entre "lig" e "ende" – o gráfico ordena os sufixos em ordem crescente de média de tempo de reação com o uso do argumento `fct_reorder`.

3.2.2. Rodando uma regressão linear bayesiana

Teoricamente, para rodarmos uma regressão linear em Stan, precisaríamos aprender a linguagem Stan, que possui uma sintaxe otimizada para modelos estatísticos—veja exemplo no apêndice, quadro A1. Contudo, graças a pacotes como `brms`, não precisamos definir manualmente nossos modelos com Stan. Em vez de escrevermos um modelo como no quadro A1, simplesmente usaremos a conhecida sintaxe de regressões em R: `y ~ x`. Ou seja, do ponto de vista técnico, você precisa de muito pouco para conseguir rodar um modelo em Bayes nos dias de hoje—desde que você já esteja familiarizado com regressões e R.

O modelo que rodaremos, diferentemente do exemplo do quadro A1, tem um preditor categórico. Ou seja, não estamos diante de $y_n = \alpha_n + \beta_n x + \epsilon_n$. Estamos diante de $y_n = \alpha_n + \beta_n x + \beta_n x +$

[7] O gráfico em questão é conhecido como "*half violin plot*", uma vez que representa a metade de um *violin plot*. Consulte o script que acompanha este artigo para ter acesso ao código que gerou a figura.

$\beta_n x + \beta_n x + \epsilon_n$, em que cada $\beta_n x$ representa um de nossos sufixos e α_n representa nosso nível de referência, "bar". O princípio, é claro, é exatamente o mesmo, embora x aqui possa representar apenas 0 ou 1.

```
25| fit = brm(LogRT ~ Affix, data = dan, family = "Gaussian",
26|         cores = 4,
27|         chains = 4,
28|         save_model = "fit.stan")

29| fit # para visualizarmos o output
```

QUADRO 2 – Regressão linear com Stan via **brms**.

No quadro 2, rodamos nossa regressão com a função brm(). Utilizamos essa função para rodar *qualquer modelo* com o pacote em questão (regressões linear, logística, ordinal etc.), inclusive versões com efeitos mistos utilizando a mesma sintaxe que já usamos em modelos frequentistas. O que especifica o tipo de modelo que estamos rodando está no argumento family, que em uma regressão linear tradicional será definido como "Gaussian", e em uma regressão robusta será definido como "student" (para a distribuição *t*)[8]. Perceba que o "coração" do modelo é apenas LogRT ~ Affix, que é bastante familiar. Em seguida, estipulamos o número de núcleos que desejamos utilizar com o argumento cores: como dissemos acima, modelos em Bayes são computacionalmente intensos, e fazer uso de múltiplos núcleos ajuda consideravelmente a acelerar o processo de amostragem. A maioria dos computadores atuais tem pelo menos 4 núcleos, e rodar um script com 4 núcleos em um computador com menos núcleos não causará nenhum problema. Alternativamente, pode-se substituir a linha 27 por mc.cores = parallel::detectCores()-1,[9] para garantir que apenas um núcleo não seja utilizado.

Em seguida, definimos quantas cadeias desejamos (*default* = 4). Lembre-se de que nosso modelo está realizando uma "caminhada" aleatória no espaço de parâmetros mais plausíveis. Como saber se uma caminhada acabou chegando aonde deveria? Simples: realizamos múltiplas caminhadas simultaneamente. Se elas atingirem aproximadamente o mesmo espaço, nosso modelo convergiu com sucesso e temos estimativas confiáveis. Por essa razão, precisamos de, pelo menos, *duas* cadeias. Aqui, utilizamos quatro, o valor padrão (ou seja, cada núcleo será responsável por uma cadeia, otimizando o processo como um todo). Por fim, salvamos nosso modelo em um arquivo Stan. Você pode abrir o arquivo mais tarde no próprio RStudio para verificar o grau de complexidade da especificação via Stan, traduzida pelo pacote brms. Existem diversos outros argumentos que podem ser passados à função brm(), especialmente priors, onde podemos especificar nossas expectativas sobre parâmetros de interesse, mas naturalmente não teremos espaço neste artigo para explorarmos todos.

---

[8] Para uma regressão logística, será "bernoulli" ou "binomial", para regressão ordinal "cumulative", e para regressão multinomial "multinomial", por exemplo. A documentação do pacote apresenta diversas outras famílias possíveis.

[9] É necessário primeiramente instalar o pacote parallel com install.packages("parallel").

Após rodarmos as linhas 25–28, a primeira grande diferença perceptível será o tempo necessário até que o modelo termine de compilar e de amostrar o *a posteriori*. Regressões frequentistas sem efeitos mistos rodam instantaneamente em praticamente qualquer computador nos dias de hoje. Uma regressão simples em Bayes, contudo, não será instantânea, e poderá levar até alguns minutos dependendo do seu computador.

Quando o modelo estiver concluído, podemos simplesmente rodar fit para termos acesso ao output (i.e., não é necessário utilizar a função summary()). O quadro 4 traz o output completo do nosso modelo. Em "Samples", vemos que o modelo possui quatro cadeias, cada uma com 2.000 iterações, sendo 1.000 delas iterações de *warmup*. Pense nas iterações como "passos" que cada cadeia dará na caminhada aleatória em busca dos valores mais plausíveis para nossos parâmetros. Ou seja, desejamos 2.000 amostras do *a posteriori* de cada cadeia. O modelo automaticamente inclui *warm-up* e *thinning*, e apenas amostras tiradas *após* ambos os processos são consideradas. Cada cadeia precisa de um certo tempo até que se aproxime dos valores de parâmetros mais estáveis (i.e., plausíveis)—esse tempo é chamado de *warm-up*. Sendo assim, é recomendável que não sejam consideradas as amostras iniciais das cadeias (neste caso, as primeiras 1.000 são ignoradas). Como temos quatro cadeias, teremos um total de 4.000 amostras válidas (*post warm-up samples*). Além disso, a amostra n+1 é tipicamente correlacionada com a amostra n. Para reduzir esse grau de autocorrelação, podemos "pular" um número x de amostras. Esse processo é chamado de *thinning*. Para o presente exemplo, utilizamos os valores *default* para ambos os processos.

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: LogRT ~ Affix
   Data: dan (Number of observations: 1040)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
       total post-warmup samples = 4000

Population-Level Effects:
             Estimate  Est.Error  l-95% CI  u-95% CI  Rhat  Bulk_ESS  Tail_ESS
Intercept        6.80       0.01      6.77      6.82  1.00      2102      2420
Affixede        -0.05       0.02     -0.09     -0.01  1.00      2347      3061
Affixende        0.05       0.02      0.01      0.09  1.00      2776      3300
Affixere         0.02       0.02     -0.02      0.06  1.00      2476      2767
Affixlig        -0.06       0.02     -0.10     -0.02  1.00      2557      2532

Family Specific Parameters:
          Estimate  Est.Error  l-95% CI  u-95% CI  Rhat  Bulk_ESS  Tail_ESS
sigma         0.20       0.00      0.19      0.21  1.00      4052      2536

Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

QUADRO 3 – Output de regressão linear em Bayes.

O principal componente de nosso output está em *Population-Level Effects*, que são relativamente familiares a quem já rodou regressões frequentistas—a coluna *Est.error* representa o desvio-padrão do *a posteriori*. A diferença, naturalmente, é que não temos valores de p, mas temos $\hat{R}$ ("Rhat" no *output*), também conhecido como diagnóstico de Gelman-Rubin, e duas colunas para ESS (*Effective Sample Size*). Idealmente, precisamos de um $\hat{R} = 1$, que indica convergência do modelo–valores acima de 1 indicam a não convergência. As colunas para ESS simplesmente nos mostram quantas amostras reais (pós *warmup/thinning*) o modelo conseguiu extrair do *a posteriori* após levarmos em conta que algumas amostras estão autocorrelacionadas e, portanto, são menos/pouco informativas.[10] Não há um número mágico que devemos almejar para ESS: quanto maior, melhor, pois teremos mais amostras. Diferentes autores recomendarão diferentes valores, dependendo dos dados e do modelo que temos em mãos, mas é seguro dizer que qualquer valor acima de 1.000 é um "bom valor"—independente do número de iterações ou cadeias utilizado no modelo.

Como em regressões frequentistas, nosso modelo estima os coeficientes de interesse assim como seus desvios-padrão. Ao contrário de modelos frequentistas, que informam intervalos de 95% de confiança, o nosso modelo informa intervalos de 95% de credibilidade, em l-95% CI (*lower 95% credible interval*) e u-95% CI (*upper 95% credible interval*). Esse intervalo contém os valores mais prováveis para o parâmetro em questão, e é comumente chamado de *highest density interval*, ou HDI. Por exemplo, o nosso *intercept* de valor $\hat{\beta} = 6.80$ indica que 6,8 é o valor mais provável do tempo de reação (em escala log(ms)) para palavras com "bar", sendo que o valor real desse parâmetro tem 95% de probabilidade de estar entre 6,77 (l-95% CI) e 6,82 (u-95% CI), mas com maior probabilidade dos valores mais próximos a 6,8. Semelhantemente, o afixo "ede" tem um tempo de reação menor que o de "bar", com 95% de probabilidade de ser entre 0,01 e 0,09 menor, com valores mais próximos a -0.05 sendo os mais prováveis. Interpretaremos os resultados em maior detalhe a seguir. Por fim, em *Family Specific Parameters*, temos a estimativa de *sigma*, isto é, o desvio-padrão do *a posteriori* de nossa variável resposta (tempo de reação).

3.3. Interpretando e reportando resultados

A melhor maneira de olharmos para os resultados de um modelo em Bayes é visualizarmos nossos *a posteriori*. Antes de fazermos isso, contudo, você deve estar se perguntando como é possível definir se um resultado é estatisticamente plausível ou não (não usamos a palavra "significativo" aqui, uma vez que não temos valores de p). Uma maneira simplista e categórica de concluirmos que efeitos são reais envolve verificar se nosso intervalo de credibilidade inclui 0. Por exemplo, o resultado de "ede" é $\hat{\beta} = 6.80, 95\%\ HDI = [-0.09, -0.01]$. Aqui, $\hat{\beta}$ representa a *média* do *a posteriori* para esse sufixo (relativo ao sufixo "bar"), e seu HDI *não inclui* zero. Portanto, concluímos que "ede" é estatisticamente

[10] A autocorrelação do *a posteriori* avalia a correlação entre valores de amostragem do *a posteriori*. Portanto, não se trata de uma medida de colinearidade entre variáveis.

diferente de "bar" no que diz respeito aos tempos de reação que elicita nos participantes. É importante lembrarmos que 95% é um valor *arbitrário*, e qualquer outro valor seria igualmente justificável para um intervalo. McElreath (2020), por exemplo, utiliza em seu livro, em seus estudos e em seu rethinking package um intervalo de 89%. A justificativa é ser um número primo, motivo tão arbitrário quanto os 95%. A função summary(..., prob = 0.95, ...) permite alterar o intervalo padrão. Também é possível alterar esse valor nas diferentes técnicas de visualização de distribuições *a posteriori*, assim como nas diferentes formas de avaliar as amostras do *a posteriori*.

A interpretação acima não está errada, mas é simplista demais, e tenta trazer à análise bayesiana uma maneira frequentista de interpretar resultados: queremos uma resposta categórica. O problema é que, diferentemente de intervalos de confiança, HDIs são distribuições. Ou seja, um zero contido em uma cauda da distribuição é bastante diferente de um zero no centro dessa mesma distribuição. É preciso, portanto, averiguar *onde* o valor zero está na distribuição. Assim, podemos saber o quão plausível é supor um efeito nulo considerando os dados que temos em mãos. Tudo isso reforça a necessidade de *visualizarmos* nossos resultados.

Primeiramente, verificaremos a convergência do modelo em questão a partir de um gráfico de traços (também conhecido como gráfico de lagartas). Queremos averiguar se todas as quatro cadeias "concordam" ao chegarem no valor mais plausível para cada um de nossos quatro parâmetros, neste caso os sufixos (além do *intercept*). No eixo x da figura 3, vemos as 1.000 amostras úteis (lembre-se de que usamos as primeiras das 2.000 amostras como *warm-up*). No eixo y, vemos o valor dos nossos coeficientes ($\hat{\beta}$). Como podemos ver, todas as cadeias se sobrepõem em torno do mesmo espaço de valores, o que nos mostra que houve convergência de cadeias. Este tipo de gráfico é geralmente omitido de uma publicação real, servindo principalmente como diagnóstico para o analista.

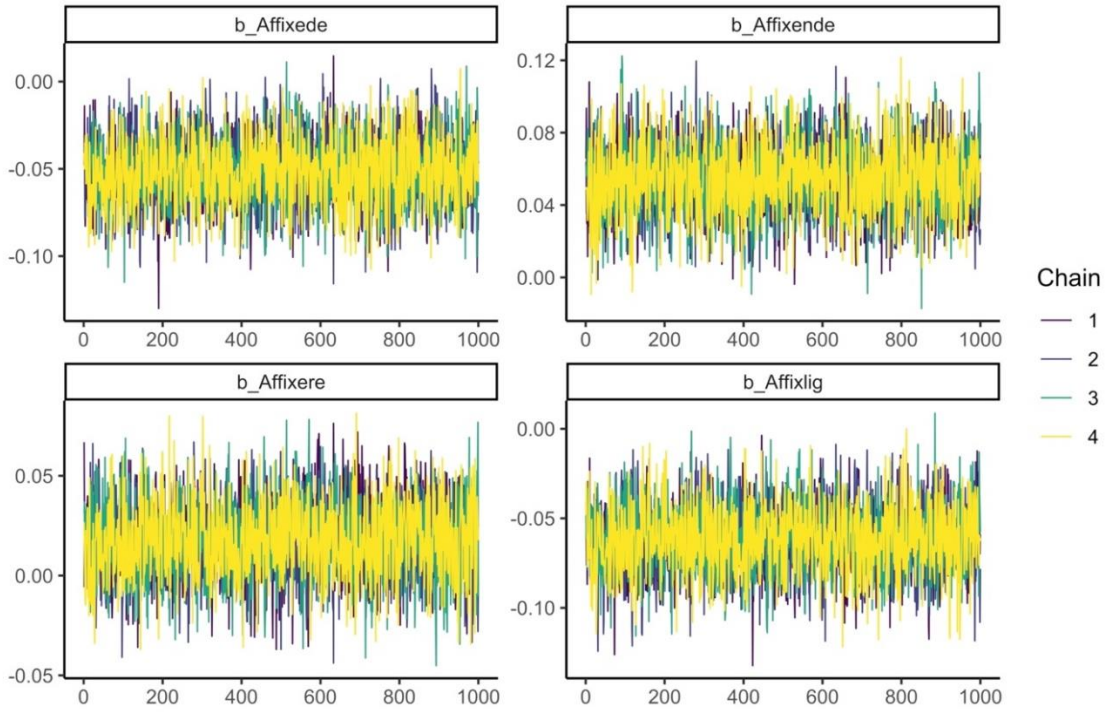

FIGURA 3 – Gráfico de traços para verificar a convergência de cadeias.

A figura 4 apresenta dois gráficos com cadeias que não convergem. Veja que há espaços visitados apenas por uma ou duas cadeias, e vários espaços não visitados por nenhuma. Casos assim exigem alguma modificação na especificação do modelo, como um número maior de iterações, um número maior de *warm-up* ou a especificação de *priors* minimamente informativos.

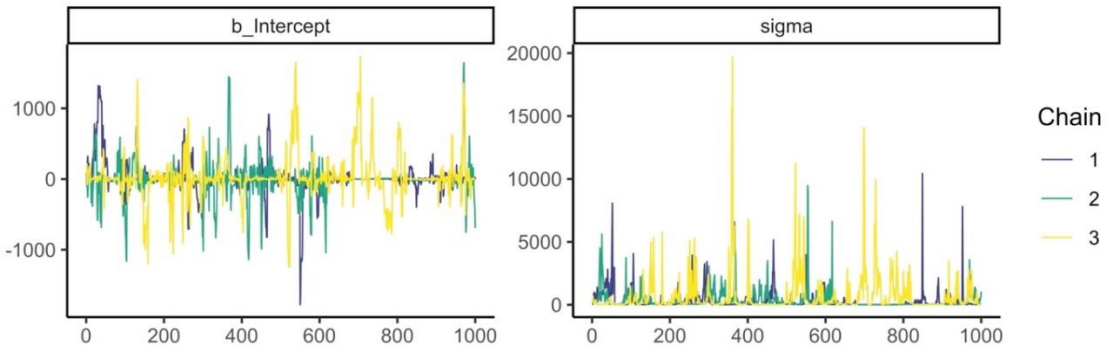

FIGURA 4 – Gráfico de traços com cadeias que não convergem.
Fonte: Adaptado de McElreath (2020).

Em seguida, visualizaremos o principal gráfico de um modelo, em que observamos as distribuições *a posteriori* para cada um de nossos parâmetros (figura 5). O gráfico é bastante intuitivo, uma

vez que delimita zero e compila tanto os *a posteriori* de interesse quanto seus HDIs (95%)—área cinza clara de cada distribuição. Como podemos ver, apenas um HDI inclui zero (sufixo "ere"). Perceba que, para esse sufixo, zero é um valor relativamente provável, já que está aproximadamente entre a média e o limite inferior do HDI. Com isso, não podemos afirmar que "ere" tem um efeito estatisticamente real relativo a "bar" nos tempos de reação no estudo em questão. As figuras 3 e 5 foram elaboradas com o código do quadro 4, em que ajustamos o tema e o esquema de cores antes de gerarmos as figuras.

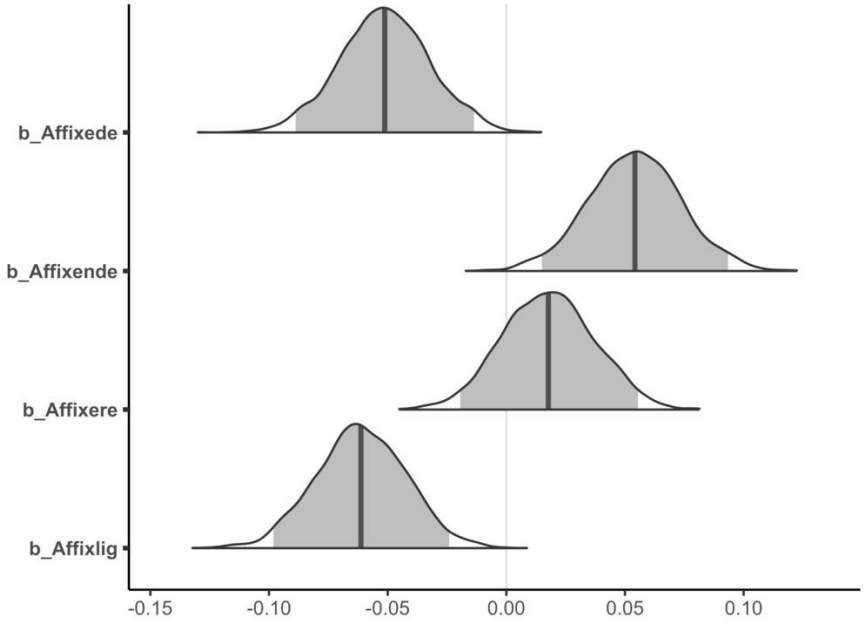

FIGURA 5 – Gráfico de com *a posteriori* para afixos.

```
31| # Definir tema minimalista para ggplot()
32| bayesplot::bayesplot_theme_set(new = theme_classic())

34| # Gráfico de traços
35| bayesplot::color_scheme_set("viridis") # definir esquema de cores para traços
36| bayesplot::mcmc_trace(fit,
          pars = c("b_Affixede", "b_Affixende", "b_Affixere", "b_Affixlig"))
37| # ggsave(filename = "trace.jpeg", height = 4.5, width = 7, dpi = 1000)

39| # A posteriori para afixos
40| bayesplot::color_scheme_set("darkgray")
41| bayesplot::mcmc_areas(fit,
          pars = c("b_Affixede", "b_Affixende", "b_Affixere", "b_Affixlig"),
42|                 prob = 0.95, point_est = "mean")
43| # ggsave(filename = "posteriors.jpeg", height = 4, width = 5.5, dpi = 1000)
```

QUADRO 4 – Gerando figuras para diagnóstico e resultados do modelo em Bayes.

O gráfico da figura 5 reforça a vantagem de modelos bayesianos em apresentar os resultados na forma de distribuições de probabilidades em vez de coeficientes pontuais (*point estimates*). Isso adiciona a dúvida e incerteza que deve ser natural quando se busca inferir parâmetros desconhecidos de uma população com base em uma amostra.

Como os dados que utilizamos envolvem múltiplas coletas dos mesmos participantes e com as mesmas palavras, o ideal é que sua análise inclua efeitos aleatórios para "Subject" e "Word". Para o pesquisador familiarizado com modelos de efeitos mistos (hierárquicos ou multinível), a tarefa é bastante simples: basta adicionar `(1 | Subject)` e `(1 | Word)` para interceptos aleatórios para falantes e para palavras, respectivamente, ao comando da linha 25[11] do quadro 2. O *script* completo disponibilizado em https://osf.io/bvj4w/ contém esse modelo e, ao rodá-lo, verifica-se que o modelo adiciona um pouco mais de dúvida aos coeficientes, alargando seus intervalos de credibilidade.

4. Considerações finais e sugestões de leitura

Neste artigo, introduzimos brevemente uma análise de dados bayesiana a partir de um modelo de regressão linear. Naturalmente, qualquer modelo estatístico pode ser rodado de forma bayesiana—o pacote `brms` está preparado para rodar os principais modelos de regressão utilizados em estudos linguísticos. Como vimos, modelos bayesianos são superiores a modelos frequentistas porque (i) apresentam um resultando mais relevante, ou seja, P(H|E) ao invés de P(E|H)—o que automaticamente remove a binaridade simplista de valores de *p*; (ii) possuem uma interpretação mais intuitiva (e.g., sem valores de *p* ou intervalos de confiança); (iii) oferecem um alto grau de personalização, especialmente através da especificação de uma distribuição *a priori*, que nos permite incorporar à análise estatística nosso conhecimento de área e resultados de estudos anteriores. Além dessas vantagens, amostras do *a posteriori* proporcionam uma "imagem de alta resolução" sobre tamanhos de efeito, uma vez que temos acesso a uma distribuição inteira sobre os efeitos mais plausíveis a partir dos dados observados (e de nosso *a priori*).

Dadas as vantagens de métodos bayesianos apontadas aqui e pela literatura, por que, afinal, deveríamos utilizar um modelo tradicional frequentista quando há uma alternativa mais vantajosa? Há, pelo menos, duas razões por que uma migração total para modelos em Bayes talvez não seja tão simples—ambas as razões são externas ao método *per se*. A primeira, mencionada ao longo deste artigo, é a intensidade computacional envolvida: um modelo bayesiano exige mais tempo para convergir. Esse problema raramente será tão grave, uma vez que não costumamos usar *big data* com frequência em linguística (compare, por exemplo, com estudos em genética)—além disso, como mencionado acima, podemos utilizar múltiplos núcleos para tirarmos amostras do *a posteriori*. Ainda assim, o problema pode ser levemente inconveniente. Uma sugestão é iniciar a análise estatística com modelos tradicionais, com o objetivo de explorar efeitos iniciais rapidamente, e, subsequentemente, migrar para um

[11] Ou (Affix | Subject) para interceptos e *slopes* aleatórios para falantes.

modelo em Bayes quando a definição das variáveis estiver mais clara—leve em conta que um modelo em Bayes com *a priori* não informativo resultará em efeitos bastante similares aos de um modelo equivalente frequentista na maioria das vezes.

O principal desafio na migração para Bayes, contudo, será a aceitação da área. Como análises bayesianas não são tão comuns em boa parte das subáreas em linguística, especialmente no Brasil, haverá certo estranhamento por parte de pareceristas e leitores, que estarão acostumados a ver valores de *p* atrelados a resultados estatísticos. Além disso, o conceito de *a priori* informativos pode ser visto como problemático por quem não está familiarizado com análises bayesianas—consulte Gelman (2008) sobre objeções comuns. Ou seja, análises em Bayes talvez precisem (a) ser acompanhadas de informações fundamentais sobre o método, e (b) apresentar uma interpretação dos resultados mais detalhada.

Ao migrarmos de modelos frequentistas para modelos bayesianos, é saudável desenvolvermos alguns costumes específicos. Por exemplo, como modelos em Bayes levam consideravelmente mais tempo para rodar, é uma excelente ideia salvarmos o output do modelo em formato `RData`—afinal, não queremos ter de rodar o mesmo modelo a cada vez que revisitarmos nosso script. Você pode ler mais sobre esse formato de dados em Garcia (2021), capítulo 10.

Por fim, uma dúvida comum é: se a análise bayesiana é superior, ainda devemos estudar ou ensinar métodos frequentistas tradicionais em programas de pós-graduação? Em primeiro lugar, estatística frequentista ainda faz parte da imensa maioria dos estudos linguísticos—no Brasil e fora dele. Em segundo lugar, sempre haverá centenas ou milhares de estudos relevantes publicados com método frequentista, e lê-los criticamente nunca deixará de ser uma habilidade fundamental a qualquer pesquisador. Em terceiro lugar, estudantes de pós-graduação em linguística frequentemente têm uma base frágil em estatística, e um foco em Bayes sem um pilar em estatística tradicional pode ser ineficiente do ponto de vista pedagógico. É preciso entender uma análise de dados bayesiana como um "próximo passo", ou um método complementar, e não como um substituto de métodos frequentistas. Dominar fundamentos bayesianos e frequentistas é naturalmente a melhor opção.

Não poderíamos encerrar este artigo sem recomendações adicionais de leitura. Garcia (2021), por exemplo, apresenta um capítulo inteiro dedicado à análise bayesiana. O capítulo parte do zero, contém códigos em R comentados, e, assim como o presente artigo, utiliza o pacote `brms`—sendo, portanto, bastante amigável. Os dois principais livros inteiramente dedicados à análise bayesiana que recomendaríamos são Kruschke (2015) e McElreath (2020), que detalham minuciosamente conceitos e implementação de modelos em Bayes. Por fim, Gelman et al. (2014) apresentam uma referência completa (embora menos amigável) sobre modelos em Bayes.

REFERÊNCIAS

ARANTES, Pablo; LIMA Jr, Ronaldo Mangueira (2021). Using a Coupled-Oscillator Model of Speech Rhythm to Estimate Rhythmic Variability In Two Brazilian Portuguese Varieties (CE and SP). Cadernos de Linguística, v. 2, n. 4, e577. http://doi.org. 10.25189/2675-4916.2021.V2.N4.ID577

BAAYEN, Rolf Harald. *languageR*: v 1.0, 2007a.

BAYES, Thomas. LII. An essay towards solving a problem in the doctrine of chances. De autoria do falecido reverendo Sr. Bayes, F.R.S. comunicado pelo Sr. Price, em uma carta para John Canton, A.M.F.R.S. *Philosophical Transactions* (1683–1775), v. 53, pp. 370–418, 1763.

BÜRKI, Audrey; ELBUY, Shereen; MADEC, Sylvain; VASISHTH, Shravan. What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis. *Journal of Memory and Language*, v. 114, pp. 104–125, 2020. http://dx.doi.org/10.1016/j.jml.2020.104125.

BÜRKNER, Paul-Christian. Why not to be afraid of priors (too much). In: Bayes@Lund 2018, Lund: 2018a. Disponível em < https://www.youtube.com/watch?v=Uz9r8eV2erQ>. Acesso em: 24 ago. 2021.

BÜRKNER, Paul-Christian. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, v. 10, n. 1, pp. 395–411, 2018b.

CARPENTER, Bob; GELMAN, Andrew.; HOFFMAN, Matthew.; LEE, Daniel; GOODRICH, Ben; BETANCOURT, Michael; BRUBAKER, Marcus; GUO, Jiqiang; LI, Peter; RIDDELL, Allen. Stan: a probabilistic programming language. *Journal of Statistical Software*, Articles, v. 76, n. 1, pp. 1–32, 2017.

CHATER, Nick; TENENBAUM, Joshua B.; YUILLE, Alan. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, v. 10, n. 7, pp. 287–344, 2006.

GARCIA, Guilherme D. When lexical statistics and the grammar conflict: learning and repairing weight effects on stress. *Language* 95(4):612–641, 2019. http://doi.org/10.1353/lan.2019.0068.

GARCIA, Guilherme D. Language transfer and positional bias in English stress. *Second Language Research*, v. 34, n. 6, pp. 445–474, 2020. http://doi.org/10.1177/0267658319882457

GARCIA, Guilherme D. *Data visualization and analysis in second language research*. Routledge, Nova York, NY, 2021.

GELMAN, Andrew. Objections to Bayesian statistics. *Bayesian Analysis*, v. 3, n. 3, pp. 445–449, 2008.

GELMAN, Andrew; CARLIN, John B.; STERN, Hal S.; DUNSON, David B.; VEHTARI, Aki; RUBIN, Donald B. *Bayesian data analysis*. 3rd ed. Chapman & Hall/CRC, Boca Raton, 2014.

HACKING, Ian. *An introduction to probability and inductive logic*. Cambridge University Press, 2001.

HAYES, Bruce; SIPTÁR, Péter; ZURAW, Kie; & LONDE, Zsuzsa. Natural and unnatural constraints in Hungarian vowel harmony. *Language*, 822-863, 2009. http://www.jstor.org/stable/40492955.

IDSARDI, William. A Bayesian approach to loanword adaptations. Poster presented at the *Annual Meeting of the Linguistic Society of America*, Albuquerque, NM, 2006.

KRUSCHKE, John K. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, v. 142, n. 2, pp. 573–603, 2013. https://doi.org/10.1037/a0029146

KRUSCHKE, John K. *Doing Bayesian data analysis*: a tutorial with R, JAGS, and Stan, 2a edição. Elsevier, 2015.

LEE, Michael D.; WAGENMAKERS, Eric-Jan. *Bayesian cognitive modeling: a practical course*. Cambridge University Press, Cambridge, 2014.

LIMA JR., Ronaldo; GARCIA, Guilherme D. Diferentes análises estatísticas podem levar a conclusões categoricamente distintas. *Revista da* ABRALIN, v. 20, n. 1, pp. 1–19, 2021. https://doi.org/10.25189/rabralin.v20i1.1790

MCELREATH, Richard. *Statistical rethinking: A Bayesian course with examples in R and Stan*, 2a edição. Boca Raton & Oxon: CRC press, 2020.

MCGRAYNE, Sharon Bertsch. *The theory that would not die*. Yale University Press, 2011.

NUZZO, Regina. Scientific method: statistical errors. *Nature News*, v. 506, n. 7487, p. 150, 2014.

R CORE TEAM. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Acesso http://www.R-project.org/ Acesso em: 15 jun. 2020.

RSTUDIO TEAM. RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, 2021.

TENENBAUM, Joshua B.; GRIFFITHS, Thomas L.; KEMP, Charles. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, v. 10, n. 7, pp. 309–318, 2006.

WICKHAM, Hadley; AVERICK, Mara; BRYAN, Jennifer; CHANG, Winston; MCGOWAN, Lucy D.; FRANÇOIS, Romain; GROLEMUND, Garrett; HAYES, Alex; HENRY, Lionel; HESTER, Jim; KUHN, Max; PEDERSEN, Thomas L.; MILLER, Evan; BACHE, Stephan M.; MÜLLER, Kirill; OOMS, Jeroen; ROBINSON, David; SEIDEL, Dana P.; SPINU, Vitalie; TAKAHASHI, kohske; VAUGHAN, Davis; WILKE, Claus; WOO, Kara; YUTANI, Hiroaki. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686, 2019. https://doi.org/10.21105/joss.01686.

ZHAN, Meilin; LEVY, Roger; KEHLER, Andrew. Pronoun Interpretation in Mandarin Chinese follows principles of Bayesian inference. *PLoS One*, v. 15, n. 8, pp. 1–42, 2020. https://doi.org/10.1371/journal.pone.0237012.

Apêndice

A1. Demonstração de um modelo simples em Stan

O quadro A1 mostra um exemplo de uma regressão linear simples com um preditor contínuo: $y_n = \alpha_n + \beta_n x + \epsilon_n$. Aqui, nosso modelo tem três partes: "data", "parameters", e "model". Em "data", especificamos que tipo de dado queremos modelar. Nossa amostra tem tamanho N (um valor que é sempre positivo, naturalmente). Em seguida, especificamos nossa variável resposta, y, e nossa variável preditora, x, ambas contínuas neste exemplo hipotético. Em "parameters", temos "alpha" (nosso intercept), "beta" (o coeficiente de nosso preditor), e "sigma" (o desvio-padrão). Perceba que, diferentemente de um modelo frequentista, aqui estamos estimando também o desvio-padrão dos nossos dados. Por fim, em "model", temos a especificação do modelo. Observe como o modelo é especificado: `y ~ normal(alpha + beta * x, sigma)`. Se você já rodou uma regressão linear, essa linha deve fazer sentido: estamos basicamente dizendo que cada observação (resposta) em nossos dados segue uma distribuição normal. A média dessa distribuição normal é exatamente o que queremos estimar com nossa regressão. A diferença aqui é que também estimaremos o desvio-padrão dessa distribuição.

```
data {
  int<lower=0> N;
  vector[N] x;
  vector[N] y;
}
parameters {
  real alpha;
  real beta;
  real<lower=0> sigma;
}
model {
  y ~ normal(alpha + beta * x, sigma);
}
```

QUADRO A1 – Exemplo simplificado de uma regressão linear em Stan.

A sintaxe no quadro A1 é bastante simplificada, e esconde algo muito importante: a linha `y ~ normal(alpha + beta * x, sigma)` é vetorizada,[12] ou seja, não precisamos adicionar um *for-loop*. Naturalmente, a notação de Stan vai muito além do exemplo simples acima.

[12] Especificações mais atuais utilizarão a função especialmente criada para regressões `normal_id_glm()` em vez de `normal()`.